Multioverlap simulations for transitions between reference configurations

Bernd A. Berg,[1,2,*] Hiroshi Noguchi,[3,†] and Yuko Okamoto[3,4,‡]
[1]*Department of Physics, Florida State University, Tallahassee, Florida 32306, USA*
[2]*School of Computational Science and Information Technology, Florida State University, Tallahassee, Florida 32306, USA*
[3]*Department of Theoretical Studies, Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan*
[4]*Department of Functional Molecular Science, Graduate University for Advanced Studies, Okazaki, Aichi 444-8585, Japan*
(Received 3 May 2003; published 23 September 2003)

We introduce a procedure to construct weight factors, which flatten the probability density of the overlap with respect to some predefined reference configuration. This allows one to overcome free-energy barriers in the overlap variable. Subsequently, we generalize the approach to deal with the overlaps with respect to two reference configurations so that transitions between them are induced. We illustrate our approach by simulations of the brain peptide Met-enkephalin with the ECEPP/2 (Empirical Conformational Energy Program for Peptides) energy function using the global-energy-minimum and the second lowest-energy states as reference configurations. The free energy is obtained as functions of the dihedral and the root-mean-square distances from these two configurations. The latter allows one to identify the transition state and to estimate its associated free-energy barrier.

DOI: 10.1103/PhysRevE.68.036126 PACS number(s): 05.10.Ln, 87.53.Wz, 87.14.Ee, 87.15.Aa

I. INTRODUCTION

Markov chain Monte Carlo (MC) simulations, for instance, by means of the Metropolis method [1], are well suited to simulate generalized ensembles. Generalized ensembles do not occur in nature, but are of relevance for computer simulations (see Refs. [2–4] for recent reviews). They may be designed to overcome free-energy barriers, which are encountered in Metropolis simulations of the Gibbs-Boltzmann canonical ensemble. Generalized ensembles do still allow for rigorous estimates of the canonical expectation values, because the ratios between their weight factors and the canonical Gibbs-Boltzmann weights are exactly known.

Umbrella sampling [5] was one of the earliest generalized-ensemble algorithms. In the multicanonical approach [6,7] one weights with a microcanonical temperature, which corresponds, in a selected energy range, to a working estimate of the inverse density of states. Expectation values of the canonical ensembles can be constructed for a wide temperature range, hence the name ''multicanonical.'' Here, ''working estimate'' means that running the updating procedure with the (fixed) multicanonical weight factors covers the desired energy range. The Markov process exhibits random walk behavior and moves in cycles from the maximum (or above) to the minimum (or below) of the chosen energy range, and back. A working estimate of the multicanonical weights allows for calculations of the spectral density and all related thermodynamical observables with any desired accuracy by simply increasing the MC statistics. Thus, we have a two-step approach: The first step is to obtain the working estimate of the weights and the second step is to perform a long production run with these weights. There is no need for that estimate to converge towards the exact inverse spectral density. Once the working estimate of the weights exists, MC simulations with frozen weights converge and allow one to calculate thermodynamical observables with, in principle, arbitrary precision. Various methods, ranging from finite-size scaling estimates [8] in case of suitable systems to general purpose recursions [9–11], are at our disposal to obtain a working estimate of the weights.

In the present paper we deal with a variant of the multicanonical approach: Instead of flattening the energy distribution, we construct weights to flatten the probability density of the overlap with a given reference configuration. This allows one to overcome energy barriers in the overlap variable and to get accurate estimates of thermodynamic observables at overlap values which are rare in the canonical ensemble. A similar concept was previously used in spin glass simulations [12], but there is a crucial difference: In Ref. [12] the weighting was done for the self-overlap of two replicas of the system and a proper name would be multi-self-overlap simulations, while in the present paper we are dealing with the overlap to a predefined configuration.

We next generalize our approach to deal with two reference configurations so that transitions between them become covered and our method then allows one to estimate the transition states and its associated free-energy barrier. We have in mind situations where experimentalists determined the reference configurations and observed transitions between them, but an understanding of the free-energy landscape between the configurations is missing. An example would be the conversion from a configuration with α helix structures to a native structure which is mostly in the β sheet, as it is the case for β-lactoglobulin [13,14].

The paper is organized as follows. In the following section we describe the algorithmic details, using first one and then two reference configurations. In particular, a two-step updating procedure is defined, which is typically more efficient than the conventional one-step updating. Moreover,

*Email address: berg@csit.fsu.edu
†Present address: Institut für Festkörperforschung, Forschungszentrum Jülich, D-52425 Jülich, Germany. Email address: hi.noguchi@fz-juelich.de
‡Email address: okamotoy@ims.ac.jp

based on the sums of uniformly distributed random numbers, a method to obtain a working estimate of the multioverlap weights is introduced. In Sec. III we illustrate the method for a simulation with the pentapeptide Met-enkephalin. Our simulations use the all-atom energy function ECEPP/2 (empirical conformational energy program for peptides [15]) and rely on its implementation in the computer package SMMP (simple molecular mechanics for proteins [16]). We use as reference configurations the global energy minimum (GEM) state, which has been determined by many authors [17–21], and the second lowest-energy state, as identified in Refs. [19,22]. While our overlap definition relies on a distance definition in the space of the dihedral angles, it turns out that for the data analysis the use of the root-mean-square (rms) distance is crucial. It is only in the latter variable that one obtains a clear picture of the transition saddle point in the two-dimensional free-energy diagram. In the final section a summary of the present results and an outlook with respect to future applications are given.

II. MULTIOVERLAP METROPOLIS ALGORITHM

In this section we explain the details of our multioverlap algorithm. The overlap of a configuration versus a reference configuration is defined in the following section. In Sec. II B we discuss details of the updating. To achieve step one of the method, i.e., the construction of a working estimate of the multioverlap weights, one could employ a similar recursion as the one used in Ref. [12] or explore the approach of Ref. [11]. Instead of doing so, we decided to test a new method: At infinite temperature, $\beta=0$, the overlap distributions can be calculated analytically (see Sec. II D). We use this as starting point and estimate the overlap weights at the desired temperature by increasing β in sufficiently small steps so that the entire overlap range remains covered. In the final section we define the overlap with respect to two distinct reference configurations to cover the transition region between them.

A. Definition of the overlap

There is a considerable amount of freedom in defining the overlap of two configurations. For instance, one may rely on the rms distance between configurations, and in Sec. III D we analyze some of our results with this variable. However, the computation of the rms distance is slow and for MC calculations it is important to rely on a computationally fast definition. Therefore, we define the overlap in the space of dihedral angles by, as it was already used in Ref. [24],

$$q=(n-d)/n, \tag{1}$$

where n is the number of dihedral angles and d is the distance between configurations defined by

$$d=||v-v^1||=\frac{1}{\pi}\sum_{i=1}^{n} d_a(v_i,v_i^1). \tag{2}$$

Here, v_i is our generic notation for the dihedral angle i, $-\pi<v_i\leqslant\pi$, and v^1 is the vector of dihedral angles of the reference configuration. The distance $d_a(v_i,v_i')$ between two angles is defined by

$$d_a(v_i,v_i')=\min(|v_i-v_i'|,2\pi-|v_i-v_i'|). \tag{3}$$

The symbol $||.||$ defines a norm in a vector space. In particular, the triangle inequality holds:

$$||v^1-v^2||\leqslant||v^1-v||+||v-v^2||. \tag{4}$$

For a single angle we have

$$0\leqslant|v_i-v_i^1|\leqslant\pi\Rightarrow 0\leqslant d\leqslant n. \tag{5}$$

At $\beta=0$ (i.e., infinite temperature)

$$d_i=\frac{1}{\pi}d_a(v_i,v_i^1) \tag{6}$$

is a uniformly distributed random variable in the range $0\leqslant d_i\leqslant 1$ and the distance d in Eq. (2) becomes the sum of n such uniformly distributed random variables, which allows for an exact calculation of its distribution.

B. Multioverlap weights

We choose a reference configuration of n dihedral angles v_i^1 $(i=1,\ldots,n)$, to define the dihedral distance (2). We want to simulate the system with weight factors that lead to a random walk (RW) process in the dihedral distance d,

$$d<d_{\min}\rightarrow d>d_{\max} \text{ and back.} \tag{7}$$

Here, $d_{\min}$ is chosen sufficiently small so that one can claim that the reference configuration has been reached, e.g., a few percent of $n/2$, which is the average d at $T=\infty$. The value of $d_{\max}$ has to be sufficiently large to introduce a considerable amount of disorder, e.g., $d_{\max}=n/2$. In the following we call one event of form (7) a random walk cycle (RWC).

One possibility is to choose weight factors which give a flat probability density in the dihedral distance range $0\leqslant d\leqslant n/2$, falling off for $d>n/2$ by keeping the d dependence of the weight constant for $d\geqslant n/2$. This is quite similar to multimagnetical simulations [8], for which the external magnetic field takes the place of the reference configuration. The analogy becomes obvious, when the external field is defined via a ghost spin, which couples to all other spins. For instance, the spins $\vec{s}$ of the Heisenberg ferromagnet are three-dimensional vectors of magnitude one. Their interaction with an external magnetic field $\vec{H}$ can be written as

$$\vec{H}\cdot\sum_i \vec{s}_i=H\sum_i \vec{s}_H\cdot\vec{s}_i=N\,H\,q, \tag{8}$$

where $\vec{s}_H$ is the unit vector in the direction of the magnetic field, $\vec{s}_i$ is the Heisenberg spin at site i, N is the number of spins, and q is the overlap of the spin configuration with the reference configuration $\vec{s}_H$:

$$q=\frac{1}{N}\sum_i \vec{s}_H\cdot\vec{s}_i. \tag{9}$$

Using the multioverlap language [12], the multimagnetical [8] weight factors may then be rewritten as

$$\exp[-\beta E+S(q)]=w_c(E)\,w_q(q), \tag{10}$$

where

$$w_c(E)=\exp(-\beta E), \tag{11}$$

and $E=-\Sigma_{\langle ij\rangle}\vec{s}_i\cdot\vec{s}_j$ is energy function of the Heisenberg ferromagnet (the sum is over nearest neighbor spins). Here, $S(q)$ has the meaning of a microcanonical entropy of the overlap parameter, which has to be determined so that the probability density becomes flat in q. Weights for other than the flat distribution have also been discussed in the literature, e.g., Ref. [25], on which we shall comment in connection with Fig. 7 below.

C. The updating procedure

In essence, there are two ways to implement the update.

(1) Combine the multioverlap and the canonical weights to one probability, which is accepted or rejected in one random step.

(2) Accept or reject the multioverlap and the canonical probabilities sequentially in two random steps.

1. One-step updating

As defined in Eqs. (10) and (11), the weight factor is a product of $w_c(E)$ and $w_q(d)$, where $w_c(E)$ is the usual canonical Gibbs-Boltzmann factor and $w_q(d)$ is the multioverlap weight factor, where we now use the distance d from the reference configuration (instead of the overlap q) as argument. As is clear from Eq. (1), the use of either q or d as argument is equivalent, while in the presentation of results the use of either variable can have intuitive advantages. In the one-step updating we combine the weights to

$$w(E,d)=w_c(E)\,w_q(d), \tag{12}$$

and accept or reject newly proposed configurations in the standard Metropolis way. Notably, the calculation of $w_q(d)$ (a simple table lookup) is very fast compared with the calculation of $w_c(E)$. Therefore, the following two-step procedure is of interest.

2. Two-step updating

Suppose that the present configuration is (d,E) and a new configuration (d',E') is proposed:

$$(d,E)\rightarrow(d',E'). \tag{13}$$

We can sequentially first accept or reject with the $w_q(d)$ probabilities and then conditionally, when the d part is accepted, with the $w_c(E)$ probabilities.

Proof. We show detailed balance for two subsequent updates of the same dihedral angle with the two-step procedure. There are four cases with probabilities of acceptance:

$$P_i,\quad i=1,2,3,4. \tag{14}$$

They are listed in the following:

$$\text{Case 1: } w_q(d')\geqslant w_q(d) \quad\text{and}\quad w_c(E')\geqslant w_c(E),$$
$$P_1=1, \tag{15}$$

$$\text{Case 2: } w_q(d')\geqslant w_q(d) \quad\text{and}\quad w_c(E')<w_c(E),$$
$$P_2=w_c(E')/w_c(E), \tag{16}$$

$$\text{Case 3: } w_q(d')<w_q(d) \quad\text{and}\quad w_c(E')\geqslant w_c(E),$$
$$P_3=w_q(d')/w_q(d), \tag{17}$$

$$\text{Case 4: } w_q(d')<w_q(d) \quad\text{and}\quad w_c(E')<w_c(E),$$
$$P_4=w_q(d')w_c(E')/[w_q(d)\,w_c(E)]. \tag{18}$$

For the inverse move

$$(d',E')\rightarrow(d,E), \tag{19}$$

with probabilities of acceptance

$$P_i',\quad i=1,2,3,4, \tag{20}$$

the cases are the following:

$$\text{Case 1: } w_q(d)\leqslant w_q(d') \quad\text{and}\quad w_c(E)\leqslant w_c(E'),$$
$$P_1'=w_q(d)\,w_c(E)/[w_q(d')\,w_c(E')], \tag{21}$$

$$\text{Case 2: } w_q(d)\leqslant w_q(d') \quad\text{and}\quad w_c(E)>w_c(E'),$$
$$P_2'=w_q(d)/w_q(d'), \tag{22}$$

$$\text{Case 3: } w_q(d)>w_q(d') \quad\text{and}\quad w_c(E)\leqslant w_c(E'),$$
$$P_3'=w_c(E)/w_c(E'), \tag{23}$$

$$\text{Case 4: } w_q(d)>w_q(d') \quad\text{and}\quad w_c(E)>w_c(E'),$$
$$P_4'=1. \tag{24}$$

For the ratios we find

$$\frac{P_i}{P_i'}=\frac{w_q(d')\,w_c(E')}{w_q(d)\,w_c(E)}, \tag{25}$$

independently of $i=1,2,3,4$. Therefore, we have constructed a valid Metropolis updating procedure.

D. Sums of a uniformly distributed random variable

To calculate the overlap weights at infinite temperature, we consider the sum

$$u^r = x_1^r + \cdots + x_n^r \tag{26}$$

of the random variables $x_j^r (j=1,\ldots,n)$, each uniformly distributed in the interval $[0,1)$ and derive a recursion formula for the probability density $f_n(u)$ of this distribution. Care is taken to cast the recursion in a form which allows for a numerically stable implementation [26] over a reasonably large range of n.

Let us recall the probability density of the uniform distribution:

$$f_1(x) = \begin{cases} 1 & \text{for } 0 \leqslant x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

To derive the recursion formula for the probability density of the random variable (26), it is convenient to cast it in the form

$$f_n(u) = \sum_{k=1}^{n} f_{n,k}(x_k) \quad \text{with} \quad x_k = u - k + 1, \tag{28}$$

where

$$f_{n,k}(x) = \begin{cases} \sum_{i=0}^{n-1} a_{n,k}^i x^i & \text{for } 0 \leqslant x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

The master formula for the recursion is obtained from the convolution

$$f_n(u) = \int_0^u f_1(u-v) f_{n-1}(v)\, dv. \tag{30}$$

The distributions of sums of random variables are often elegantly obtained from the inverse transformation of their product in Fourier space (e.g., Ref. [26]). However, for the uniform distribution this approach leads to a rather complicated inverse transformation. Let now $u = x + k - 1$ with $0 \leqslant x < 1$, and Eqs. (27)–(29) imply

$$\begin{aligned} f_{n,k}(x) &= \int_{k-2+x}^{k-1+x} f_{n-1}(v)\, dv \\ &= \int_x^1 f_{n-1,k-1}(y)\, dy + \int_0^x f_{n-1,k}(y)\, dy. \end{aligned} \tag{31}$$

Using Eq. (29) and performing the integrations, we obtain

$$\begin{aligned} f_{n,k}(x) = &\sum_{i=0}^{n-2} a_{n-1,k-1}^i \frac{1}{i+1} - \sum_{i=0}^{n-2} a_{n-1,k-1}^i \frac{x^{i+1}}{i+1} \\ &+ \sum_{i=0}^{n-2} a_{n-1,k}^i \frac{x^{i+1}}{i+1}. \end{aligned} \tag{32}$$

Expanding in powers of x and comparing Eq. (29) with Eq. (32) allows one to calculate the coefficients $a_{n,k}^i$ recursively in a numerically robust way:

$$a_{n,k}^0 = \sum_{j=0}^{n-1} \frac{a_{n-1,k-1}^j}{j+1}, \quad a_{n,k}^i = \sum_{j=0}^{n-1} \frac{a_{n-1,k}^j - a_{n-1,k-1}^j}{j+1}. \tag{33}$$

Once the coefficients $a_{n,k}^i$ are available, one can easily evaluate the probability densities $f_n(u)$ and the corresponding cumulative distribution functions.

The probability density (28) takes its maximum value for $u = n/2$. Due to the central limit theorem the fall-off behavior is Gaussian as long as u stays sufficiently close to $n/2$. In the tails, for $u \to 0$ or $u \to n$, the fall off is much faster than Gaussian, namely, an exponential of an exponential as follows from extreme value statistics [27].

E. Combination of two weights

In the following the weights with superscript j, $w_q^j(d_j)$, correspond to two distinct reference configurations v^j, $(j=1,2)$, and d_j is the distance from the configuration at hand to the configuration v^j. Let us assume that multioverlap simulations with respect to the two reference configurations have been carried out and that the weights, $w_q^1(d_1)$ and $w_q^2(d_2)$, have been determined so that they sample their distance distributions approximately uniformly.

We want to construct combined weights $w_q^{12}(d_1,d_2)$ which lead to a RW process between the configurations v^1 and v^2. Our choice is

$$w_q^{12}(d_1,d_2) = \begin{cases} w_q^1(d_1) & \text{for } d_1 < d_2, \\ c_j\, w_q^2(d_2) & \text{for } d_1 \geqslant d_2. \end{cases} \tag{34}$$

The constant c_j, with j either 1 or 2, is introduced to allow for smooth transitions from $d_1 < d_2$ to $d_1' \geqslant d_2'$ and vice versa. We determine c_j from the analysis of either run 1 (or run 2), which are the (one reference configuration) simulations leading to the weights $w_q^1(d_1)$ [or $w_q^2(d_2)$]. The constant c_1 is found from run 1 by scanning the time series for configuration for which $d_1 \geqslant d_2$ holds and which have a one-update transition $(d_1,d_2) \to (d_1',d_2')$ with $d_1' < d_2'$. From these configurations k we determine the constant c_1 so that

$$\sum_k w_q^1[d_1(k)] = c_1 \sum_k w_q^2[d_2(k)] \tag{35}$$

holds. Similarly, run 2 may be used to get c_2. It turns out that the normalized weights almost agree in the transition region and, therefore, the patching (34) works. The dependence of the constant on the run used for its determination is small, and it appears not worthwhile to explore more sophisticated methods.

It is straightforward to implement the Metropolis updating with respect to weights (34). For the transition

$$(d_1,d_2) \to (d_1',d_2'), \tag{36}$$

one has to distinguish four more cases as follows:

$$\text{Case 1:} \quad d_1 < d_2 \quad \text{and} \quad d_1' < d_2', \tag{37}$$

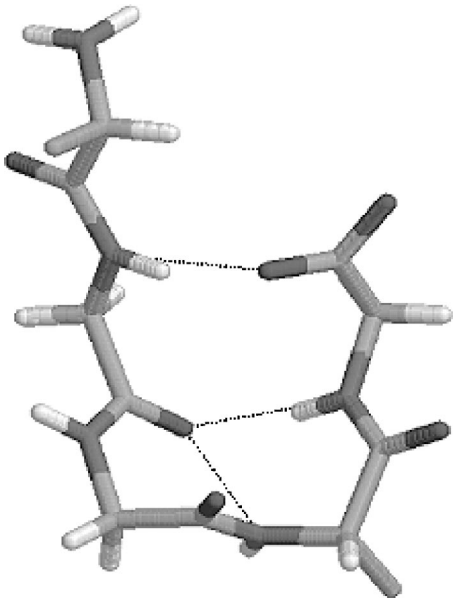

FIG. 1. Reference configuration 1. Only backbone structure is shown. The *N*-terminus is on the left-hand side and the C-terminus on the right-hand side. The dotted lines stand for hydrogen bonds. The figure was created with RasMol [23].

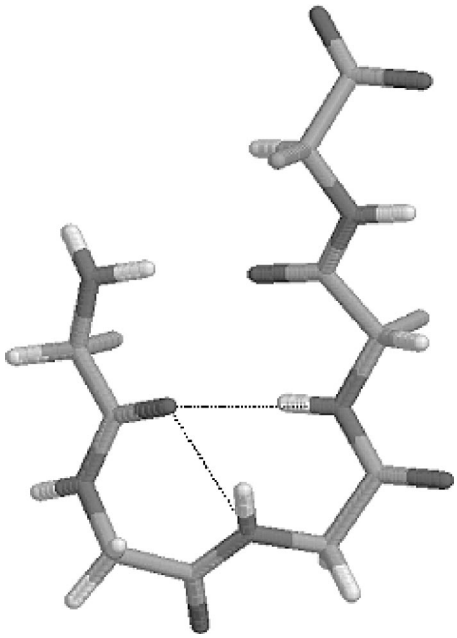

FIG. 2. Reference configuration 2. See the caption of Fig. 1 for details.

$$\text{Case 2: } d_1 < d_2 \quad \text{and} \quad d_1' \geqslant d_2', \tag{38}$$

$$\text{Case 3: } d_1 \geqslant d_2 \quad \text{and} \quad d_1' < d_2', \tag{39}$$

$$\text{Case 4: } d_1 \geqslant d_2 \quad \text{and} \quad d_1' \geqslant d_2'. \tag{40}$$

Alternatively to the approach outlined, one may combine d_1 and d_2 into a new variable θ_d for which the weights are then calculated as in the one-dimensional case. A suitable choice along this line is

$$\theta_d = \frac{2}{\pi}\arctan\left(\frac{d_1}{d_2}\right). \tag{41}$$

III. MET-ENKEPHALIN SIMULATIONS

In the following we introduce two reference configurations. Subsequently, we discuss first the results for simulations with one reference configuration and then those involving both reference configurations.

A. The reference configurations

Met-enkephalin has the amino-acid sequence Tyr-Gly-Gly-Phe-Met. We fix the peptide-bond dihedral angles ω to 180°, which implies that the total number of variable dihedral angles is $n=19$. We neglect the solvent effects as in previous works. The low-energy configurations of Met-enkephalin in the gas phase have been classified into several groups of similar structures [19,22]. Two reference configurations, called configuration 1 and configuration 2, are used in the following and depicted in Figs. 1 and 2, respectively. Configuration 1 has a β-turn structure with hydrogen bonds between Gly-2 and Met-5 and configuration 2 a β-turn with a hydrogen bond between Tyr-1 and Phe-4 [22].

For our present work the two reference configurations were obtained by minimizing the GEM and the second lowest-energy state of previous literature with respect to the ECEPP/2 energy function. The minimization was performed with the SMMP minimizer [16] and by quenching. Both methods gave identical final energies. In Table I we list the vari-

TABLE I. Met-enkephalin reference configurations. The columns GEM_{min} and B_{min} correspond to configuration 1 and configuration 2, respectively.

Residue	Angle	GEM [21]	GEM_{min}	B [19]	B_{min}
1	χ_1	−179.9	−179.8	−179	+179.4
1	χ_2	−111.3	−111.4	−95	−94.3
1	χ_6	+145.3	+145.3	+169	−179.9
1	ϕ	−86.4	−86.3	+111	+55.7
2	ψ	+153.7	+153.7	+157	+157.6
2	ϕ	−161.6	−161.5	−71	−70.7
3	ψ	+71.2	+71.1	+78	+78.0
3	ϕ	+64.1	+64.1	159	+156.5
4	ψ	−93.5	−93.5	−37	−35.7
4	χ_1	+179.8	+179.8	+59	+55.3
4	χ_2	+80.0	+80.0	+87	+86.8
4	ϕ	−81.7	−81.7	−154	−155.7
5	ψ	−29.2	−29.2	+151	+151.6
5	χ_1	−65.1	−65.1	−68	−69.4
5	χ_2	−179.2	−179.2	+177	−176.3
5	χ_3	−179.3	−179.3	−179	−179.7
5	χ_4	−60.0	−59.9	+60	+59.9
5	ϕ	−80.8	−80.7	−140	−140.0
5	ψ_t	+143.9	+143.5	−29	−30.6

TABLE II. Energies (in kcal/mol) of the Met-enkephalin reference configurations 1 and 2.

	Total	Coulomb	Lennard-Jones	*H* Bond	Torsion
1	−10.72	+21.41	−27.10	−6.21	+1.19
2	−8.42	+22.59	−26.38	−4.85	+0.23

able dihedral angles of the configurations before and after this minimization. The initial dihedral angles for the GEM are taken from Table 1 of Ref. [21] and the initial dihedral angles for the second lowest-energy state *B* are from Table I of Ref. [19]. In Table I we give the angles in degrees, while for the MC simulations radians were used as in Eqs. (1) and (2) for the overlap. Our labeling of the residues follows the SMMP convention and deviates from those of Refs. [21,19].

The distance between the two minimized configurations is $d=6.62$ $(q=0.652)$ and their energies are given in Table II.

B. Simulations with one reference configuration

Each of our multioverlap simulations at fixed temperature relies on a statistics of 16 777 216 sweeps for which data are recorded in a time series of 524 288 events, i.e., with a step-size of 32 sweeps. We started most of our simulations with the GEM configuration, but some random starts were also performed and no noticeable differences were encountered.

Starting with the analytical result (28), valid at $\beta=0$, the weights are calculated by increasing β (i.e., decreasing the temperature) between simulations slowly so that the RW of each simulation still covers the desired overlap range when using the weight estimates from the previous temperature. Discretization errors due to histograming can be severe and instead of weights which are piecewise constant within each one histogram interval, we used the interpolation of Ref. [6]:

$$\ln w(d)=(1-\alpha)\ln w(d_i)+\alpha \ln w(d_{i+1}) \quad \text{for } d_i \leqslant d < d_{i+1}, \tag{42}$$

where

$$\alpha=\frac{d-d_i}{d_{i+1}-d_i}. \tag{43}$$

Figure 3 depicts the thus obtained weight function estimates from simulations with reference configuration 1. After five simulations we arrive at the physical temperature $T=300$ K. The same iteration works with reference configuration 2.

For the values $d_{\min}=0.025\,n$ and $d_{\max}=0.495\,n$, where $n=19$ is the number of angels in Eq. (2), we list in Table III the number of RWCs (7) achieved at each temperature. We also list the CPU time ratios for the one-step versus the two-step updating procedures, which we discussed in the preceding section. Especially at high temperatures, which are needed in our approach, the two-step updating turns out to be more efficient than the one-step updating and all of our production runs were done with it.

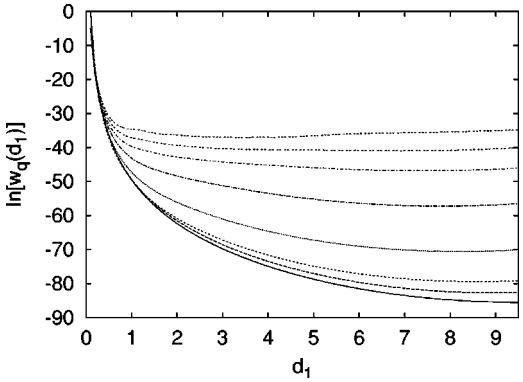

FIG. 3. Weight estimates from simulations with reference configuration 1. From up to down the weight functions correspond to the following temperatures: 230, 300, 400, 700, 2 000, 10 000, 100 000 K, and infinity $(\beta=0)$.

We next rely on the peaked distribution function [26] to visualize some of the data kept in the time series of our simulations. The peaked distribution function of a probability density $f(x)$ is defined by

$$F_{\text{peaked}}(x)=\begin{cases} F(x) & \text{for } F(x)\leqslant 0.5, \\ 1-F(x) & \text{for } F(x)>0.5, \end{cases} \tag{44}$$

where

$$F(x)=\int_{-\infty}^{x} dx'\, f(x') \tag{45}$$

is the usual cumulative distribution function (see, for instance, Ref. [28]).

To visualize how the canonical energy distribution moves when we lower the temperature, we plot in Fig. 4 the peaked energy distributions as obtained by reweighting some of the multioverlap simulations of Fig. 3 to the canonical ensemble of their simulation temperature. Due to the reweighting the distributions look precisely as one expects for energies from canonical MC simulations. In contrast to conventional canonical simulations, the raw data feature a considerably larger number of events at low energies. This is illustrated in Fig. 5, where we plot the 300 K and 400 K peaked distribution functions of Fig. 4 together with their raw multioverlap peaked distributions

TABLE III. Number of random walk cycles in the simulations with our two reference configurations. The last column lists the CPU time ratios for one-step versus two-step updating.

T (K)	Configuration 1	Configuration 2	One-step/two-step
100 000	9458	9514	3.0
10 000	3122	3149	1.8
2000	2893	2741	1.6
700	2169	2227	1.5
400	1342	1693	1.3
300	462	610	1.2
230	46	41	1.2

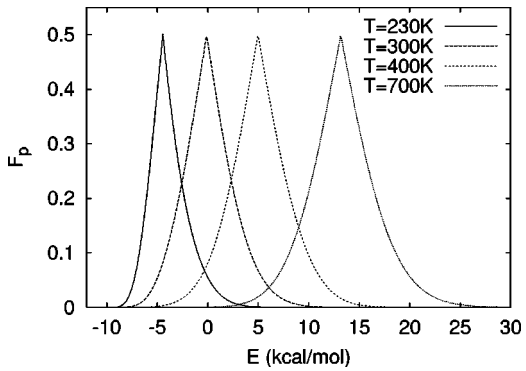

FIG. 4. Canonical, peaked energy distributions obtained by reweighting multioverlap simulations. From left to right the temperatures used are 230, 300, 400, and 700 K.

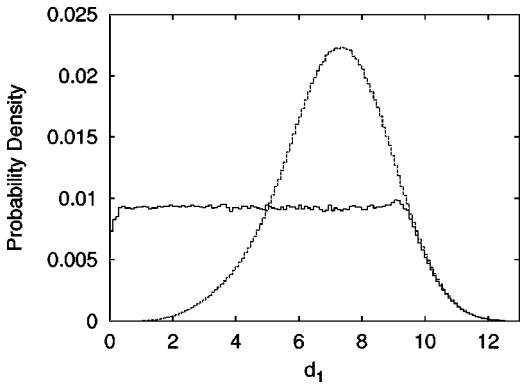

FIG. 6. Probability density of the distance from a multioverlap simulation at $T=400$ K (flat) and its canonically reweighted probability density (peaked).

In Fig. 6 we give an example of the probability density of the distance. For the 400 K simulation with reference configuration 1 we plot the probability density of d_1 as obtained from the multioverlap simulation together with its canonically reweighted probability density. The simulation itself is run with the multioverlap weights from the 700 K simulations and the multioverlap histogram shown is reweighted to the multioverlap 400 K weights. As expected, we have a flat distribution between 0 and $n/2=9.5$ (the latter is the average value of the distance at $T=\infty$). Moreover, there is a good coverage of configurations close to the GEM, which are highly suppressed in the 400 K canonical ensemble. The maximum ratio of the multioverlap density divided by the canonical density is 6×10^{16} in this plot.

For the same simulation Fig. 7 depicts separately the peaked distribution function of the forward and backward RWCs (7). A considerable asymmetry is noticeable and it turns out that the weights of the $1/k$ ensemble [25] lead to more RWCs than the flat distribution of Fig. 6. In connection with our simulations this is a lucky circumstance, because the $1/k$ distribution of weights is in essence the distribution at a somewhat higher temperature than that of the simulation. This increases the flexibility when estimating good weights at a lower temperature from the already existing simulation results at a higher temperature.

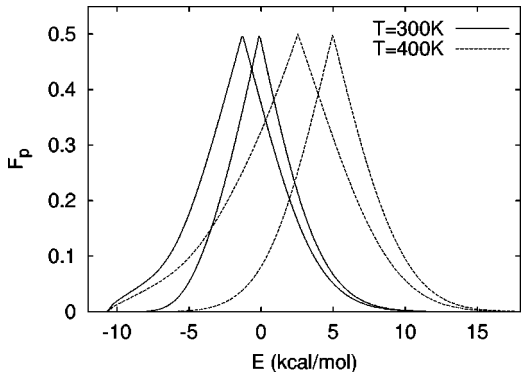

FIG. 5. Peaked multioverlap (left shifted) and canonical energy distributions at $T=300$ K and $T=400$ K.

For multioverlap simulations the reweighting towards low temperatures can work much better than for canonical simulations. This is due to the fact that the low-energy configurations close to low-energy reference configuration are already in the ensemble. This is illustrated in Fig. 8, where we reweight the data from a multioverlap simulation with reference configuration 1 at $T=300$ K and compare with a conventional multicanonical simulation based on the SMMP package [16]. The specific heat C_V and the derivative of the overlap with respect to the temperature are shown. From 200 K to 400 K the deviations of the results are of the order of the statistical errors, which are not shown for clarity of the figure. Below 200 K deviations of the reweighted overlap simulation from the correct behavior become visible, first in dq_1/dT then in C_V. Such deviations are expected as the low-energy attractor does not lead to a uniform coverage of all low-energy states. The successful reweighting from high simulation temperatures to lower temperatures is an improvement, because the Metropolis dynamics at high temperatures is faster. But the reweighting of a multioverlap simulation to a lower temperature will fail at some point, because the reference configuration introduces a bias towards particular low-energy configurations.

The temperature at which C_V and $-dq_1/dT$ take peak values correspond to the coil-globule transition temperature

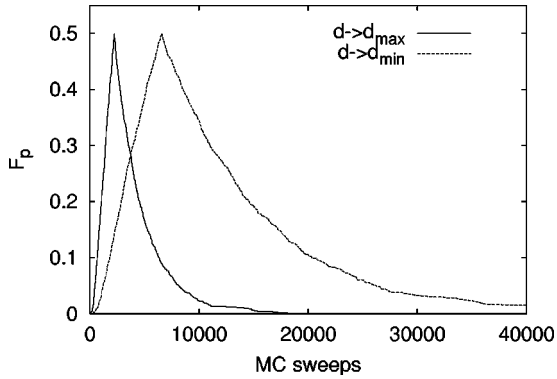

FIG. 7. Peaked distribution functions for the forward ($d\rightarrow d_{\max}$) and backward ($d\rightarrow d_{\min}$) parts of the random walk cycles from a multioverlap simulation at $T=400$ K.

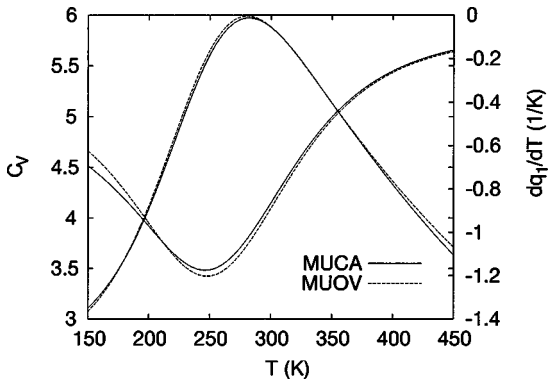

FIG. 8. Left-hand side ordinate: specific heat reweighted from a multicanonical (MUCA) and from a 300 K multioverlap (MUOV) simulation with reference configuration 1. Right-hand side ordinate: dq_1/dT reweighted from the same simulations, where q_1 is the overlap with reference configuration 1.

T_θ and the folding temperature T_f [24]. From Fig. 8 we read off the following approximate values:

$$T_\theta = 280 \text{ K} \quad \text{and} \quad T_f = 245 \text{ K}. \tag{46}$$

C. Simulations with two reference configurations

At 300 K we combine the weights from the runs with reference configurations 1 and 2 to one weight function according to our Eq. (34). We record now three different RWCs.

(1) With respect to reference configuration 1 from $d_{\min}$ to $d_{\max}$ and back, found 315 times.

(2) With respect to reference configuration 2 from $d_{\min}$ to $d_{\max}$ and back, found 545 times.

(3) From $d_{\min}$ of reference configuration 1 to $d_{\min}$ of reference configuration 2 and back, found 196 times.

In Fig. 9 we show the probability densities of this simulation with respect to the distances from our reference configurations. They are no longer flat, but a satisfactory coverage in the variables d_1 and d_2 is still achieved. Note that both probability densities have peaks at $d=6.62$, which is the distance between configurations 1 and 2. This implies

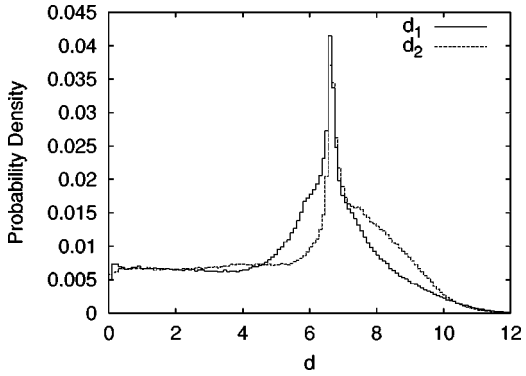

FIG. 9. Combined weight simulation at $T=300$ K: probability densities with respect to the distances d_1 and d_2.

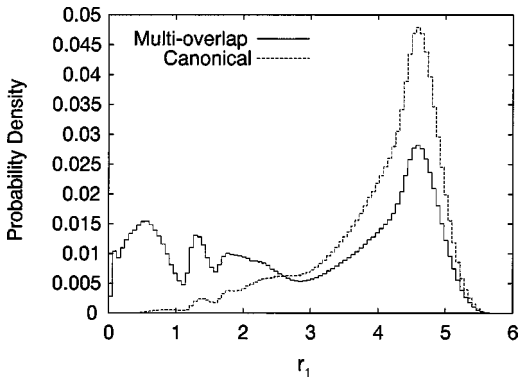

FIG. 10. Probability density of the rms distance from the multioverlap simulation at $T=400$ K of Fig. 6, and its canonically reweighted probability density. The abscissa is the rms distance (Å) in Eq. (47) from the reference configuration 1.

that both reference configurations have been visited with high probability.

D. Physics results

We would like to analyze the transitions between our two reference configurations in some detail. For this purpose we use the rms distance, which is defined by

$$d_{\text{rms}} = \min\left[\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\vec{x}_i - \vec{x}_i^{\,j})^2} \right], \tag{47}$$

where N is the number of atoms, $\{\vec{x}_i^{\,j}\}$ are the coordinates of the reference configuration j, and the minimization is over the translations and rotations of the coordinates of the configuration $\{\vec{x}_i\}$.

Distance (2) and the rms distance (47) are quite distinct. The reason is that a change of a single dihedral angle in the central parts of the molecule can cause a large deviation in the rms distance. Although the two configurations are then close-by from the point of view of the MC algorithm, physically they are rather far apart, as the similarity of the three-dimensional structures is governed by the rms distance. Therefore, the rms distance distribution deviates considerably from the dihedral distance distribution. We illustrate this by plotting in Fig. 10 the rms probability density of the 400 K simulation for which the dihedral distance probability density is shown in Fig. 6. Note that the rms distribution has a few peaks, i.e., stays kind of rough, despite the flat dihedral distance distribution.

We now analyze the free-energy landscape [29] from the results of our simulation with combined weights at 300 K in some detail. We study the landscape with respect to some reaction coordinates (and hence it should be called the potential of mean force). In order to study the transition states between reference configurations 1 and 2, we first plotted the free-energy landscape with respect to the distances d_1 and d_2. However, we did not observe any transition saddle point. A satisfactory analysis of the saddle point becomes possible when the rms distance (instead of the dihedral distance) is

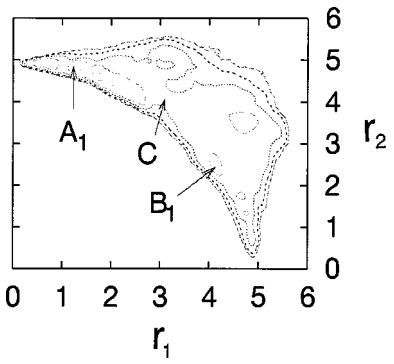

FIG. 11. Free-energy landscape at $T=250$ K with respect to rms distances (Å) from the two reference configurations, $F(r_1,r_2)$. Contour lines are drawn every $2k_BT$. The labels A_1 and B_1 indicate the positions for the local-minimum states at $T=250$ K that originate from the reference configuration 1 and the reference configuration 2, respectively. The label C stands for the saddle point that corresponds to the transition state.

used. Figure 11 shows contour lines of the free energy reweighted to $T=250$ K, which is close to the folding temperature (46). Here, the free energy $F(r_1,r_2)$ is defined by

$$F(r_1,r_2)=-k_BT\ln P(r_1,r_2), \tag{48}$$

where r_1 and r_2 are the rms distances defined in Eq. (47) from the reference configuration 1 and the reference configuration 2, respectively, and $P(r_1,r_2)$ is the (reweighted) probability at $T=250$ K to find the peptide with values r_1,r_2. The probability was calculated from the two-dimensional histogram of bin size 0.06×0.06 Å^2. The contour lines were plotted every $2k_BT$ ($=0.99$ kcal/mol for $T=250$ K).

Note that the reference configurations 1 and 2, which are, respectively, located at $(r_1,r_2)=(0,4.95)$ and $(4.95,0)$, are not local minima in free energy at the finite temperature ($T=250$ K) because of the entropy contributions. The corresponding local-minimum states at A_1 and B_1 still have the characteristics of the reference configurations in that they have backbone hydrogen bonds between Gly-2 and Met-5 and between Tyr-1 and Phe-4, respectively. We remark that we observe in Fig. 11 another well-defined local-minimum state around $(r_1,r_2)=(4.7,3.5)$. This state can also be considered to correspond to configuration 2 because we again observe the backbone hydrogen bond between Tyr-1 and Phe-4. The side-chain structures are, however, more deviated from configuration 2 than B_1, resulting in a larger value of r_2.

The transition state C in Fig. 11 should have intermediate structure between configurations 1 and 2. In Fig. 12 we show a typical backbone structure of this transition state. We see the backbone hydrogen bond between Gly-2 and Phe-4. This is precisely the expected intermediate structure between configurations 1 and 2, because going from configuration 1 to configuration 2 we can follow the backbone hydrogen-bond rearrangements: The hydrogen bond between Gly-2 and Met-5 of configuration 1 is broken, Gly-2 forms a hydrogen bond with Phe-4 (the transition state), this new hydrogen bond is broken, and finally Phe-4 forms a hydrogen bond with Tyr-1 (configuration 2).

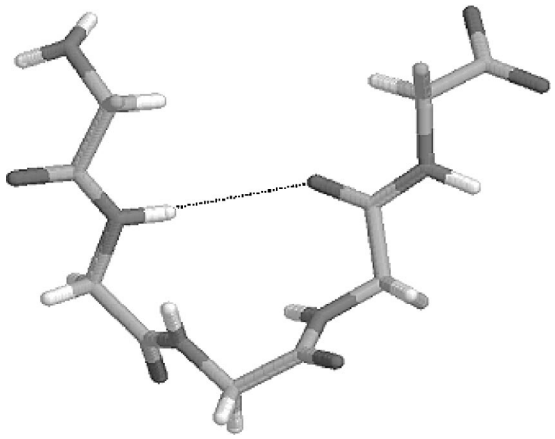

FIG. 12. The transition state between reference configurations 1 and 2. See the caption of Fig. 1 for details.

It is interesting to see in Fig. 11 that there is only one saddle point in the free-energy landscape that connects configurations 1 and 2. Hence, the transition between configurations 1 and 2 always passes through the state C.

In Ref. [22] the low-energy conformations of Met-enkephalin were studied in detail and they were classified into several groups of similar structures based on the pattern of backbone hydrogen bonds. It was found there that below $T=300$ K there are two dominant groups, which correspond to configurations 1 and 2 in the present paper. Although much less conspicuous, the third most populated structure is indeed the group that is identified to be the transition state in the present work.

In Figs. 13 and 14 we show the internal energy landscape and the entropy landscape at $T=250$ K, respectively. Here, the internal energy U is defined by the (reweighted) average ECEPP/2 potential energy:

$$U(r_1,r_2)=\langle E(r_1,r_2)\rangle. \tag{49}$$

Here, the average was again calculated from the two-dimensional histogram of bin size 0.06×0.06 Å^2. The entropy S was then calculated by

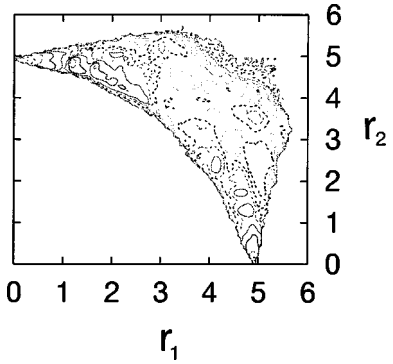

FIG. 13. Internal energy landscape at $T=250$ K with respect to rms distances (Å) from the two reference configurations, $U(r_1,r_2)$. Contour lines are drawn every $2k_BT$.

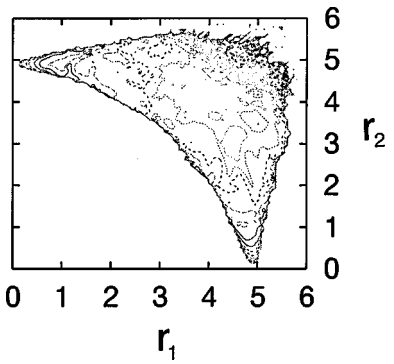

FIG. 14. Entropy landscape at $T=250$ K with respect to rms distances (Å) from the two reference configurations, $-TS(r_1,r_2)$. Contour lines are drawn every $2k_BT$.

$$S(r_1,r_2)=\frac{1}{T}[U(r_1,r_2)-F(r_1,r_2)]. \tag{50}$$

The landscape in Fig. 14 is actually $-TS(r_1,r_2)$.

Both internal energy and entropy landscapes are more rugged than free-energy landscape (we observe much more number of contour lines in Figs. 13 and 14 than in Fig. 11). The internal energy has clear local minima at the points $(r_1,r_2)=(0,4.95)$ and $(4.95,0)$, which, respectively, correspond to configurations 1 and 2, while the entropy landscape has local maxima at these points. These two terms tend to cancel each other, and the free-energy landscape is smoothed out.

In Table IV we list the numerical values of the free energy, internal energy, and entropy multiplied by temperature at the two local-minimum states (A_1 and B_1 in Fig. 11) and the transition state (C in Fig. 11). The internal energy is just the average of the ECEPP/2 potential energy (without any shift of zero point). The free energy was normalized so that the value at A_1 is zero. The values at the coordinates of reference configurations 1 and 2, which are, respectively, referred to as A_0 and B_0 in the table, are also listed.

Among the five points, A_0 and B_0 are unfavored in free energy mainly due to the large entropy effects, although they are energetically most favored. This means that at this temperature the exact conformations of the reference configurations 1 and 2 are not populated much. The relevant states are rather A_1, B_1, and C. The state A_1 can be considered to be ''deformed'' configuration 1, and B_1 deformed configuration 2 due to the entropy effects, whereas C is the transition state between A_1 and B_1. Among these three points, the free energy F and the internal energy U are the lowest at A_1, while the entropy contribution $-TS$ is the lowest at C. The free energy difference ΔF, internal energy difference ΔU, and entropy contribution difference $-T\Delta S$ are 1.0 kcal/mol, 1.9 kcal/mol, and -0.9 kcal/mol between B_1 and A_1, 2.2 kcal/mol, 4.6 kcal/mol, and -2.4 kcal/mol between C and A_1, and 1.2 kcal/mol, 2.7 kcal/mol, and -1.5 kcal/mol between C and B_1. Hence, the internal energy contribution and the entropy contribution to free-energy are opposite in sign and the magnitude of the former is roughly twice as that of the latter at this temperature.

TABLE IV. Free energy, internal energy, entropy multiplied by temperature at $T=250$ K (all in kcal/mol) at the two local-minimum states A_1 and B_1 and the transition state C in Fig. 11. The values at the coordinates of reference configurations 1 and 2, which are, respectively, referred to as A_0 and B_0, are also listed. The rms distances are in angstroms.

Coordinate (r_1,r_2)	F	U	$-TS$
A_1 (1.23, 4.83)	0	-5.4	5.4
B_1 (4.17, 2.43)	1.0	-3.5	4.5
C (3.09, 4.05)	2.2	-0.8	3.0
A_0 (0.03, 4.95)	15	-10.5	26
B_0 (4.95, 0.03)	20	-8.1	28

IV. SUMMARY AND CONCLUSIONS

We have outlined an approach to perform MC simulations which yield the free-energy distribution between two reference configurations. The multioverlap weights for this purpose were obtained by a novel, iterative process. The main point of this iterative process is not that it is supposed to be more efficient than the recursion that was used in the multi-self-overlap simulations of Ref. [12], but that it is an entirely independent approach, which starts from an analytically controlled limit. Recursions such as the one used in Ref. [12] are not ''foolproof.'' For instance, while most of the spin glass replica in Ref. [12] were well behaved, a few did not complete their recursion after more than an entire year of single processor CPU time. Similar situations could be encountered in all-atom simulations of larger peptides, where the normal multicanonical weight recursion as well as similar multioverlap weight recursion could fail. The present method provides then an alternative, approaching the physical region from a different limit.

Noticeable, our multioverlap approach is well-suited to be combined with a recently introduced, biased Metropolis sampling [30]. Namely, the required configurations at higher temperatures are as well necessary for our particular multioverlap recursion, so that no extra simulations are required in this respect.

On the physical side, we have found that entropy effects are rather important for a small peptide. The effects of entropy on the folding of real proteins in realistic solvent have yet to be studied in detail.

We have also performed the analysis of this paper for Met-enkephalin with variable ω angles and, in particular, simulated with combined weights at a number of temperatures. The results found are quite similar to those reported in this paper. In future work we intend to analyze the transition between reference configuration for larger systems of actual interest such as β-lactoglobulin.

ACKNOWLEDGMENTS

We are grateful for the financial support from the Joint Studies Program of the Institute for Molecular Science (IMS). One of the authors (B.B.) would like to thank the

IMS faculty and staff for their kind hospitality during his stay. In part, this work was supported by grants from the U.S. Department of Energy under Contract No. DE-FG02-97ER40608 (for B.A.B.), the Japan Society for the Promotion of Science for Young Scientists (for H.N.), and from the Research for the Future Program of the Japan Society for the Promotion of Science (Grant No. JSPS-RFTF98P01101) (for Y.O.).

[1] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[2] U.H. Hansmann and Y. Okamoto, in *Annual Reviews of Computational Physics VI*, edited by D. Stauffer (World Scientific, Singapore, 1999), p. 129.
[3] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[4] B.A. Berg, Comput. Phys. Commun. **104**, 52 (2002).
[5] G.M. Torrie and J.P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[6] B.A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).
[7] B.A. Berg and T. Celik, Phys. Rev. Lett. **69**, 2292 (1992).
[8] B.A. Berg, U.H. Hansmann, and T. Neuhaus, Phys. Rev. B **47**, 497 (1993).
[9] B.A. Berg, J. Stat. Phys. **82**, 323 (1996).
[10] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).
[11] F. Wang and D.P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[12] B.A. Berg, A. Billoire, and W. Janke, Phys. Rev. B **61**, 12143 (2000).
[13] K. Kuwajima, H. Yamaya, S. Miwa, S. Sugai, and T. Nagamura, FEBS Lett. **221**, 115 (1987).
[14] D. Hamada, S. Segawa, and S. Goto, Nat. Struct. Biol. **3**, 868 (1996).
[15] M.J. Sippl, G. Némethy, and H.A. Scheraga, J. Phys. Chem. **88**, 6231 (1984), and references given therein.
[16] F. Eisenmenger, U.H. Hansmann, S. Hayryan, and C.-K. Hu, Comput. Phys. Commun. **138**, 192 (2001).
[17] Z. Li and H.A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[18] B. von Freyberg and W.J. Braun, J. Comput. Chem. **12**, 1065 (1991).
[19] Y. Okamoto, T. Kikuchi, and H. Kawai, Chem. Lett. **1992**, 1275 (1992).
[20] U.H. Hansmann and Y. Okamoto, J. Comput. Chem. **14**, 1333 (1993).
[21] H. Meirovitch, E. Meirovitch, A.G. Michel, and M. Vásquez, J. Phys. Chem. **98**, 6241 (1994).
[22] A. Mitsutake, U.H. Hansmann, and Y. Okamoto, J. Mol. Graphics Modell. **16**, 226 (1998).
[23] R.A. Sayle and E.J. Milner-White, Trends Biochem. Sci. **20**, 374 (1995).
[24] U.H. Hansmann, M. Masuya, and Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. **94**, 10 652 (1997).
[25] B. Hesselbo and R. Stinchcombe, Phys. Rev. Lett. **74**, 2151 (1995).
[26] B.A. Berg (unpublished).
[27] E.J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).
[28] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes in Fortran*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[29] U.H. Hansmann, Y. Okamoto, and J.N. Onuchic, Proteins: Struct., Funct., Genet. **34**, 472 (1999).
[30] B.A. Berg, Phys. Rev. Lett. **90**, 180601 (2003).