Thermodynamics of lattice heteropolymers

Michael Bachmann[a) and Wolfhard Janke[b)
Institut für Theoretische Physik, Universität Leipzig, Augustusplatz 10/11, D-04109 Leipzig, Germany

We calculate thermodynamic quantities of hydrophobic-polar (HP) lattice proteins by means of a multicanonical chain-growth algorithm that connects the new variants of the Pruned-Enriched Rosenbluth Method and flat histogram sampling of the entire energy space. Since our method directly simulates the density of states, we obtain results for thermodynamic quantities of the system for all temperatures. In particular, this algorithm enables us to accurately simulate the usually difficult accessible low-temperature region. Therefore, it becomes possible to perform detailed analyses of the low-temperature transition between ground states and compact globules. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1651055]

I. INTRODUCTION

The native conformation of a protein is strongly correlated with the sequence of amino acid residues building up the heteropolymer. The sequence makes the protein unique and assigns it a specific function within a biological organism. The reason is that the different types of amino acids vary in their response to the environment and in their mutual interaction. It is a challenging task to reveal on what general principles the folding process of a protein is based. Models differ extremely in their level of abstraction, ranging from simple and purely qualitative lattice models to highly sophisticated all-atom off-lattice formulations with explicit solvent that partially yield results comparable with experimental data. Due to the enormous computational effort required for simulations of realistic proteins, usually characteristic properties of a protein with a given sequence are studied in detail. Much simpler, but by no means trivial, lattice models enjoy a growing interest, since they allow a more global view on, for example, the analysis of the relation between sequence and structure.

In this paper, we shall focus ourselves on thermodynamic properties of lattice proteins at all temperatures. In particular, this includes the investigation of the transitions between the different classes of states: lowest-energy states, compact globules, and random coils. Since the ground-state-globule transition occurs at rather low temperatures, a powerful algorithm is required that in particular allows a reasonable sampling of the low-lying energy states. To this end we combined multicanonical strategies[1–3] with chain-growth algorithms[4–7] to a new method[8] which works temperature-independent and directly simulates the density of states. This quantity contains all energetic information necessary for computing the mean energy, free energy, entropy, and specific heat for all temperatures. In the following, we present results obtained from the application of this method to different lattice proteins with lengths up to 103 monomers,

modeled by the simplest lattice formulation for heteropolymers, the hydrophobic-polar (HP) model.[9] In this model, only two types of monomers enter, hydrophobic (H) and polar (P) residues. The model is based on the assumption that the hydrophobic interaction is one of the fundamental principles in protein folding. An attractive hydrophobic interaction provides for the formation of a compact hydrophobic core that is screened from the aqueous environment by a shell of polar residues. Therefore the energy function reads

$$E = - \sum_{\langle i,j < i-1 \rangle} \sigma_i \sigma_j, \qquad (1)$$

where $\langle i,j < i-1 \rangle$ denotes summation over nearest lattice neighbors that are nonadjacent along the self-avoiding chain of monomers. A hydrophobic monomer at the ith position in the chain has $\sigma_i = 1$ and a polar monomer is assigned $\sigma_i = 0$.

The first part of this paper, where we explain how our new method works, will be of a more technical nature, while the second part is devoted to applications of this algorithm to heteropolymers. In Sec. II we discuss the main differences of Monte Carlo methods based on move sets for updating conformations and chain-growth algorithms based on the Pruned-Enriched Rosenbluth Method (PERM) as well as their peculiarities. This is followed by Sec. III about the thermodynamic quantities that will be estimated with our method. Then, in Sec. IV, we enter into the description of the multicanonical chain-growth algorithm. This technical part is preluded by recalling the essential ingredients of the original[4,5] PERM and the recently proposed improved variants[6,7] nPERM$_{\mathrm{is}}^{\mathrm{ss}}$, respectively, as these are fundamental for setting up our algorithm. Then we proceed with the explanation of multicanonical chain growth and the determination of the multicanonical weight factors. Section V is devoted to the validation of our method, and in Sec. VI we present the results obtained with our algorithm. There we focus on thermodynamic properties of heteropolymers with sequences of more than 40 monomers. Finally, we summarize the main aspects of this paper in Sec. VII.

[a)Electronic mail: michael.bachmann@itp.uni-leipzig.de
[b)Electronic mail: wolfhard.janke@itp.uni-leipzig.de; homepage: http://www.physik.uni-leipzig.de/CQT

II. MOVE SETS VERSUS CHAIN GROWTH

Polymers fold on a lattice into conformations that are by definition self-avoiding. This takes into account the finite volume and the uniqueness of the monomers. A lattice site can hence be occupied by a single monomer only. This has the consequence that the number of very dense conformations of a polymer is by orders of magnitude lower than that of random coil states. In Monte Carlo simulations, particular attention must therefore be devoted to efficient update procedures which also allow the sampling of dense conformations. In polymer simulations, so-called move sets were applied with some success to study the behavior near the Θ point, which denotes the phase transition, where polymers subject to an attractive interaction collapse from random coils to compact conformations. Move sets being widely used usually consist of transformations that change the position of a single monomer and a single bond vector (end flips), a single monomer position but two bonds (corner flips), two positions and three bonds (crankshaft) or moves with more changes, and pivot rotations, where the ith monomer serves as pivot point and one of the two partial chains connected with it is rotated about any axis through the pivot.[10] For (a-thermal) self-avoiding walks the latter method is known to be very efficient.[11] It becomes inefficient, however, the more dense the conformation is. At low temperatures, the acceptance rate of locally changing a dense conformation decreases drastically and the simulation threatens to get stuck in a specific conformation or to oscillate between two states. Since the search for ground states is an essential aspect of studying lattice proteins, the application of move sets is not very useful, at least for chains of reasonable length.

A more promising alternative is the completely different approach based on chain growth. The polymer grows by attaching the nth monomer at a randomly chosen next-neighbor site of the $(n-1)$th monomer. The growth is stopped, if the total length N of the chain is reached or the randomly selected continuation of the chain is already occupied. In both cases, the next chain is started to grow from the first monomer. This simple chain growth is also not yet very efficient, since the number of discarded chains grows exponentially with the chain length. The performance can be improved with the Rosenbluth chain-growth method,[12] where first the free next neighbors of the $(n-1)$th monomer are determined and then the new monomer is placed to one of the unoccupied sites. Since the probability of each possibility for the next monomer to be set varies with the number of free neighbors, this implies a bias given by

$$p_n \sim \left(\prod_{l=2}^{n} m_l \right)^{-1}, \qquad (2)$$

where m_l is the number of free neighbors to place the lth monomer. The bias is corrected by assigning a Rosenbluth weight factor $W_n^R \sim p_n^{-1}$ to each chain that has been generated by this procedure. Nevertheless, this method suffers from attrition too: If all next neighbors are occupied, i.e., the

chain was running into a "dead end" (attrition point), the complete chain has to be discarded and the growth process has to be started anew.

Combining the Rosenbluth chain-growth method with population control, however, as is done in PERM,[4,5] leads to a further considerable improvement of the efficiency by increasing the number of successfully generated chains. This method is particularly useful for studying the Θ point of polymers, since the Rosenbluth weights of the statistically relevant chains approximately cancel against their Boltzmann probability. The (a-thermal) Rosenbluth weight factor W_n^R is therefore replaced by

$$W_n^{\text{PERM}} = \prod_{l=2}^{n} m_l e^{-(E_t - E_{l-1})/k_B T},$$
$$\qquad (3)$$
$$2 \leq n \leq N \quad (E_1 = 0, \ W_1^{\text{PERM}} = 1),$$

where T is the temperature and E_l is the energy of the partial chain $\mathbf{X}_l = (\mathbf{x}_1, \ldots, \mathbf{x}_l)$ created with Rosenbluth chain growth. In PERM, population control works as follows. If a chain has reached length n, its weight W_n^{PERM} is calculated and compared with suitably chosen upper and lower threshold values, $W_n^>$ and $W_n^<$, respectively. For $W_n^{\text{PERM}} > W_n^>$, *identical* copies are created which then grow independently. The weight is equally divided among them. If $W_n^{\text{PERM}} < W_n^<$, the chain is pruned with some probability, say 1/2, and in case of survival, its weight is doubled. For a value of the weight lying between the thresholds, the chain is simply continued without enriching or pruning the sample.

In the recently developed new variants[6,7] nPERM$_{\text{is}}^{\text{ss}}$, the number of copies is not constant and depends on the ratio of the weight W_n^{PERM} compared to the upper threshold value $W_n^>$ and the copies are necessarily chosen to be different [the method of selecting the copies is based on simple sampling (ss) in nPERMss and a kind of importance sampling (is) in nPERMis]. This proves quite useful in producing highly compact polymers and therefore these new methods are very powerful in determining lowest-energy states of lattice proteins.

III. DENSITY OF STATES AND THERMODYNAMIC QUANTITIES

In order to investigate the thermodynamic properties of lattice proteins accurate simulations are necessary. Due to the difficulties with the update of conformations at low temperatures and a primary interest in detecting lowest-energy conformations, only a few results of thermodynamic quantities are found in the literature. Nevertheless, the understanding of the conformational transitions and the dependence of their sharpness on the sequence is only possible with algorithms that yield good results for low as well as for high temperatures. Consequently, reasonable results can only be obtained, if the method allows the sampling of the entire energy space.

All energetic statistical quantities of the protein can be expressed by means of the density of energetic states $g(E)$, since the partition function of a lattice protein with given sequence can be written as

$$Z = \sum_{\{\mathbf{x}\}} e^{-\beta E(\{\mathbf{x}\})} = \sum_i g(E_i) e^{-\beta E_i}, \tag{4}$$

where $\beta = 1/T$ is the inverse of the thermal energy in natural units. The sum in the first representation runs over all possible realizations of self-avoiding walks on the lattice, while in the second expression the sum is taken over all energetic states a lattice protein can adopt. Then, the expectation value of any energetic observable $O(E)$ is simply

$$\langle O(E) \rangle (T) = \frac{1}{Z} \sum_i O(E_i) g(E_i) e^{-\beta E_i}, \tag{5}$$

and the mean energy as the negative logarithmic derivative of Z with respect to β is given by

$$\langle E \rangle (T) = \frac{1}{Z} \sum_i E_i g(E_i) e^{-\beta E_i}. \tag{6}$$

With these expressions, the specific heat $C_V = d\langle E \rangle / dT$ obeys the fluctuation formula

$$C_V(T) = \frac{1}{T^2} (\langle E^2 \rangle - \langle E \rangle^2). \tag{7}$$

Moreover, knowing $g(E)$, the Helmholtz free energy is obtained from

$$F(T) = -T \ln \sum_i g(E_i) e^{-\beta E_i} \tag{8}$$

and the entropy can be calculated as

$$S(T) = \frac{1}{T} [\langle E \rangle (T) - F(T)]. \tag{9}$$

In addition, nonenergetic structural quantities are of interest for discussing the compactness of conformations, such as the end-to-end distance

$$R_{ee} = |\mathbf{x}_N - \mathbf{x}_1| \tag{10}$$

and the radius of gyration

$$R_{gyr} = \sqrt{\frac{1}{N} \sum_{l=1}^N (\mathbf{x}_l - \mathbf{x}_0)^2}, \tag{11}$$

where $\mathbf{x}_0 = \sum_{l=1}^N \mathbf{x}_l / N$ is the center of mass of the conformation (with all monomers having equal mass). For the calculation of mean values of nonenergetic quantities O, Eq. (5) is replaced by the general formula

$$\langle O \rangle (T) = \frac{1}{Z} \sum_{\{\mathbf{x}\}} O(\{\mathbf{x}\}) e^{-\beta E(\{\mathbf{x}\})}. \tag{12}$$

In the following, we shall focus on the study of these thermodynamic quantities for different HP lattice proteins and develop an algorithm that allows a direct simulation of the density of states.

IV. MULTICANONICAL CHAIN-GROWTH ALGORITHM

Before we describe the idea behind the multicanonical chain-growth algorithm and the iterative determination of the multicanonical weights which are related to the density of states, we first recall the canonical chain growth variants nPERM_{is}^{ss}, on which our algorithm builds up.

A. Canonical chain growth

In the original chain-growth algorithm PERM,[4] the sample of chains with length $n < N$ is enriched by identical copies if the weight factor (3) is bigger than a threshold value $W_n^>$. In order to obey correct Boltzmann statistics, the weight is divided among the clones. If, however, $W_n^{\text{PERM}} < W_n^<$, the chain is pruned with, e.g., probability 1/2, requiring the weight of a surviving chain to be taken twice. Then one attaches a new monomer at a randomly chosen free next-neighbor site of the previous one. This is done for all chains that still exist and the procedure is repeated until the total chain length N is reached or the growth of a chain was terminated by a dead end or due to pruning. After all chains created within this tree have grown until their end is reached and thus the present so-called *tour* is finished, a new growth process starts from the first monomer, i.e., a new tour begins. Having created an appropriate number of chains with length N, they will be canonically distributed at the given temperature T. In fact, this is also true for all partial chains with intermediate lengths $n < N$, but there are strong correlations between chains with different lengths n.

In the recently proposed new PERM variants nPERM_{is}^{ss} [new PERM with simple/importance sampling (ss/is)] a considerable improvement is achieved by creating different copies, i.e., the chains are identical in $(n-1)$ monomers but have different continuations, instead of completely identical ones, since identical partial chains usually show a similar evolution. Because of the different continuations, the weights of the copies can differ. Therefore it is not possible to decide about the number of copies on the basis of a joint weight. The suggestion is to calculate first a predicted weight which is then compared with the upper threshold $W_n^>$ in order to determine the number of clones. Another improvement of PERM being followed up since first applications to lattice proteins is that the threshold values $W_n^>$ and $W_n^<$ are no longer constants, but are dynamically adapted with regard to the present estimate for the partition sum and to the number of successfully created chains with length n. The partition sum is proportional to the sum over the weight factors of all conformations of chains with length n, created with a Rosenbluth chain-growth method like, for instance, nPERM_{is}^{ss}:

$$Z_n = \frac{1}{M_{\text{tours}}} \sum_i W_n^{\text{nPERM}_{is}^{ss}}(\mathbf{X}_{n,t}). \tag{13}$$

Here, $\mathbf{X}_{n,t}$ denotes the tth generated conformation of length n. The proportionality constant is the inverse of the number of chain growth starts M_{tours}. Note that due to this normalization it is possible to estimate the degeneracy of the energy states. This is in striking contrast to importance sampling Monte Carlo methods, where the overall constant on the right-hand side of Eq. (13) cannot be determined and hence only *relative* degeneracies can be estimated.

Since nPERMss and nPERMis, respectively, are possible fundamental ingredients for our algorithm, it is useful to recall in some detail how these chain-growth algorithms work. The main difference in comparison with the original PERM

is that, if the sample of chains of length $n-1$ shall be enriched, the continuations to an unoccupied next-neighbor site have to be different, i.e., the weights of these chains with length n can differ. Therefore it is impossible to calculate a uniform weight like W_n^{PERM} as given in Eq. (3) *before* deciding whether to enrich, to prune, or simply to continue the current chain of length $n-1$. As proposed in Ref. 6, it is therefore useful to control the population on the basis of a predicted weight W_n^{pred} which is introduced as

$$W_n^{\mathrm{pred}} = W_{n-1}^{\mathrm{nPERM_{is}^{ss}}} \sum_{\alpha=1}^{m_n} \chi_\alpha^{\mathrm{nPERM_{is}^{ss}}}, \qquad (14)$$

where m_n denotes the number of free neighboring sites to continue with the nth monomer. The "importances" $\chi_\alpha^{\mathrm{nPERM_{is}^{ss}}}$ differ for nPERMss and nPERMis. Due to its characterization as a simple sampling algorithm (nPERMss), where all continuations are equally probable, and as a method with importance sampling (nPERMis), the importances may be defined as

$$\chi_\alpha^{\mathrm{nPERMss}} = 1, \qquad (15)$$

$$\chi_\alpha^{\mathrm{nPERMis}} = (m_n^{(\alpha)} + \tfrac{1}{2}) e^{-\beta(E_n^{(\alpha)} - E_{n-1})}. \qquad (16)$$

The expression for nPERMis involves the energy $E_n^{(\alpha)}$ of the choice $\alpha \in [1, m_n]$ for placing the nth monomer and the number of free neighbors $m_n^{(\alpha)}$ of this choice which is identical to m_{n+1}, provided the αth continuation was indeed selected for placing the nth monomer. Since $\chi_\alpha^{\mathrm{nPERMis}}$ contains informations beyond the nth continuation of the chain, nPERMis controls the further growth better than nPERMss. The predicted weight for the nth monomer is now used to decide how the growth of the chain is continued. If the predicted weight is bigger than the current threshold, $W_n^{\mathrm{pred}} > W_n^>$, and $m_n > 1$, the sample of chains is enriched and the number of copies k is determined according to the empirical rule $k = \min[m_n, \mathrm{int}(W_n^{\mathrm{pred}}/W_n^>)]$. Thus, $2 \leq k \leq m_n$ different continuations will be followed up. Using nPERMss, the k continuations are chosen randomly with equal probability among the m_n possibilities, while for nPERMis the probability of selecting a certain k-tuple $A = \{\alpha_1, \ldots, \alpha_k\}$ of different continuations is given by

$$p_A = \frac{\sum_{\alpha \in A} \chi_\alpha}{\sum_A \sum_{\alpha \in A} \chi_\alpha}. \qquad (17)$$

Considering the probabilities p_A as partial intervals of certain length, arranging them successively in the total interval $[0,1]$ (since $\sum_A p_A = 1$), and drawing a random number $r \in [0,1)$, one selects the tuple whose interval contains r. This tuple of different sites is then chosen to continue the chain. The corresponding weights are[6]

$$W_{n,\alpha_j}^{\mathrm{nPERM_{is}^{ss}}} = W_{n-1}^{\mathrm{nPERM_{is}^{ss}}} \frac{m_n}{k \binom{m_n}{k} p_A} e^{-\beta(E_n^{(\alpha_j)} - E_{n-1})}, \qquad (18)$$

where $j \in \{1, \ldots, k\}$ is the index of the α_jth continuation within the tuple A. In the special case of simple sampling this expression reduces to $W_{n,\alpha_j}^{\mathrm{nPERMss}} = W_{n-1}^{\mathrm{nPERMss}} m_n \exp[-\beta(E_n^{(\alpha_j)}$

$- E_{n-1})]/k$. If the predicted weight is less than the lower threshold, $W_n^{\mathrm{pred}} < W_n^<$, however, the growth of this chain is stopped with probability 1/2. In this case, one traces the chain back to the last branching point, where the growth can be continued again, or, if there are no branching points, a new tour is started. If the chain survives, the continuation of the chain follows the same procedure as described above, but now with $k = 1$, where Eq. (17) simplifies to $p_A = \chi_\alpha / \sum_\alpha \chi_\alpha$ since $A = \{\alpha\}$. In this case the weight of the chain is taken twice. For $W_n^< \leq W_n^{\mathrm{pred}} \leq W_n^>$, the chain is continued without enriching or pruning (once more with $k = 1$).

The first tour, where the nth monomer is attached for the first time, is started with bounds set to $W_n^> = \infty$ and $W_n^< = 0$, thus avoiding enrichment and pruning. For the following tours, we use[13]

$$W_n^> = C \frac{Z_n}{Z_1} \frac{c_n^2}{c_1^2}, \qquad (19)$$

where c_i is the number of created chains with i monomers. The constant $C \leq 1$ is some positive number and controls the number of successfully generated chains per tour. For the lower bound we use $W_n^< = 0.2 W_n^>$. All these choices are in correspondence to Ref. 6.

B. Multicanonical sampling of Rosenbluth-weighted chains

The idea behind our multicanonical chain-growth method is to flatten the canonical energy distribution provided by $\mathrm{nPERM_{is}^{ss}}$. For a given temperature, the latter algorithms yield accurate canonical distributions over some orders of magnitude. In order to construct the entire density of states, standard reweighting procedures may be applied, requiring simulations for different temperatures.[3] The low-temperature distributions are, however, very sensitive against fluctuations of weights which inevitably occur because the number of energetic states is low, but the weights are high. Thus, it is difficult to obtain a correct distribution of energetic states, since this requires a reasonable number of hits of low-energy states. Therefore we assign the chains an additional weight, the multicanonical weight factor W_n^{flat}, chosen such that all possible energetic states of a chain of length n possess almost equal probability of realization. The first advantage is that states having a low Boltzmann probability compared to others are hit more frequently. Second, the multicanonical weights introduced in that manner are proportional to the inverse canonical distribution at temperature T, $W_n^{\mathrm{flat}}(E) \sim 1/P_n^{\mathrm{can},T}(E)$, with respect to the inverse density of states

$$W_n^{\mathrm{flat}}(E) \sim g_n^{-1}(E) \qquad (20)$$

for $T \to \infty$. Thus, only one simulation is required and a multi-histogram reweighting is not necessary. An important conceptual aspect is the fact that the multicanonical weight factors are unknown in the beginning and have to be determined iteratively.

Before we discuss the technical aspects regarding our method, we first explain it more formally. The energy-

dependent multicanonical weights are trivially introduced into the partition sum (13) as suitable "decomposition of unity" in the following way:

$$Z_n = \frac{1}{M_{\text{tours}}} \sum_t W_n^{\text{nPERM}_{\text{is}}^{\text{ss}}}(\mathbf{X}_{n,t}) W_n^{\text{flat}}(E(\mathbf{X}_{n,t}))$$
$$\times [W_n^{\text{flat}}(E(\mathbf{X}_{n,t}))]^{-1}. \qquad (21)$$

Since we are going to simulate at infinite temperature, we express with Eq. (20) the partition sum which then coincides with the total number of all possible conformations as

$$Z_n = \frac{1}{M_{\text{tours}}} \sum_t g_n(E(\mathbf{X}_{n,t})) W_n(\mathbf{X}_{n,t}) \qquad (22)$$

with the combined weight

$$W_n(\mathbf{X}_{n,t}) = W_n^{\text{nPERM}_{\text{is}}^{\text{ss}}}(\mathbf{X}_{n,t}) W_n^{\text{flat}}(E(\mathbf{X}_{n,t})). \qquad (23)$$

Taking this as the probability for generating chains of length n, $p_n \sim W_n$, leads to the desired flat distribution $H_n(E)$, from which the density of states is obtained by

$$g_n(E) \sim \frac{H_n(E)}{W_n^{\text{flat}}(E)}. \qquad (24)$$

The canonical distribution at *any* temperature T is calculated by simply reweighting the density of states to this temperature, $P_n^{\text{can},T}(E) \sim g_n(E) \exp(-E/T)$.

C. Iterative determination of the density of states

In the following, we describe our procedure for the iterative determination of the multicanonical weights, from which we obtain an estimate for the density of states. Since there are no informations about an appropriate choice for the multicanonical weights in the beginning, we set them in the zeroth iteration for all chains $2 \leq n \leq N$ and energies E equal to unity, $W_n^{\text{flat},(0)}(E) = 1$, and the histograms to be flattened are initialized with $H_n^{(0)}(E) = 0$. These assumptions render the zeroth iteration a pure nPERM$_{\text{is}}^{\text{ss}}$ run.

Since we set $\beta = 0$ from the beginning, the accumulated histogram of all generated chains of length n,

$$H_n^{(0)}(E) = \sum_t W_{n,t}^{\text{nPERM}_{\text{is}}^{\text{ss}}} \delta_{E_t E}, \qquad (25)$$

is a first estimate of the density of states. In order to obtain a *flat* histogram in the next iteration, we update the multicanonical weights

$$W_n^{\text{flat},(1)}(E) = \frac{W_n^{\text{flat},(0)}(E)}{H_n^{(0)}(E)} \quad \forall n, E \qquad (26)$$

and reset the histogram, $H_n^{(1)}(E) = 0$.

The first and all following iterations are multicanonical chain growth runs and proceed along similar lines as described earlier, with some modifications. The prediction for the new weight again follows Eq. (14), but the importances χ_α^{is} (15) are in the ith iteration introduced as

$$\chi_\alpha^{\text{is},(i)} = \left(m_n^{(\alpha)} + \frac{1}{2} \right) \frac{W_n^{\text{flat},(i)}(E_n^{(\alpha)})}{W_{n-1}^{\text{flat},(i)}(E_{n-1})}. \qquad (27)$$

In the simple sampling case, we still have $\chi_\alpha^{\text{ss},(i)} = 1$. If the sample is enriched ($W_n^{\text{pred}} > W_n^>$) the weight (18) of a chain with length n choosing the α_jth continuation is now replaced by

$$W_{n,\alpha_j}^{\text{ss,is}} = W_{n-1}^{\text{ss,is}} \frac{m_n}{k \binom{m_n}{k} p_A} \frac{W_n^{\text{flat},(i)}(E_n^{(\alpha_j)})}{W_{n-1}^{\text{flat},(i)}(E_{n-1})}, \qquad (28)$$

where in the simple sampling case (ss) p_A and the binomial factor again cancel each other. If $W_n^< \leq W_n^{\text{pred}} \leq W_n^>$, an nth possible continuation is chosen (selected as described for the enrichment case, but with $k=1$) and the weight is as in Eq. (28). Assuming that $W_n^{\text{pred}} < W_n^<$ and that the chain has survived pruning (as usual with probability 1/2), we proceed as in the latter case and the chain is assigned twice that weight. The upper threshold value is now determined in analogy to Eq. (19) via

$$W_n^> = C \frac{Z_n^{\text{flat}}}{Z_1^{\text{flat}}} \frac{c_n^2}{c_1^2}, \qquad (29)$$

where

$$Z_n^{\text{flat}} = \frac{1}{M_{\text{tours}}} \sum_t W_{n,t}^{\text{ss,is}} \qquad (30)$$

is the estimated partition sum according to the new distribution provided by the weights (28) for chains with n monomers. Whenever a new iteration is started, Z_n^{flat}, c_n, $W_n^<$ are reset to zero, and $W_n^>$ to infinity (i.e., to the upper limit of the data type used to store this quantity). If a chain of length n with the energy E was created, the histogram is increased by its weight:

$$H_n^{(0)}(E) = \sum_t W_{n,t}^{\text{ss,is}} \delta_{E_t E}. \qquad (31)$$

From iteration to iteration, this histogram approaches the desired flat distribution $H_n(E)$ and after the final iteration $i = I$, the density of states is estimated by

$$g_n^{(I)}(E) = \frac{H_n^{(I)}(E)}{W_n^{\text{flat},(I)}(E)}, \quad 2 \leq n \leq N, \qquad (32)$$

in analogy to Eq. (24).

In our simulations, we typically performed up to 30 iterations. The runs 0 to $I-1$ were terminated after $10^5 - 10^6$ chains of total length N had been produced, while in the measuring run ($i = I$) usually $10^7 - 10^9$ conformations are sufficient to obtain reasonable statistics. The parameter C in Eq. (29) that controls the pruning/enrichment statistics and thus how many chains of complete length N are generated per tour was set to $C = 0.01$, such that on average ten complete chains were successfully constructed within each tour. With this choice, the probability for pruning the current chain or enriching the sample was about 20%. In almost all started tours M_{tours} at least one chain achieved its complete length. Thus the ratio between successfully finished tours M_{succ} and started tours M_{tours} is very close to unity, assuring that our algorithm performed with quite good efficiency.

Unlike typical applications of multicanonical or flat histogram algorithms in importance sampling schemes, where all energetic states become equally probable such that the dynamics of the simulation corresponds to a random walk in energy space, the distribution to be flattened in our case is the histogram that accumulates the weights of the conformations. Hence, if the histogram is flat, a small number of high-weighted conformations with low energy E has the same probability as a large number of appropriate conformations with energy $E' > E$ carrying usually lower weights. Therefore the number of actual low-energy hits remains lower than the number of hits of states with high energy. In order to accumulate enough statistics in the low-energy region, the comparatively large number of generated conformations in the measuring run is required.

We have also implemented multicanonical chain growth simulations, where we were going to flatten the "naked" energy distribution, i.e., we tried to equalize the number of hits for all energetic states. The problem is that this contradicts the philosophy of Rosenbluth chain-growth methods, where the bias connected with the Rosenbluth weight controls the population of samples. Therefore lowest-energy states were not "tuned" by this bias and not hit accordingly. For applications without special focus to the low-temperature region, it may be, however, an appropriate alternative to the above-described procedure and should be pursued further.

V. VALIDATION AND PERFORMANCE

Before we discuss the physical results obtained with the multicanonical chain-growth algorithm, we first remark on tests validating the method. We compared the specific heat for very short chains with data from exact enumeration and found that our method reproduces the exact results with high accuracy. For a chain with 42 monomers, where exact results are not available, we performed a multi-histogram reweighting[14] from canonical distributions at different temperatures obtained from original nPERMis runs. Here, it turned out that our method shows a considerably higher performance (higher accuracy in spite of lower statistics at comparable CPU times). We also compared with implementations based on sophisticated importance sampling Monte Carlo schemes, e.g., we have also performed multicanonical sampling[1,2] and Wang–Landau simulations[15] in combination with conformational updates different from chain growth (e.g., move sets as described in Sec. II). For the present applications, however, all of these attempts proved to be less efficient.

A. Comparison with results from exact enumeration

As a first validation of our method, we apply it to a set of 14mers with some interesting properties (see Table I) regarding the relation between their ground-state degeneracy and the strength of the low-temperature conformational transition between lowest-energy states and compact globules.[8,16] In finite-size systems, (pseudo-)transitions are usually identified through structural peculiarities (maxima for strong transitions or "shoulders" for weak transitions) in the temperature-dependent behavior of fluctuations of thermodynamic quantities. Usually, it is hard to obtain a quite accurate estimation

TABLE I. Sequences, hydrophobicity n_H, and global minimum energy E_{min} with degeneracy g_0^{ex} (without rotations, reflections, and translations) of the exactly enumerated 14mers used for validation of our algorithm. The last column contains the predictions for the ground-state degeneracy obtained with our method.

No.	Sequence	n_H	E_{min}	g_0^{ex}	g_0
14.1	$HPHPH_2PHPH_2P_2H$	8	-8	1	0.98 ± 0.03
14.2	$H_2P_2HPHPH_2PHPH$	8	-8	2	2.00 ± 0.07
14.3	$H_2PHPHP_2HPHPH_2$	8	-8	2	2.00 ± 0.06
14.4	$H_2PHP_2HPHPH_2PH$	8	-8	4	3.99 ± 0.13

of the fluctuations in the low-temperature region. Thus it is a good test of our method to calculate fluctuating quantities for the 14mers listed in Table I and to compare with results that are still available by exactly enumerating all possible 943 974 510 conformations (except translations).[16] Therefore we determined with our method the densities of states for these 14mers and calculated the fluctuations of the energy around the mean value in order to obtain the specific heat according to Eq. (7). We generated 10^9 chains and the results for the specific heat turned out to be highly accurate. This is demonstrated in Fig. 1, where we have plotted for the exemplified 14mers the relative errors $\varepsilon(T) = |C_V^{ex}(T) - C_V(T)|/C_V^{ex}(T)$ of our estimates $C_V(T)$ compared with the specific heats $C_V^{ex}(T)$ obtained by the exact enumeration procedure. We see that, except for very low temperatures, the relative error is uniformly smaller than 10^{-3}.

B. Multiple histogram reweighting

The calculation of the density of states by means of *canonical* stochastic algorithms cannot be achieved by simply reweighting *one* canonical histogram, obtained for a given temperature, to as many as necessary distributions to cover the whole temperature region, since the overlap between the sampled distribution and most of the reweighted histograms is too small.[3] As this simple reweighting only works in a certain region around the temperature the simulation was performed, a multiple application of the reweighting procedure at different sampling temperatures is necessary.[14] For the sequence of a 42mer to be studied in detail in Sec. VI A, we performed the multiple reweighting of five overlapping histograms obtained by separate nPERMis runs at tempera-
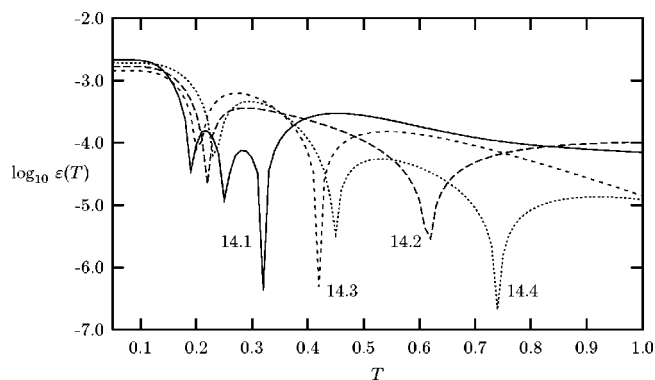

FIG. 1. Logarithmic plot of the relative errors of our estimates for the specific heats of the 14mers given in Table I.
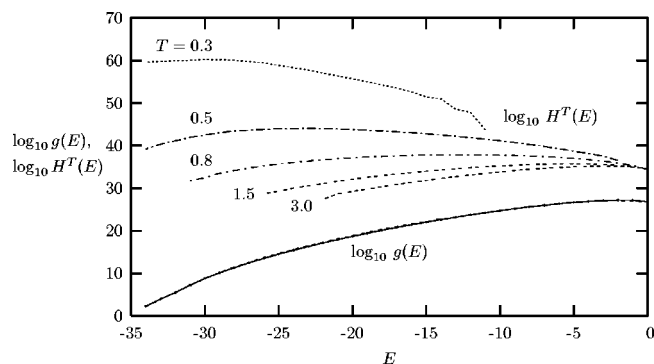
FIG. 2. Histograms $H^T(E)$ obtained by single nPERMis runs for five different temperatures $T=0.3$, 0.5, 0.8, 1.5, and 3.0 (dashed lines). The resulting density of states $g(E)$ obtained by multiple histogram reweighting (long dashed line) lies within the error bars of the density of states calculated by means of our method (solid line).

tures 0.3, 0.5, 0.8, 1.5, and 3.0 in order to estimate the density of states. The histograms as well as the resulting density of states are shown in Fig. 2, where we have also plotted the density of states being obtained by means of our multicanonical sampling algorithm. Each of the histograms contains statistics of 8×10^7 chains. This number was adequately chosen such that the density of states from histogram reweighting matches within the error bars of the density of states obtained with our algorithm that also inherently supplies us with the *absolute* density of states. Note that these absolute values cannot be obtained by means of the multiple-histogram reweighting procedure, where the normalization is initially arbitrary. Our density of states was obtained by accumulating statistics of 5×10^7 chains. This means that eight times more chains were necessary to approximately achieve the accuracy with the multiple histogram reweighting method. The iterative period for the determination of the multicanonical weights is no drawback, as it takes in our implementation only 10% compared to the production run. Therefore we conclude that our dynamical method is more efficient and also more elegant than a static reweighting scheme, where also a reliable estimation of statistical errors is extremely cumbersome.

C. Flat-histogram algorithms with update mechanisms different from chain growth

The calculation of the density of states for heteropolymers with less than 40 monomers does not represent a big challenge. It is still possible to combine generalized ensemble methods like multicanonical sampling[1,2] or the Wang–Landau method[15] with move sets including pivot rotations to yield reasonable statistics in the low-energy sector. In order to avoid the sophisticated implementation of the move sets, an alternative method can be used, for instance, where the conformational information is encoded in a string of letters denoting the directions F(orward), B(ackward), R(ight), L(eft), U(p), and D(own) the walker may follow in an embedded coordinate system. This structural conformation is then updated by simply changing a letter. All these methods require a time-consuming self-avoidance check following each update. We have tested these combinations of

sampling and update methods for 14mers, where we could compare with exact results from enumeration, and applied it to a 30mer with 20 energy states with rather high degeneracies. All these states were frequently hit such that the results were reasonable for all temperatures. Remarkably, it turned out, however, that the sampling of low-energy states becomes more problematic the lower the degeneracy of these states is. Either the algorithm got stuck after hitting such a state, or it took a long time to find it for the first time. These were indications for a "hidden" conformational barrier that could not be circumvented with these procedures.

Applying these methods to sequences with more than 40 monomers did not yield reliable results. Low-energy states were too rarely or never hit in long-term simulations. Performing a biased simulation by explicitly starting from a state with lowest energy, i.e., initializing with a very dense conformation, it took much too long until a new self-avoiding conformation was found and accepted. Comparing this with applications of the multicanonical chain-growth method to these examples led us to the conclusion that, in the application to lattice proteins, chain-growth methods are much more capable of avoiding such barriers.

VI. RESULTS

In the following we focus on results which we obtained with the multicanonical chain-growth algorithm for heteropolymers with HP sequences of more than 40 monomers.

A. Lattice model for parallel β helix with 42 monomers

We consider a 42mer with the sequence $PH_2\,PHPH_2\,PHPHP_2\,H_3\,PHPH_2\,PHPH_3\,P_2\,HPHPH_2\,PHPH_2\,P$ that forms a parallel helix in the ground state. It was designed to serve as a lattice model of the parallel β helix of *pectate lyase C*.[17] But there are additional properties that make it an interesting and challenging system to which we want to apply our method first. The ground-state energy is known to be $E_{\min}=-34$. Moreover, the specific heat has a very pronounced low-temperature peak that indicates a (pseudo-)transition between the lowest-energy states possessing compact hydrophobic cores and the regime of the globule conformations. This transition is in addition to the usual one between globules and random coils.

Figure 3 shows how the estimate for the density of states of this 42mer evolves with increasing number of iterations. The zeroth iteration is the initial pure nPERMis run at $\beta = 0$. This does not render, however, a proper image of the abilities of nPERMis, which works much better at finite temperatures. Iterations 1–8 are used to determine the multicanonical weights over the entire energy space $E \in [-34, 0]$. Then, the ninth iteration is the measuring run which gives a very accurate estimate for the density of states covering about 25 orders of magnitude. Our estimate for the ground-state degeneracy is $g_0 = 3.9 \pm 0.4$, which is in perfect agreement with the known value $g_0^{ex} = 4$ (except translational, rotational, and reflection symmetries).[18]

The average structural properties at finite temperatures can be best characterized by the mean end-to-end distance $\langle R_{ee} \rangle(T)$ and the mean radius of gyration $\langle R_{gyr} \rangle(T)$. As
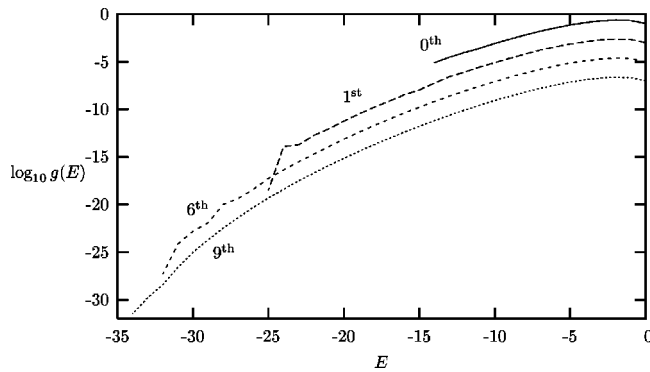
FIG. 3. Estimates for the density of states $g_{42}^{(i)}(E)$ for the 42mer after different levels of iteration. Since the curves would fall on top of each other, we have added, for better distinction, a suitable offset to the curves of the first, sixth, and ninth runs. The estimate of the zeroth run is normalized to unity.

FIG. 5. Specific heat C_V and derivatives with respect to temperature of mean end-to-end distance $\langle R_{ee} \rangle$ and radius of gyration $\langle R_{gyr} \rangle$ as functions of temperature for the 42mer. The ground-state–globule transition occurs between $T_0^{(1)} \approx 0.24$ and $T_0^{(2)} \approx 0.28$, while the globule–random coil transition takes place between $T_1^{(1)} \approx 0.53$ and $T_1^{(2)} \approx 0.70$ (shaded areas).

these quantities carry shape informations, their calculation is not exclusively based on the density of states $g_N(E(\mathbf{X}_{N,t}))$ and hence Eq. (5) cannot be applied. Therefore expectation values of such quantities O are obtained from the time series of the measuring run of the multicanonical chain growth simulation at infinite temperature by using the general formula

$$\langle O \rangle(T) = \frac{1}{Z_N} \sum_t O(\mathbf{X}_{N,t})$$
$$\times W_N(\mathbf{X}_{N,t}) g_N(E(\mathbf{X}_{N,t})) e^{-\beta E(\mathbf{X}_{N,t})}, \qquad (33)$$

with the estimate for the partition sum $Z_N = \sum_t W_N(\mathbf{X}_{N,t}) g_N(E(\mathbf{X}_{N,t})) \exp\{-\beta E(\mathbf{X}_{N,t})\}$. Our results for $\langle R_{ee} \rangle(T)$ and $\langle R_{gyr} \rangle(T)$ of the 42mer are shown in Fig. 4. Here and in the following figures, the statistical errors were estimated by using the jackknife binning method.[19] The pronounced minimum in the end-to-end distance can be interpreted as an indication of the transition between the lowest-energy states and globules: The low number of ground states have similar and highly symmetric shapes (due to the reflection symmetry of the sequence) but the ends of the chain are polar and therefore they are not required to reside close to each other. Increasing the temperature allows the protein to fold into conformations different from the ground states and contacts between the ends become more likely. Therefore,
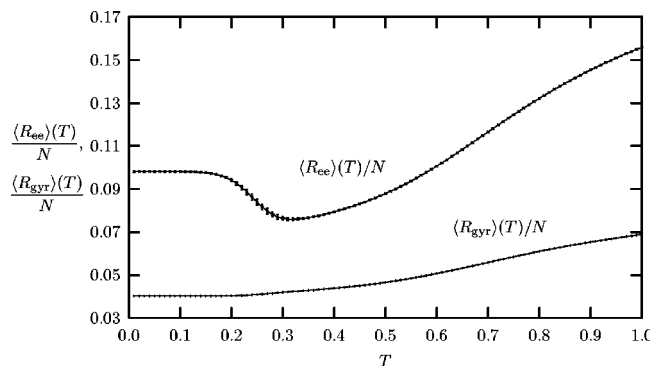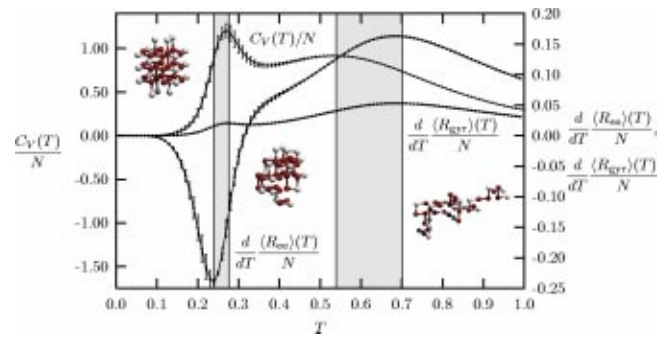
the mean end-to-end distance decreases and the protein has entered the globule "phase." Further increasing the temperature then leads to a disentangling of the globules and random coil conformations with larger end-to-end distances dominate. From Fig. 5, where we have plotted the specific heat and the derivatives of the mean end-to-end distance and of the mean radius of gyration with respect to the temperature,

$$\frac{d}{dT} \langle R_{ee} \rangle(T) = \frac{1}{T^2} (\langle E R_{ee} \rangle - \langle E \rangle \langle R_{ee} \rangle), \qquad (34)$$

$$\frac{d}{dT} \langle R_{gyr} \rangle(T) = \frac{1}{T^2} (\langle E R_{gyr} \rangle - \langle E \rangle \langle R_{gyr} \rangle), \qquad (35)$$

we estimate the temperature region of the ground-state–globule transition to be within $T_0^{(1)} \approx 0.24$ and $T_0^{(2)} \approx 0.28$. The globule–random coil transition takes place between $T_1^{(1)} \approx 0.53$ and $T_1^{(2)} \approx 0.70$.

In Fig. 6 we show variances $\sigma_O^2 = \langle O^2 \rangle - \langle O \rangle^2$ as functions of temperature for $O = E$, R_{ee}, and R_{gyr}. We observe weak "shoulders" around $T = 0.3$ (to see this for σ_{gyr}^2 would require, however, an even higher resolution of the plot), close to the interval, where the low-temperature transition is expected. The situation is much more diffuse in the temperature region, where the globule–random coil transition should
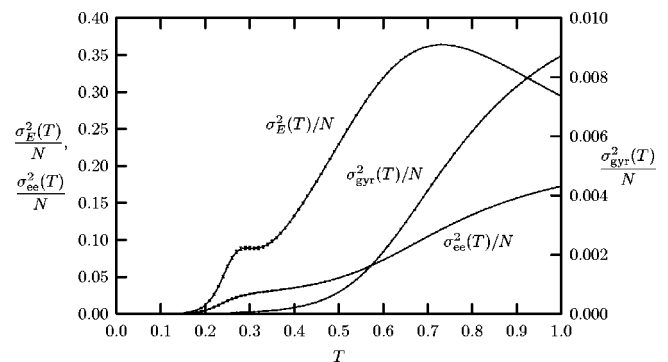

FIG. 4. Mean end-to-end distance $\langle R_{ee} \rangle$ and mean radius of gyration $\langle R_{gyr} \rangle$ of the 42mer.

FIG. 6. Temperature dependence of the fluctuations of energy σ_E^2, end-to-end distance σ_{ee}^2, and radius of gyration σ_{gyr}^2 for the 42mer. Note the different scales for σ_{ee}^2 and σ_{gyr}^2. In the temperature interval plotted the variance of the radius of gyration (right scale) is much smaller than the variance of the end-to-end distance (left scale), $\sigma_{gyr}^2 < \sigma_{ee}^2$.
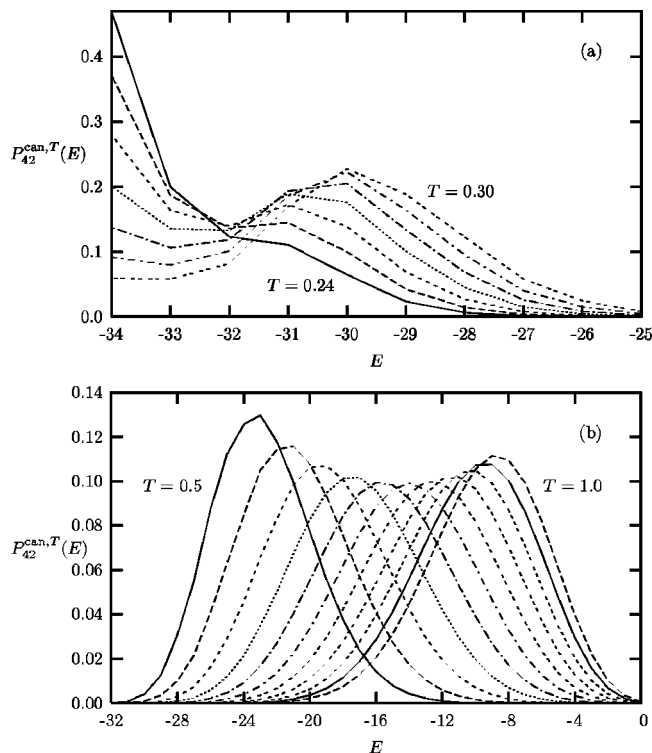
FIG. 7. Canonical distributions for the 42mer at temperatures (a) $T = 0.24, 0.25,..., 0.30$ close to the ground-state–globule transition region between $T_0^{(1)} \approx 0.24$ and $T_0^{(2)} \approx 0.28$, (b) $T = 0.50, 0.55,..., 1.0$. The high-temperature peak of the specific heat in Fig. 5 is near $T_1^{(1)} \approx 0.53$, but at $T_1^{(2)} \approx 0.73$ the distribution has the largest width, cf. Fig. 6. Near this temperature, the mean radius of gyration and the mean end-to-end distance (see Figs. 4 and 5) have their biggest slope.

take place. The variance of the energy σ_E^2 has a peak at $T = 0.73$, near the temperatures of the corresponding peaks of the derivatives (34) and (35) plotted in Fig. 5. The variances of the end-to-end distance σ_{ee}^2 and the radius of gyration σ_{gyr}^2, however, do not exhibit at all a peak near this temperature. Obviously, there is no unique behavior of these quantities that is usually used to identify conformational transitions in this temperature region. In Fig. 7 we have plotted the canonical distributions $P_{42}^{can,T}(E)$ for different temperatures in the vicinity of the two transitions. From Fig. 7(a) we read

off that the distributions possess two peaks at temperatures within that region where the ground-state–globule transition takes place. This is interpreted as an indication of a "first-order-like" transition.[20] The behavior in the vicinity of the globule–random coil transition is less spectacular as can be seen in Fig. 7(b), and since the energy distribution shows one peak only, this transition could be denoted as being "second-order-like." The width of the distributions grows with increasing temperature until it has reached its maximum value, which is located near $T \approx 0.7$, cf. Fig. 6. For higher temperatures, the distributions become narrower again.

Since finite-size scaling is impossible because of the noncontinuable sequences of different types of monomers, "transitions" between classes of protein shapes are, of course, to be distinguished from phase transitions in the strict thermodynamic sense. In conclusion, conformational transitions for polymers of finite size, such as proteins, are usually weak and therefore difficult to identify. Thus, these considerations are, of course, of limited thermodynamic significance. From a technical point of view, however, it is of some importance since Markovian Monte Carlo algorithms can have problems with sampling the entire energy space, as the probability in the gap between the two peaks can be suppressed by many orders of magnitude (what is obviously not the case in our example of the 42mer) and tunnelings are extremely rare. Just for such situations, flat histogram algorithms have primarily been developed.[1,2]

B. Ten designed 48mers

We have also analyzed the ten designed sequences with 48 monomers given in Ref. 21. The ratio between the numbers of hydrophobic and polar residues in one half for these HP proteins, i.e., the hydrophobicity is $n_H = 24$. In Table II we have listed the sequences and ground-state properties. The minimum energies we found coincide with the values given in Refs. 6, 7, and 21. Figure 8 shows the densities of states for selected 48mers and the multicanonical histograms of the production run. Note that for Rosenbluth chain-growth methods (a-thermal or at $\beta = 0$) the histogram for chains of length N is obtained by accumulating their individual Rosenbluth weights W_N^R, which explains the poorer performance near the minimum energy, where a small number of states

TABLE II. Properties of the 48mers. For each of the sequences we have listed the ground-state energy E_{min} and the ground-state degeneracy g_0 estimated with our algorithm. For comparison, we have also quoted the lower bounds on native degeneracies $g_{CHCC}^<$ obtained by means of the CHCC (constrained-based hydrophobic core construction) method (Ref. 22) as given in Ref. 21. In both cases the constant factor 48 from rotational and reflection symmetries of conformations spreading into all three spatial directions was divided out.

| No.	Sequence	E_{min}	$g_0 (\times 10^3)$	$g_{CHCC}^< (\times 10^3)$
48.1	$HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3HP_8H_2$	-32	5226 ± 812	1500
48.2	$H_4PH_2PH_5P_2HP_2H_2P_2HP_6HP_2HP_3HP_2H_2P_2H_3PH$	-34	17 ± 8	14
48.3	$PHPH_2PH_6P_2HPHP_2HPH_2PHPHP_3HP_2H_2P_2H_2PHP_2HP$	-34	6.6 ± 2.8	5.0
48.4	$PHPH_2P_2HPH_3P_2H_2PH_2P_3H_5P_2HPH_2PHPHP_4HP_2HPHP$	-33	60 ± 13	62
48.5	$P_2HP_3HPH_4P_2H_4PH_2P_3P_2HPHPHP_2HP_6H_2PH_2PH$	-32	1200 ± 332	54
48.6	$H_3P_3H_2PHPH_2PH_2PHP_7HPHP_2HP_3HP_2H_6PH$	-32	96 ± 19	52
48.7	$PHP_4HPH_3PHPH_4PH_2PH_2P_3HPHP_3H_3P_2H_2P_2H_2P_3H$	-32	58 ± 21	59
48.8	$PH_2PH_3PH_4P_2H_3P_6HPH_2P_2H_2PHP_3H_2PHPHPH_2P_3$	-31	22201 ± 6594	306
48.9	$PHPHP_4HPHPHP_2HPH_6P_2H_3PHP_2HPH_2P_2HPH_3P_4H$	-34	1.4 ± 0.5	1.0
48.10	$PH_2P_6H_2P_3H_3PHP_2HPH_2P_2HP_2HP_2H_2P_2H_7P_2H_2$	-33	187 ± 87	188
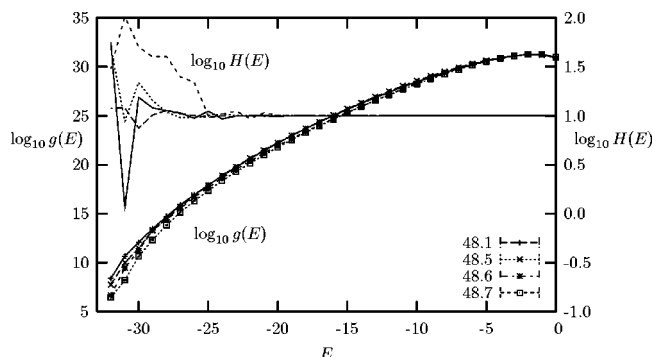
FIG. 8. Logarithmic plots of the densities of states $g(E)$ and "flat" histograms $H(E)$ for the sequences 48.1, 48.5, 48.6, and 48.7 from Table II that have the same lowest energy $E_{min}=-32$. The normalization of the histograms of these examples was chosen such that they coincide at maximum energy, $\log_{10} H(E_{max}=0)=1$.

enters with big weights. This differs from the usual procedure in algorithms with importance sampling, where the counter of an energy bin being hit by an appropriate state is incremented by unity.

In Fig. 9 we have plotted the mean energy, free energy, and entropy as functions of temperature for these lattice proteins. These results were obtained by means of the density of states as sampled with our algorithm. For $T \to 0$ the curves for both $\langle E \rangle$ and F merge into the four different values of E_{min} ($=-34, -33, -32, -31$) while the entropy exhibits the ground-state degeneracies, $S \to \ln g_0$. Our estimates for the degeneracies g_0 of the ground-state energies, and for comparison, the lower bounds $g_{CHCC}^<$ given in Ref. 21, are listed in Table II. The lower bounds were obtained with the constrained-based hydrophobic core construction (CHCC) method.[22] Our values lie indeed above these lower bounds or include it within the range of statistical errors. Notice that for the sequences 48.1, 48.5, and 48.8, our estimates for the ground-state degeneracy are much higher than the bounds $g_{CHCC}^<$. In these cases the smallest frame containing the entire hydrophobic core is rather large (cube containing $4 \times 3 \times 3 = 36$ monomers with surface area $A = 32$ [bond length]^2) such that enumeration of this frame is cumbersome. For 48.5 and 48.8, we further found ground-state conformations lying in less compact frames (48.5: $A = 32, 40, 42, 48, 52, 54$ [bond length]^2, 48.8: $A = 32, 40,$
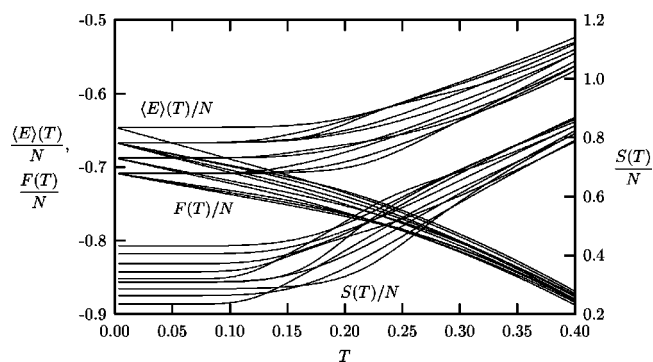

FIG. 9. Mean energy $\langle E \rangle (T)$, Helmholtz free energy $F(T)$, and entropy $S(T)$ for 48mers with the sequences given in Table II.

42 [bond length]^2) and those conformations would require still more effort to be identified with the CHCC algorithm, which was designed to locate global energy minima and therefore starts the search beginning from the most compact hydrophobic frames. The ground-state energies of these examples are rather high ($E_{min}=-31$ for 48.8, and $E_{min}=-32$ for 48.1 and 48.5) and therefore a higher degeneracy seems to be natural. This is, however, only true, if there does not exist a conformational barrier that separates the compact H-core low-energy states from the general compact globules. Comparing the ground-state degeneracies and the low-temperature behavior of the specific heats for the sequences 48.1, 48.5, 48.6, and 48.7 (all of them having global energy minima with $E_{min}=-32$) as shown in Figs. 8 and 10, respectively, we observe that 48.6 and 48.7 with rather low ground-state degeneracy actually possess a pronounced low-temperature peak in the specific heat, while the higher-degenerate proteins 48.1 and 48.5 only show a weak indication of a structural transition at low temperatures. The HP proteins 48.2, 48.3, and 48.9, which have the lowest minimum energy $E_{min}=-34$ among the examples in Table II, also have the lowest ground-state degeneracies. These three candidates indeed seem to exhibit a rather strong ground-state–globule transition, as can be read off from the associated specific heats in Fig. 10.

We have again measured the mean end-to-end distances and mean radii of gyration, which are also plotted as functions of temperature into Fig. 10. Both quantities usually serve to interpret the conformational compactness of polymers. For HP proteins, the end-to-end distance is strongly influenced, however, by the types of monomers attached to the ends of the chain. It is easily seen from the figures that the 48mers with sequences starting and ending with a hydrophobic residue (48.1, 48.2, and 48.6) have a smaller mean end-to-end distance at low temperatures than the other examples from Table II. The reason is that the ends can form hydrophobic contacts and therefore a reduction of the energy can be achieved. Thus, in these cases contacts between ends are usually favorable and the mean end-to-end distance is close to the mean radius of gyration. Interestingly, there indeed exists a crossover region, where $\langle R_{ee} \rangle < \langle R_{gyr} \rangle$. Compared with the behavior of the specific heat, this interval is close to the region, where the phase dominated by low-energy states crosses over to the globule-favored phase. The hydrophobic contact between the ends is strong enough to resist the thermal fluctuations in that temperature interval. The reason is that once such a hydrophobic contact between the ends is established, usually other in-chain hydrophobic monomers are attracted and form a hydrophobic core surrounding the end-to-end contact. Thus, before the contact between the ends is broken, an increase of the temperature first leads to a melting of the surrounding contacts. The entropic freedom to form new conformations is large since the low-energy states are all relatively high-degenerate and do not possess symmetries requiring an appropriate amount of heat to be broken. For sequences possessing mixed or purely polar ends, the mean end-to-end distance and mean radius of gyration differ much more strongly, as there is no energetic reason why the ends should be next neighbors.
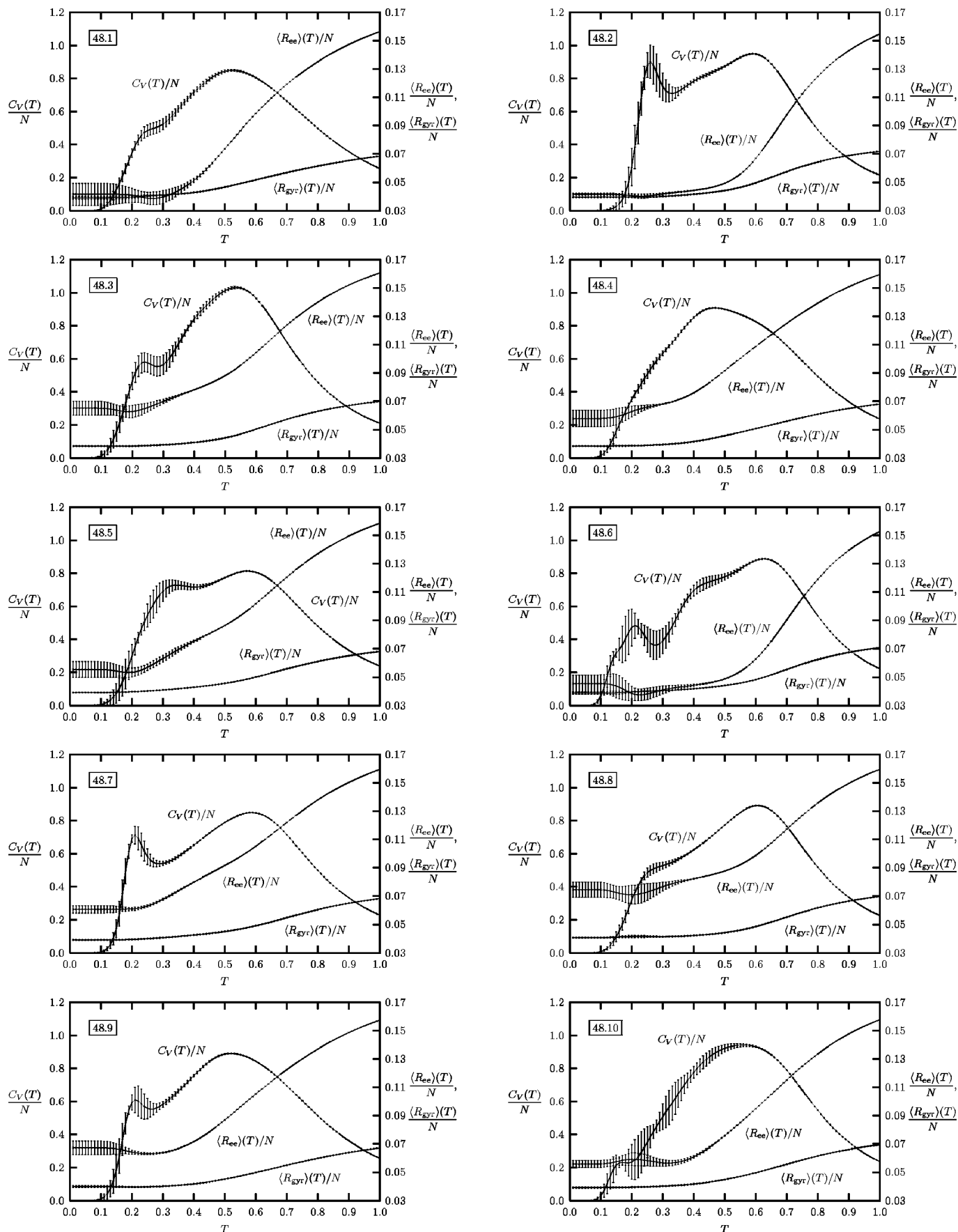
FIG. 10. Heat capacities $C_V(T)$, mean end-to-end distances $\langle R_{ee}\rangle(T)$, and mean radii of gyration $\langle R_{gyr}\rangle(T)$ of the ten designed 48mers from Table II.
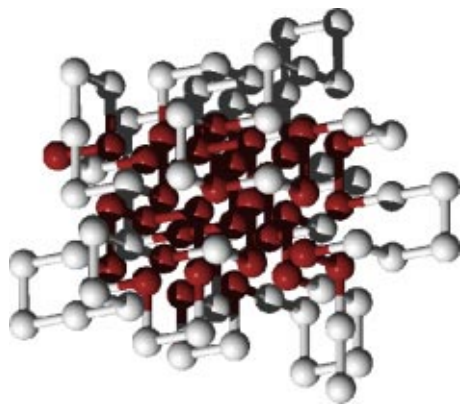
FIG. 11. Conformation of the 103mer with the lowest energy found, $E_{min} = -56$.

FIG. 13. Specific heat of the 103mer.

In conclusion, we see that for longer chains the strength of the low-temperature transition does not only depend on low ground-state degeneracies as it does for short chains.[16] Rather, the influence of the higher-excited states cannot be neglected. A striking example is sequence 48.4 with rather low ground-state degeneracy, but only weak signals for a low-temperature transition.

C. Beyond 100 monomers...

The final example we applied our algorithm to was a 103mer as proposed in Ref. 23. Until recently, the "ground state" was believed to have energy $E_{min} = -49$ (Ref. 24). The best estimate up to now was found with nPERMis to be $E_{min} = -54$ (Ref. 6) and, with an additional bias suppressing contacts between H and P monomers, even $E_{min} = -55$ (Ref. 7). Our algorithm not only decreased the lowest-energy value to $E_{min} = -56$ (see Fig. 11), but also enabled us to obtain results for the thermodynamic quantities as in the previous examples. Figure 12 shows the density of states which covers more than 50 orders of magnitude from which we determined the specific heat shown in Fig. 13. The degeneracy of the lowest-energy state $E_{min} = -56$ was determined to be of order 10^{16} such that it seems likely that there still exist one or more even lower-lying energetic states.
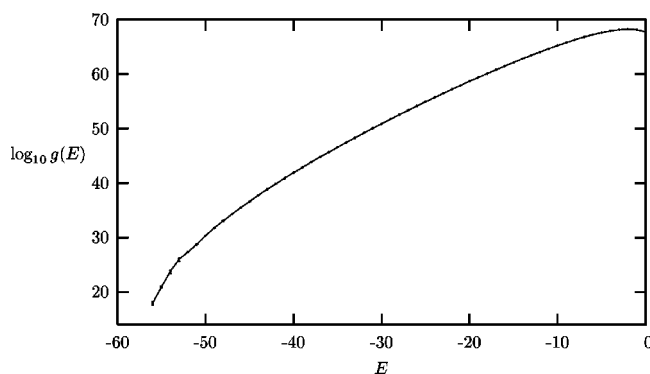

FIG. 12. Density of states for the 103mer, ranging over more than 50 orders of magnitude.
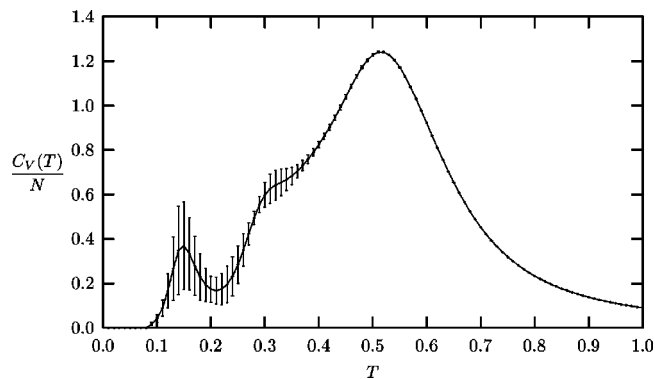
VII. SUMMARY

In order to study heteropolymers at very low temperatures with reasonable accuracy, we developed a multicanonical chain-growth algorithm based on recently improved variants of PERM. Comparing with exact enumeration data for HP lattice proteins with 14 monomers, we validated that our method is suitable to accurately determine thermodynamic quantities for all temperatures. Then, we applied this algorithm to lattice proteins with sequences of more than 40 monomers known from literature. Additionally, we determined statistical properties for all temperatures for examples with up to 103 monomers. Since our algorithm allows the estimation of the degeneracy of the energy states, we determined for all sequences the ground-state degeneracy, as it is an indication of the "uniqueness" of the native state.

In particular, we presented a detailed investigation of a sequence of 42 monomers that has interesting characteristic properties, e.g., a quite low ground-state degeneracy. Since some results regarding the ground states and thermodynamic properties were available,[20] it was a good candidate for testing our algorithm and for checking the performance of our method. For this sequence we analyzed in detail the temperature-dependent behavior of radius of gyration, end-to-end distance, as well as their fluctuations, and compared it with the specific heat in order to elaborate relations between characteristic properties of these curves (peaks, "shoulders") and conformational transitions not being transitions in a strict thermodynamic sense due to the impossibility to formulate a thermodynamic limit for proteins. Therefore, we identified temperature regions, where global changes of protein conformations occur. These transition regions separate "phases," where random coils, maximally compact globules, or states with compact hydrophobic core dominate. The random-coil–globule transition is well understood, since the heteropolymers behave qualitatively similar to homopolymers at the Θ collapse transition. For finite systems this (pseudo)transition occurs at the Boyle temperature T_N being defined through the vanishing of the second virial coefficient in the homopolymer state equation for small monomer concentrations.[25] The description of the conformational changes happening at the low-temperature transition is, however, more difficult. From exact enumeration for 14mers,[16] simulations of the 42mer,[8,20] and our results for the 42mer and the 48mers in this paper, we conclude that the indica-

tions exhibiting this transition are no numerical artifacts. Qualitatively, we expect that maximally compact globules refold into conformations with very compact hydrophobic core as this increases the number of hydrophobic contacts, partially at the expense of maximal globular compactness. Conformational indicators such as mean end-to-end distance (that depends too much on the monomer types at the ends) and mean radius of gyration are not sufficient to explain this transition. In polymer models without disorder[26,27] and with random interactions between monomers,[28] which behave similarly, there are indications that this transition could be of glassy type, with dynamics very similar to the one observed in supercooled liquids and structural glasses.[29] Approving this observation for the low-temperature behavior of the HP model would require, however, the introduction of an appropriate conformational order parameter being more relevant than end-to-end distance and radius of gyration.

For comparing properties of lattice proteins with the same number of monomers and identical hydrophobicity, but different sequence—thus allowing for a kind of conformational–sequential analysis—we studied energetic and conformational thermodynamic quantities for the famous list of ten 48mers given in Ref. 21 in great detail. This also allowed us to study how conformational properties and the strength of shape transitions depend on the protein sequence. As an interesting by-product, we not only confirmed the known global-minimum energies for these examples, but we even found a new minimum for the 103mer being the longest sequence under consideration.

ACKNOWLEDGMENT

[1] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).

[2] W. Janke, Physica A **254**, 164 (1998); B. A. Berg, Fields Inst. Comm. **26**, 1 (2000).

[3] W. Janke, in *Computer Simulations of Surfaces and Interfaces*, NATO Science Series, II. Mathematics, Physics, and Chemistry, Vol. 114, Proceedings of the NATO Advanced Study Institute, Albena, Bulgaria, September 2002, edited by B. Dünweg, D. P. Landau, and A. I. Milchev (Kluwer, Dordrecht, 2003), p. 137.

[4] P. Grassberger, Phys. Rev. E **56**, 3682 (1997); H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, Phys. Rev. Lett. **80**, 3149 (1998); U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, Proteins **32**, 52 (1998).

[5] P. Grassberger and W. Nadler, in *Computational Statistical Physics—From Billiards to Monte Carlo*, edited by K. H. Hoffmann and M. Schreiber (Springer, Berlin, 2002), p. 169, and references therein.

[6] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003).

[7] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, Phys. Rev. E **68**, 021113 (2003).

[8] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003).

[9] K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[10] See, e.g., L. W. Lee and J.-S. Wang, Phys. Rev. E **64**, 056112 (2001).

[11] N. Madras and A. D. Sokal, J. Stat. Phys. **50**, 109 (1988).

[12] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).

[13] Note that with the normalization in Eq. (13) $Z_1 = 1$. Nonetheless, we have kept it in Eq. (19) to retain the form of corresponding expressions in previous publications (Refs. 6–8), where the normalization was left open.

[14] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[15] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[16] M. Bachmann and W. Janke, Acta Phys. Pol. B **34**, 4689 (2003).

[17] M. D. Yoder, N. T. Keen, and F. Jurnak, Science **260**, 1503 (1993).

[18] T. C. Beutler and K. A. Dill, Protein Sci. **5**, 2037 (1996).

[19] R. G. Miller, Biometrika **61**, 1 (1974); B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (SIAM, Philadelphia, 1982).

[20] G. Chikenji, M. Kikuchi, and Y. Iba, Phys. Rev. Lett. **83**, 1886 (1999), and references therein.

[21] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).

[22] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993); Proc. Natl. Acad. Sci. U.S.A. **92**, 146 (1995).

[23] E. E. Lattman, K. M. Fiebig, and K. A. Dill, Biochemistry **33**, 6158 (1994).

[24] L. Toma and S. Toma, Protein Sci. **5**, 147 (1996).

[25] P. Grassberger and R. Hegger, J. Chem. Phys. **102**, 6881 (1995).

[26] E. Pitard and J.-P. Bouchaud, Eur. Phys. J. E **5**, 133 (2001).

[27] N. V. Dokholyan, E. Pitard, S. V. Buldyrev, and H. E. Stanley, Phys. Rev. E **65**, 030801(R) (2002).

[28] E. Pitard and E. I. Shakhnovich, Phys. Rev. E **63**, 041501 (2001).

[29] W. Kob, in *Soft and Fragile Matter*, edited by M. E. Cates and M. R. Evans (IOP, London, 2000), p. 249.