Multicanonical study of coarse-grained off-lattice models for folding heteropolymers

Michael Bachmann,[1,*] Handan Arkın,[1,2,†] and Wolfhard Janke[1,‡]

[1]*Institut für Theoretische Physik, Universität Leipzig, Augustusplatz 10/11, D-04109 Leipzig, Germany*
[2]*Department of Physics Engineering, Hacettepe University, 06532 Ankara, Turkey*

We have performed multicanonical simulations of hydrophobic-hydrophilic heteropolymers with two simple effective, coarse-grained off-lattice models to study the influence of specific interactions in the models on conformational transitions of selected sequences with 20 monomers. Another aspect of the investigation was the comparison with the purely hydrophobic homopolymer and the study of general conformational properties induced by the "disorder" in the sequence of a heteropolymer. Furthermore, we applied an optimization algorithm to sequences with up to 55 monomers and compared the global-energy minimum found with lowest-energy states identified within the multicanonical simulation. This was used to find out how reliable the multicanonical method samples the free-energy landscape, in particular for low temperatures.

PACS number(s): 87.15.Aa, 87.15.Cc, 05.10.−a

I. INTRODUCTION

The understanding of protein folding is one of the most challenging objectives in biochemically motivated research. Although the physical principles are known, the complexity of proteins as being macromolecules consisting of numerous atoms, the influence of quantum chemical details on long-range interactions, as well as the role of the solvent, etc., make an accurate analysis of the folding process of realistic proteins extremely difficult. Therefore, one of the most important questions in this field is how much detailed information can be neglected to establish effective, coarse-grained models yielding reasonable, at least qualitative, results that allow for, e.g., a more global view on the relationship between the sequence of amino acid residues and the existence of a global, funnel-like energy minimum in a rugged free-energy landscape [1].

Within the past two decades, much work has been devoted to introduce minimalistic models based on general principles that are believed to primarily control the structure formation of proteins. One of the most prominent examples is the HP model of lattice proteins [2], which has been exhaustively investigated without revealing all secrets, despite its simplicity. In this model, only two types of monomers are considered, with hydrophobic (H) and polar (P) character. Chains on the lattice are self-avoiding to account for the excluded volume. The only explicit interaction is between nonadjacent but next-neighbored hydrophobic monomers. This interaction of hydrophobic contacts is attractive to force the formation of a compact hydrophobic core which is screened from the (hypothetic) aqueous environment by the polar residues. Statistical mechanics simulations of this model are still the subject of studies requiring the application of sophisticated algorithms [3–5].

A manifest off-lattice generalization of the HP model is the AB model [6], where the hydrophobic monomers are labeled by A and the polar or hydrophilic ones by B. The contact interaction is replaced by a distance-dependent Lennard-Jones type of potential accounting for short-range excluded volume repulsion and long-range interaction, the latter being attractive for AA and BB pairs and repulsive for AB pairs of monomers. An additional interaction accounts for the bending energy of any pair of successive bonds. This model was first applied in two dimensions [6] and generalized to three-dimensional AB proteins [7,8], partially with modifications taking implicitly into account additional torsional energy contributions of each bond.

More knowledge-based coarse-grained models are often of Gō type, where, in the simplest lattice formulation, the energy is proportional to the number of native contacts. These models usually require knowledge of the native conformation and serve, e.g., as models for studies of folding pathways [9,10] and native topology [11,12]. There is also growing interest in modeling the folding behavior of single-domain proteins with simplified models as many of them exhibit a simple two-state kinetics [13] without intermediary states that would slow down the folding dynamics ("traps"). It seems, however, that native-centric models such as those of Gō type require modifications for a qualitatively correct description of this sharp folding transition [14,15].

In this paper, we report studies of thermodynamic and ground-state properties of AB sequences known from the literature [7,16,17] for two representations of the AB model [6,7] which are described in Sec. II. While, compared to all-atom formulations, the interactions in these coarse-grained models are greatly simplified and hence can be computed much faster, they still preserve a complicated rugged free-energy landscape, where naive simulations would easily get trapped. In order to accurately resolve the low-temperature behavior, we, therefore, applied a multicanonical Monte Carlo algorithm [18,19] with an appropriate update mechanism. For simulations of systems with complex free-energy landscapes, multicanonical sampling has become very popular and its application in protein simulations has a long tradition [20]. The ground-state search was achieved by means of the energy landscape paving minimizer (ELP) [21]. Section III is devoted to a description of these methods. We

*Email address: Michael.Bachmann@itp.uni-leipzig.de
†Email address: Handan.Arkin@itp.uni-leipzig.de
‡Email address: Wolfhard.Janke@itp.uni-leipzig.de

present the results for the global energy minima and the thermodynamic quantities in Sec. IV. The paper is concluded by the summary in Sec. V.

II. EFFECTIVE OFF-LATTICE MODELS

We investigated two effective off-lattice models of AB type for heteropolymers with N monomers. The first one is the original AB model as proposed in Ref. [6] with the energy function

$$E_I = \frac{1}{4}\sum_{k=1}^{N-2}(1-\cos\vartheta_k) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\left(\frac{1}{r_{ij}^{12}} - \frac{C_I(\sigma_i,\sigma_j)}{r_{ij}^6}\right),$$
(1)

where the first sum runs over the $(N-2)$ angles $0 \leq \vartheta_k \leq \pi$ of successive bond vectors. This term is the bending energy and the coupling is "ferromagnetic," i.e., it costs energy to bend the chain. The second term partially competes with the bending barrier by a potential of Lennard-Jones type depending on the distance between monomers being nonadjacent along the chain. It also accounts for the influence of the AB sequence ($\sigma_i=A$ for hydrophobic and $\sigma_i=B$ for hydrophilic monomers) on the energy of a conformation as its long-range behavior is attractive for pairs of like monomers and repulsive for AB pairs of monomers,

$$C_I(\sigma_i,\sigma_j) = \begin{cases} +1, & \sigma_i,\sigma_j = A, \\ +1/2, & \sigma_i,\sigma_j = B, \\ -1/2, & \sigma_i \neq \sigma_j. \end{cases}$$
(2)

We will refer to this model as AB model I throughout the paper.

The other model we have studied has been introduced in Ref. [7] and is a variant of AB model I in that it also consists of angular and distance-dependent energy terms, but with some substantial modifications. In the following, we denote it as AB model II. The energy is given by

$$E_{II} = -\kappa_1\sum_{k=1}^{N-2}\mathbf{b}_k\cdot\mathbf{b}_{k+1} - \kappa_2\sum_{k=1}^{N-3}\mathbf{b}_k\cdot\mathbf{b}_{k+2}$$
$$+ 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}C_{II}(\sigma_i,\sigma_j)\left(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^6}\right),$$
(3)

where $\mathbf{b}_k$ is the bond vector between the monomers k and $k+1$ with length unity. In Ref. [7], different values for the parameter set (κ_1,κ_2) were tested and finally set to $(-1,0.5)$ as this choice led to distributions for the angles between bond vectors $\mathbf{b}_k$ and $\mathbf{b}_{k+1}$ as well as the torsion angles between the surface vectors $\mathbf{b}_k\times\mathbf{b}_{k+1}$ and $\mathbf{b}_{k+1}\times\mathbf{b}_{k+2}$ that agreed best with distributions obtained for selected functional proteins. Since $\mathbf{b}_k\cdot\mathbf{b}_{k+1}=\cos\vartheta_k$, the choice $\kappa_1=-1$ makes the coupling between successive bonds "antiferromagnetic" or "antibending" contrary to what was chosen in Eq. (1) for AB model I. The second term in Eq. (3) takes torsional interactions into account without being an energy associated with the pure torsional barriers in the usual sense. The third term contains now a pure Lennard-Jones

potential, where the $1/r_{ij}^6$ long-range interaction is attractive whatever types of monomers interact. The monomer-specific prefactor $C_{II}(\sigma_i,\sigma_j)$ only controls the depth of the Lennard-Jones valley,

$$C_{II}(\sigma_i,\sigma_j) = \begin{cases} +1, & \sigma_i,\sigma_j = A, \\ +1/2, & \sigma_i,\sigma_j = B \quad\text{or}\quad \sigma_i \neq \sigma_j. \end{cases}$$
(4)

For technical reasons, we have introduced in both models a cutoff $r_{ij}=0.5$ for the Lennard-Jones potentials below which the potential is hard-core repulsive (i.e., the potential is infinite). For both models, only a few results are given in the literature. For the first model, these are estimates for the global energy minima of certain AB sequences [16], while for AB model II, primarily thermodynamic quantities were determined [7]. We have performed ELP optimizations [21] in order to find deeper-lying energy minima than the values quoted in the literature and, in particular, multicanonical sampling [18,19] for enabling us to focus on thermodynamic properties of the sequences given in Ref. [7].

III. METHODS

In this section, we describe the computational sampling methods and the update procedure we applied to obtain results for off-lattice AB heteropolymers.

A. Energy-landscape paving optimization procedure

In order to locate global energy minima of a complex system, it is often useful to apply specially biased algorithms that only serve this purpose. We used the energy landscape paving (ELP) minimizer [21] to find global energy minima of the sequences under consideration. The ELP minimization is a Monte Carlo optimization method, where the energy landscape is locally flattened. This means that if a state $\mathbf{x}$ with energy $E(\mathbf{x})$ is hit, the energy is increased by a "penalty" which itself depends on the histogram of any suitably chosen order parameter. The simplest choice is the energy distribution $H(E)$ such that we define $\tilde{E}_t=E_t+f(H(E_t))$. Thus, the Boltzmann probability $\exp(-\tilde{E}_t/k_BT)$ for a METROPOLIS update, where k_BT is the thermal energy at the temperature T, becomes a function of "time" t. The advantage of the ELP method is that local energy minima are filled up and the likelihood of touching again recently visited regions decreases. This method has successfully proved to be applicable to find global energy minima in rough energy landscapes of proteins [21,22]. Of course, as a consequence of the biased sampling, stochastic methods along the line of ELP violate detailed balance and it is therefore inappropriate to apply those for uncovering thermodynamic properties of a statistical system.

B. Multicanonical method

To obtain statistical results, we applied a multicanonical Monte Carlo algorithm, where the energy distribution is flattened artificially, allowing, in principle, for a random walk of successive states in energy space. This flattening is controllable and therefore reproducible. For this purpose, the
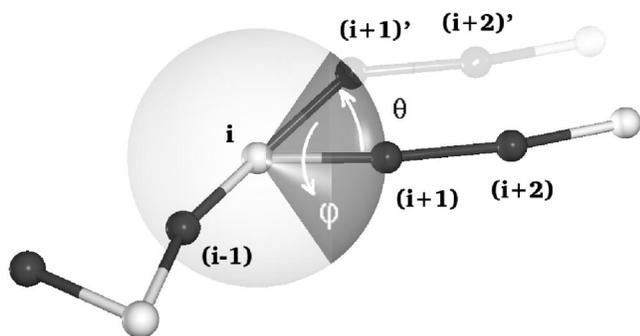
FIG. 1. Spherical update of the bond vector between the ith and $(i+1)$th monomer.

Boltzmann probability is multiplied by a weight factor $W(E)$, which in our case is a function of the energy. Then the probability for a state with energy E reads $p_M(E) = \exp(-E/k_BT)W(E)$. In order to obtain a multicanonical or "flat" distribution, the initially unknown weight function $W(E)$ has to be determined iteratively: In the beginning, the weights $W^{(0)}(E)$ are set to unity for all energies letting the first run be a usual METROPOLIS simulation which yields an estimate $H^{(0)}(E)$ for the canonical distribution. The histogram is used to determine the next guess for the weights; the simplest update is to calculate $W^{(1)}(E) = H^{(0)}(E)/W^{(0)}(E)$. Then the next run is performed with probabilities $p_M^{(1)}(E) = \exp(-E/k_BT)W^{(1)}(E)$ of states with energy E. The iterative procedure is continued until the weights are appropriate in a way that the histogram is "flat." The introduction of a flatness criterion, for instance, that the fluctuations around the average value of the histogram are less than 20%, is useful, but not necessary, if the number of iterations is very large. After having determined accurate weights $W(E)$, they are kept fixed, and following some thermalization sweeps a long production run is performed, where statistical quantities O are obtained multicanonically, $\langle O \rangle_M = \Sigma_{\{\mathbf{x}\}} p_M(E(\{\mathbf{x}\}))O(\{\mathbf{x}\})/Z_M$ with the multicanonical partition function $Z_M = \Sigma_{\{\mathbf{x}\}} p_M(E(\{\mathbf{x}\}))$. The canonical statistics is obtained by reweighting the multicanonical to the canonical distribution, i.e., mean values are computed as $\langle O \rangle = \langle OW^{-1} \rangle_M / \langle W^{-1} \rangle_M$.

For the determination of the multicanonical weights, we performed 200 iterations with at least 10^5 sweeps each. The histograms obtained in the iteration runs were accumulated error-weighted [19] such that the estimation of the weights was based on increasing statistics and not only on the relatively small number of sweeps per run. In the production period, 5×10^7 sweeps were generated to have reasonable statistics for estimating the thermodynamic quantities. Statistical errors are estimated with the standard Jackknife technique [23,24].

C. Spherical update mechanism

For updating a conformation, we use the procedure displayed in Fig. 1. Since the length of the bonds is fixed ($|\mathbf{b}_k| = 1$, $k = 1, \ldots, N-1$), the $(i+1)$th monomer lies on the surface of a sphere with radius unity around the ith monomer. Therefore, spherical coordinates are the natural choice for calculating the new position of the $(i+1)$th monomer on this sphere. For reasons of efficiency, we do not select any point on the sphere but restrict the choice to a spherical cap with maximum opening angle $2\theta_{max}$ (the dark area in Fig. 1). Thus, to change the position of the $(i+1)$th monomer to $(i+1)'$, we select the angles θ and φ randomly from the respective intervals $\cos\theta_{max} \leqslant \cos\theta \leqslant 1$ and $0 \leqslant \varphi \leqslant 2\pi$, which ensure a uniform distribution of the $(i+1)$th monomer positions on the associated spherical cap. After updating the position of the $(i+1)$th monomer, the following monomers in the chain are simply translated according to the corresponding bond vectors which remain unchanged in this type of update. Only the bond vector between the ith and the $(i+1)$th monomers is rotated; all others keep their direction. This is similar to single spin updates in local-update Monte Carlo simulations of the classical Heisenberg model with the difference that in addition to local energy changes, long-range Lennard-Jones interactions of the monomers, changing their relative position to each other, have to be computed anew after the update. In our simulations of the AB models, we used a very small opening angle, $\cos\theta_{max} = 0.99$, in order to be able to sample also very narrow and deep valleys in the landscape of angles.

IV. RESULTS

We applied the multicanonical algorithm primarily to study thermodynamic properties, e.g., the "phase" behavior of off-lattice heteropolymers. Before discussing these results, however, we analyze the capability of multicanonical sampling to find lowest-energy conformations by comparison with the dedicated minimization algorithm ELP, in particular the native fold, because the identification of these structures is not only interesting as a by-product of the simulation. Rather, since they dominate the low-temperature behavior, it is necessary that they are generated frequently in the multicanonical sampling.

A. Search for global energy minima

In order to be able to investigate the low-temperature behavior of AB off-lattice heteropolymers, we first have to assess the capabilities of the multicanonical method to sample lowest-energy conformations and, in particular, to approach closely the global energy minimum conformation. In Table I, we list the Fibonacci sequences [6] that have already been studied in Ref. [16] by means of an off-lattice variant of the improved version [4] of the celebrated chain-growth algorithm with population control, PERM [25]. In Ref. [16], first estimates for the putative ground-state energies of the Fibonacci sequences with 13 to 55 monomers were given for AB model I [6] in three dimensions. We compare these results with the respective lowest energies found in our multicanonical simulations and with the results obtained with the minimization algorithm ELP [21]. It turns out that the ground-state energies found by multicanonical sampling agree well with what comes out by the ELP minimizer, cf.

TABLE I. The four Fibonacci sequences, for which we analyzed the ground states in both models.

N	sequence
13	$AB_2AB_2ABAB_2AB$
21	$BABAB_2ABAB_2AB_2ABAB_2AB$
34	$AB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB$
55	$BABAB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB$

Table II. Another interesting result is that our estimates for the ground-state energies lie significantly below the energies quoted in Ref. [16], obtained with the off-lattice PERM variant, and our values are even lower than the energies obtained by PERM and subsequent conjugate-gradient minimization in the attraction basin. Not unexpectedly, this is particularly pronounced for the longest chain considered.

We have also performed this test with AB model II [7]. Here we can compare our results with minimum energies listed in Ref. [17], where the so-called annealing contour Monte Carlo (ACMC) method was applied to these Fibonacci sequences. From Table II, we see that we find with MUCA and ELP runs lower energies for the sequences with 21, 34, and 55 monomers, while the results for the short 13mer are comparable.

Note that the multicanonical algorithm is not tuned to give good results in the low-energy sector only. For all sequences studied in this work, the same algorithm also yielded the thermodynamic results to be discussed in the following section.

In order to check the structural similarities of the lowest-energy conformations obtained with multicanonical sampling and those from the ELP optimization runs, $\mathbf{X}^{\mathrm{MUCA,ELP}} = (\mathbf{x}_1^{\mathrm{MUCA,ELP}}, \mathbf{x}_2^{\mathrm{MUCA,ELP}}, \ldots, \mathbf{x}_N^{\mathrm{MUCA,ELP}})$, we calculated the root mean square deviation D_{rms} of the respective pairs,

$$D_{\mathrm{rms}} = \min \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\tilde{\mathbf{x}}_i^{\mathrm{MUCA}} - \tilde{\mathbf{x}}_i^{\mathrm{ELP}}|^2}. \qquad (5)$$

Here, $\tilde{\mathbf{x}}_i^{\mathrm{MUCA,ELP}} = \mathbf{x}_i^{\mathrm{MUCA,ELP}} - \mathbf{x}_0^{\mathrm{MUCA,ELP}}$ denotes the position with respect to the respective centers of mass $\mathbf{x}_0 = \sum_{j=1}^{N}\mathbf{x}_j/N$ of the ith monomer of the lowest-energy conformations

found with multicanonical sampling and ELP minimization, respectively. Obviously, D_{rms} is zero for exactly coinciding conformations, and the larger the value, the worse the coincidence. The minimization of the sum in Eq. (5) is performed with respect to a global relative rotation of the two conformations in order to find the best match. For the explicit calculation, we used the exact quaternion-based optimization procedure described in Ref. [26].

As an alternative, we introduce here another parameter that enables us to compare conformations. Instead of performing comparisons of positions, it is much simpler and less time-consuming to calculate the so-called overlap between two conformations by comparing their bond and torsion angles. As an extension of the torsion-angle based variant [27,28], we define the more general overlap parameter

$$Q(\mathbf{X},\mathbf{X}') = \frac{N_t + N_b - d(\mathbf{X},\mathbf{X}')}{N_t + N_b}, \qquad (6)$$

where (with $N_t = N-3$ and $N_b = N-2$ being the numbers of torsional angles Φ_i and bond angles $\Theta_i = \pi - \vartheta_i$, respectively)

$$d(\mathbf{X},\mathbf{X}') = \frac{1}{\pi}\left(\sum_{i=1}^{N_t} d_t(\Phi_i,\Phi_i') + \sum_{i=1}^{N_b} d_b(\Theta_i,\Theta_i')\right),$$

$$d_t(\Phi_i,\Phi_i') = \min(|\Phi_i - \Phi_i'|, 2\pi - |\Phi_i - \Phi_i'|),$$

$$d_b(\Theta_i,\Theta_i') = |\Theta_i - \Theta_i'|.$$

Since $-\pi \le \Phi_i \le \pi$ and $0 \le \Theta_i \le \pi$, it follows immediately that $0 \le d_{t,b} \le \pi$. The overlap is unity if all angles of the conformations $\mathbf{X}$ and $\mathbf{X}'$ coincide, otherwise $0 \le Q < 1$.

TABLE II. Estimates for global energy minima obtained with multicanonical (MUCA) sampling and ELP minimization [21] for the Fibonacci sequences of Table I using both models. The values for AB model I are compared with results quoted in Ref. [16] employing off-lattice PERM and after subsequent conjugate-gradient minimization (PERM+). Minimum energies found with MUCA and ELP for AB model II are compared with the lowest energies listed in Ref. [17], obtained with annealing contour Monte Carlo (ACMC) and improved by METROPOLIS quenching (ACMC+).

	AB model I				AB model II			
N	$E_{\min}^{\mathrm{MUCA}}$	$E_{\min}^{\mathrm{ELP}}$	$E_{\min}^{\mathrm{PERM}}$ [16]	$E_{\min}^{\mathrm{PERM+}}$ [16]	$E_{\min}^{\mathrm{MUCA}}$	$E_{\min}^{\mathrm{ELP}}$	$E_{\min}^{\mathrm{ACMC}}$ [17]	$E_{\min}^{\mathrm{ACMC+}}$ [17]
13	−4.967	−4.967	−3.973	−4.962	−26.496	−26.498	−26.363	−26.507
21	−12.296	−12.316	−7.686	−11.524	−52.915	−52.917	−50.860	−51.718
34	−25.321	−25.476	−12.860	−21.568	−97.273	−97.261	−92.746	−94.043
55	−41.502	−42.428	−20.107	−32.884	−169.654	−172.696	−149.481	−154.505

TABLE III. Root mean square deviations D_{rms} and overlap Q of lowest-energy conformations found with multicanonical sampling and with the ELP optimizer for the Fibonacci sequences of length N given in Table I.

N	AB model I		AB model II	
	D_{rms}	Q	D_{rms}	Q
13	0.015	0.994	0.006	0.998
21	0.025	0.992	0.009	0.997
34	0.162	0.979	1.412	0.840
55	2.271	0.766	1.904	0.857

Note that both models are energetically invariant under reflection symmetry. This means that also the landscape of the free energy as a function of the bond and torsion angles is trivially symmetric with respect to the torsional degrees of freedom. Consequently, unless exceptional cases, there is thus a trivial twofold energetic degeneracy of the global energy minimum, but the respective root mean square deviation and overlap parameter are different for the associated conformations. The values quoted throughout the paper were obtained by comparing the lowest-energy conformation found with both reference conformations and quoting the value indicating better coincidence (i.e., lower D_{rms} and higher Q). This obvious ambiguity can be circumvented by adding a symmetry-breaking term to the model that disfavors, e.g., left-handed helicity [15].

In Table III, we list the values of both parameters for the lowest-energy conformations found for the Fibonacci sequences of Table I. We see that for the shortest sequences (with 13 and 21 monomers), the coincidence of the lowest-energy structures found is extremely good and we are pretty sure that we have found the ground states. In the case of the 34mer, modeled with AB model I, both structures also coin-cide still very well and it seems that the attraction valley towards the ground state was found within the multicanonical simulation. In the simulation of the 34mer with model II, the situation is different. As seen from Table II, we found surprisingly the marginally lower energy value in the multicanonical simulation, but the associated conformation differs significantly from that identified with ELP. It is likely that both conformations do not belong to the same attraction basin and it is a future task to reveal whether this is a first indication of metastability. The situation is even more complex for the 55mer. As the parameters tell us, the lowest-energy conformations identified by multicanonical simulation and ELP minimization show significant structural differences in both models.

Answering the question of metastability is strongly related with the problem of identifying the folding path or an appropriate parametrization of the free energy landscape. Due to hidden barriers, it is practically impossible that ordinary stand-alone multicanonical sampling will be able to achieve this for such relatively long sequences. Note that, in our model I multicanonical simulation for the 55mer, we precisely sampled the density of states over *120 orders of magnitude*, i.e., the probability for finding randomly the lowest-energy conformation (that we identified with multicanonical sampling) in the conformational space is even less than 10^{-120}! In Fig. 2, we show two views of the global energy minimum conformation with energy $E_{min} \approx -42.4$ for the 55mer found by applying the ELP minimizer to AB model I. The hydrophobic core is tubelike and the chain forms a helical structure (which is here an intrinsic property of the model and is not due to hydrogen-bonding obviously being not supplied by the model).

B. Thermodynamic properties of heteropolymers

Our primary interest of this study is focused on thermodynamic properties of heteropolymers, in particular on con-
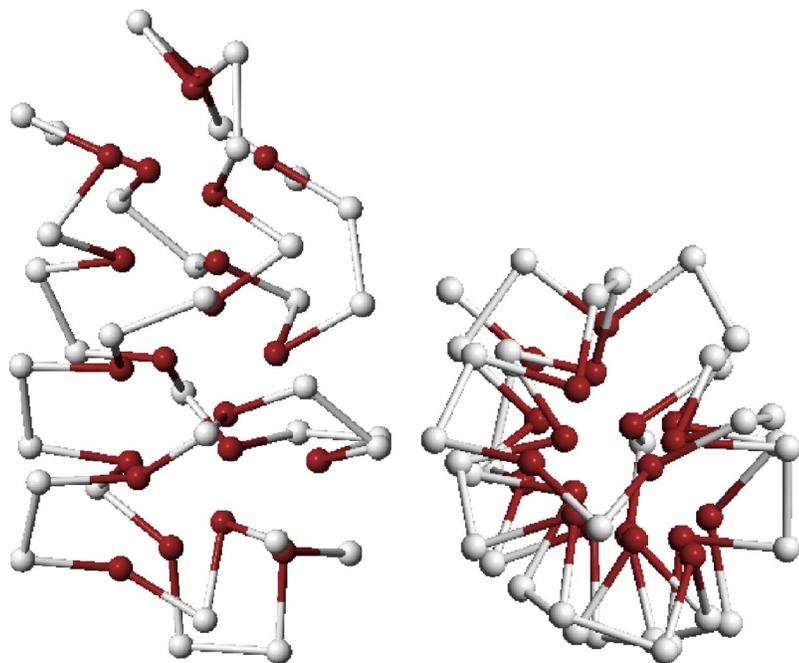

FIG. 2. (Color online) Side (left) and top view (right) of the global energy minimum conformation of the 55mer (AB model I) found with the ELP minimization algorithm (dark spheres: hydrophobic monomers—A, light spheres: hydrophilic—B).

TABLE IV. Sequences used in the study of thermodynamic properties of heteropolymers as introduced in Ref. [7]. The number of hydrophobic monomers is denoted by no. A.

No.	sequence	no. A
20.1	$BA_6BA_4BA_2BA_2B_2$	14
20.2	$BA_2BA_4BABA_2BA_5B$	14
20.3	$A_4B_2A_4BA_2BA_3B_2A$	14
20.4	$A_4BA_2BABA_2B_2A_3BA_2$	14
20.5	$BA_2B_2A_3B_3ABABA_2BAB$	10
20.6	$A_3B_2AB_2ABAB_2ABABABA$	10

formational transitions heteropolymers pass from random coils to native conformations with compact hydrophobic core. We investigated six heteropolymers with 20 monomers as studied by Irbäck *et al.* in Ref. [7] with AB model II. The associated sequences are listed in Table IV. Notice that the hydrophobicity (=no. A monomers in the sequence) is identical (=14) for the first four sequences 20.1–20.4, while sequences 20.5 and 20.6 possess only 10 hydrophobic residues. In Ref. [7], the thermodynamic behavior of these sequences was studied by means of the simulated tempering method. For revealing low-temperature properties, an additional quenching procedure was performed. In our simulations, we used multicanonical sampling over the entire range of temperatures without any additional quenching.

1. Multicanonical sampling of heteropolymers with 20 monomers

As described in Sec. IV A, the multicanonical method is capable of finding even the lowest-energy states without any biasing or quenching. We proved this also for the 20mers of Table IV by comparing once more with the ELP minimization method. In Tables V and VI, we present the estimates of the global energy minima in both models we found in these simulations. Once more, the values obtained with multicanonical sampling agree pretty well with those from ELP minimization. The respective structural coincidences are confirmed by the values for the root mean square deviation

and the overlap also being given in these tables. In order to identify conformational transitions, we calculated the specific heat $C_V(T) = (\langle E^2 \rangle - \langle E \rangle^2)/k_B T^2$ with $\langle E^k \rangle = \Sigma_E g(E) E^k \exp(-E/k_B T)/\Sigma_E g(E)\exp(-E/k_B T)$ from the density of states $g(E)$. The density of states was found (up to an unimportant overall normalization constant) by reweighting the multicanonical energy distribution obtained with multicanonical sampling to the canonical distribution $P^{\mathrm{can},T}(E)$ at infinite temperature ($\beta \equiv 1/k_B T = 0$), since $g(E) = P^{\mathrm{can},\infty}(E)$. Figure 3 shows, as an example, the density of states $g(E)$ (normalized to unity) and the multicanonical histogram for sequence 20.1 simulated with AB model II. We sampled conformations with energy values lying in the interval $[-60.0, 50.0]$, discretized in bins of size 0.01, and required the multicanonical histogram to be flat for at least 70% of the core of this interval, i.e., within the energy range $[-43.5, 33.5]$. Within and partly beyond this region, we achieved almost perfect flatness, i.e., the ratios between the mean and maximal histogram value, $H_{\mathrm{mean}}/H_{\mathrm{max}}$, as well as the ratio between minimum and mean, $H_{\mathrm{min}}/H_{\mathrm{mean}}$, exceeded 0.9. As a consequence of this high-accurate sampling of the energy space, we were able to calculate the density of states very precisely over about 70 orders of magnitude. The energy scale is, of course, bounded from below by the ground-state energy E_{min}, and the fact that we closely approximated this value can be seen by the strong decrease of the logarithm of the density of states for the lowest energies. This strong decrease of the density of states curves near the ground-state energy is common to all short heteropolymer sequences studied. It reflects the isolated character of the ground state within the energy landscape.

We used the density of states to calculate the specific heats of the 20mers in both models. The results are shown in Fig. 4. A first observation is that the specific heats obtained from simulations of AB model I show up two distinct peaks with the low-temperature peak located at $T_C^{(1)}$ and the high-temperature peak at $T_C^{(2)}$ compiled in Table V. The AB model II, on the other hand, favors only one pronounced peak at T_C with a long-range high-temperature tail. In Table VI, we compare our peak temperatures T_C with the values given in Ref. [7]. They are in good correspondence.

TABLE V. Minimal energies and temperatures of the maximum specific heats for the six 20mers of Table IV using AB model I as obtained by multicanonical sampling. The global maximum of the respective specific heats is indicated by a star ($\star$). For comparison, we have also given the globally minimal energies found from minimization with ELP as well as the respective root mean square deviation D_{rms} and the structural overlap parameter Q of the corresponding minimum energy conformations.

No.	$E_{\mathrm{min}}^{\mathrm{MUCA}}$	$E_{\mathrm{min}}^{\mathrm{ELP}}$	D_{rms}	Q	$T_C^{(1)}$	$T_C^{(2)}$
20.1	−33.766	−33.810	0.048	0.954	0.27(3)$\star$	0.61(5)
20.2	−33.920	−33.926	0.015	0.992	0.26(4)$\star$	0.69(4)
20.3	−33.582	−33.578	0.025	0.990	0.25(3)$\star$	0.69(3)
20.4	−34.496	−34.498	0.030	0.985	0.26(3)	0.66(2)$\star$
20.5	−19.647	−19.653	0.017	0.988	0.15(2)	0.41(1)$\star$
20.6	−19.322	−19.326	0.047	0.989	0.15(2)[a]	0.35(1)$\star$

[a]Specific heat of sequence 20.6 possesses only one maximum at $T_C^{(2)} \approx 0.35$. The value given for $T_C^{(1)}$ belongs to the pronounced turning point.

TABLE VI. Same as Table V, but using AB model II. For comparison, the specific-heat maximum locations T_s, estimated in Ref. [7], are also given.

No.	E_{min}^{MUCA}	E_{min}^{ELP}	D_{rms}	Q	T_C	T_s [7]
20.1	−58.306	−58.317	0.006	0.999	0.35(4)	0.36
20.2	−58.880	−58.914	0.009	0.997	0.33(4)	0.32
20.3	−59.293	−59.338	0.010	0.997	0.29(3)	0.30
20.4	−59.068	−59.079	0.007	0.998	0.27(4)	0.27
20.5	−51.525	−51.566	0.012	0.998	0.33(5)	0.33
20.6	−53.359	−53.417	0.014	0.996	0.25(2)	0.26

The sequences considered here are very short and the native fold contains a single hydrophobic core. Interpreting the curves for the specific heats in Fig. 4 in terms of conformational transitions, we conclude that the heteropolymers simulated with AB model I tend to form, within the temperature region $T_C^{(1)} < T < T_C^{(2)}$, intermediate states (often also called traps) comparable with globules in the collapsed phase of polymers. For sequences 20.5 and, in particular, 20.6, the smaller number of hydrophobic monomers causes a much sharper transition at $T_C^{(2)}$ than at $T_C^{(1)}$ (where, in fact, the specific heat of sequence 20.6 possesses only a turning point). The pronounced transition near $T_C^{(2)}$ is connected with a dramatic change of the radius of gyration, as can be seen later in Fig. 7, indicating the collapse from stretched to highly compact conformations with decreasing temperature. The conformations dominant for high temperatures $T > T_C^{(2)}$ are random coils, while for temperatures $T < T_C^{(1)}$ primarily conformations with compact hydrophobic core are favored. The intermediary globular "phase" is not at all present for the exemplified sequences when modeled with AB model II, where only the latter two "phases" can be distinguished. We have to remark that what we denote as "phases" are not phases in the strict thermodynamic sense, since for heteropolymers of the type we used in this study (this means we are not focused on sequences that have special symmetries, as, for example diblock copolymers $A_n B_m$), a thermodynamic limit is *in principle* nonsensical. Therefore conformational transitions of heteropolymers are not true phase transitions. As a conse-

quence, fluctuating quantities, for example the derivatives with respect to the temperature of the mean radius of gyration, $d\langle R_{gyr}\rangle / dT$, and the mean end-to-end distance, $d\langle R_{ee}\rangle / dT$, do not indicate conformational activity at the same temperatures, as well as when compared with the specific heat.

We conclude that conformational transitions of heteropolymers happen within a certain interval of temperatures, not at a fixed critical temperature. This is a typical finite-size effect and, for this reason, the peak temperatures $T_C^{(1)}$ and $T_C^{(2)}$ (for model I) and T_C (for model II) defined above for the specific heat are only representatives for the entire intervals. In order to make this more explicit, we consider sequence 20.3 in more detail. In Figs. 5(a) (for AB model I) and 5(b) (AB model II), we compare the energetic fluctuations (in form of the specific heat C_V) with the respective fluctuations of radius of gyration and end-to-end distance, $d\langle R_{ee,gyr}\rangle / dT = (\langle R_{ee,gyr}E\rangle - \langle R_{ee,gyr}\rangle\langle E\rangle) / k_B T^2$. Obviously, the temperatures with maximal fluctuations are not identical and the shaded areas are spanned over the temperature intervals, where the strongest activity is expected. We observe for this example that in model I [Fig. 5(a)], two such centers of activity can be separated, linked by an intermediary interval of globular traps. In fact, there is a minimum of the specific heat at T_C^{min} between $T_C^{(1)}$ and $T_C^{(2)}$, but the height of the barrier at $T_C^{(2)}$ is rather small and the globules are not very stable. From Fig. 5(b), we conclude for this sequence that in model II no peculiar intermediary conformations occur, and the folding of the heteropolymer is a one-step process.

It is widely believed and experimentally consolidated that realistic short single-domain proteins are usually two-state folders [13]. This means, there is only one folding transition and the protein is either in the folded or an unfolded (or denatured) state. Therefore, AB model II could indeed serve as a simple effective model for two-state heteropolymers. The main difference between the two models under study is that model II contains an implicit torsional energy which is not present in model I. This is in correspondence with more knowledge-based Gō-like models with explicit torsional energy contributions for the study of small proteins with known typical two-state folding-unfolding kinetics [14].

Nonetheless, there are also examples of small peptides exhibiting two clear peaks in the specific heat. In Ref. [29], the artificial peptide $Ala_{10}Gly_5Ala_{10}$ was studied in detail and it turned out that two transitions separate the ground-state
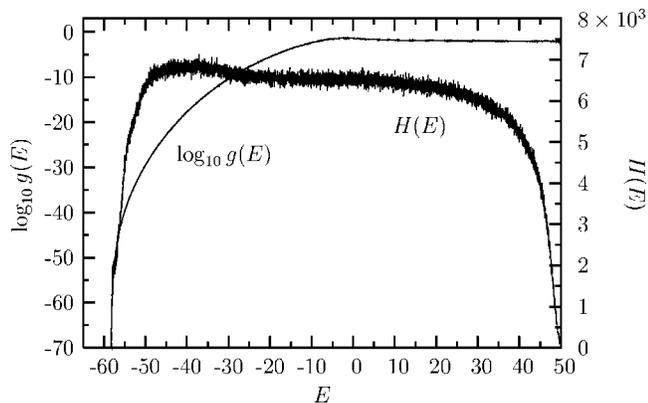

FIG. 3. Density of states $g(E)$ (normalized to unity over the plotted energy interval) and flat multicanonical histogram $H(E)$ for sequence 20.1 (AB model II).
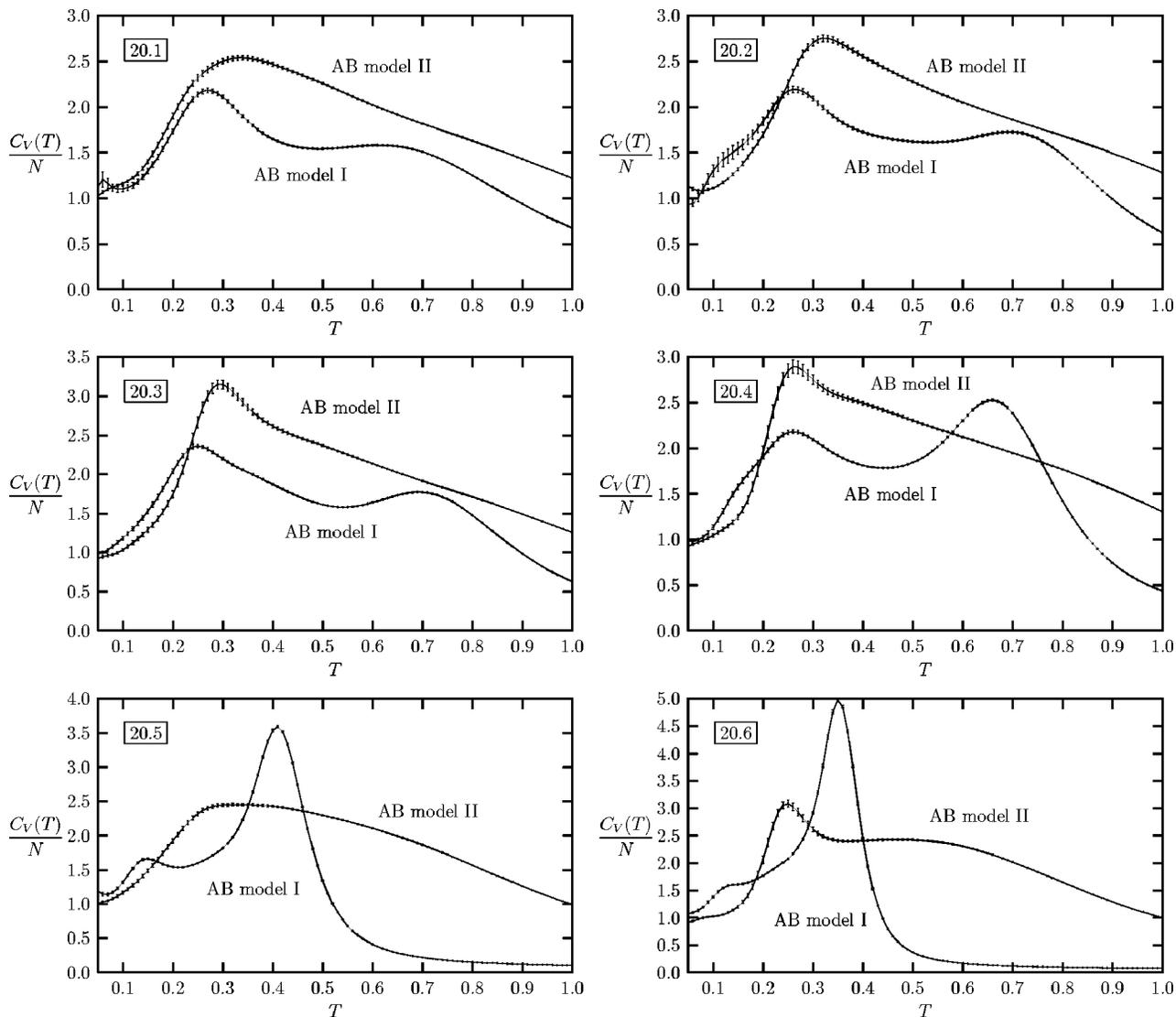
FIG. 4. Specific heats of the 20mers listed in Table IV.

conformation and random coil states. One is the alanine mediated helix-coil transition and the second the formation of a glycine hairpin that leads to a more compact conformation.

2. Comparison with a homopolymer

It is also interesting to compare the thermodynamic behavior of the heteropolymers with 20 monomers as described in the previous section with the homopolymer consisting of 20 A-type monomers, A_{20}. This is the consequent off-lattice generalization of self-avoiding interacting walks on the lattice (ISAW) that have been extensively studied over the past decades. In contrast to heteropolymers, where, because of the associated sequence of finite length, a thermodynamic limit does not exist, homopolymers display a characteristic tricritical second-order phase transition between random coil conformations ("good solvent") and compact globules ("poor solvent"), the so-called Θ transition [30].

In this study, our interest was not focused on the Θ transition but more on the direct comparison of the finite-size homopolymer and the different heteropolymer sequences. In

Fig. 6, we have plotted the specific heats of this homopolymer for both models under study. The first observation is that, independent of the model, the collapse from random coils (high temperatures) to globular conformations (low temperatures) happens, roughly, in one step. There is only one energetic barrier, as indicated by the single peak of the specific heats.

In Fig. 7, we have plotted for both models the mean radii of gyration as a function of the temperature for the sequences from Table IV in comparison with the homopolymer. For all temperatures in the interval plotted, the homopolymer obviously takes more compact conformations than the heteropolymers, since its mean radius of gyration is always smaller. This different behavior is an indication for a rearrangement of the monomers that is particular for heteropolymers: the formation of the hydrophobic core surrounded by the hydrophilic monomers. Since the homopolymer trivially also takes in the ground state a hydrophobic core conformation (since it only consists of hydrophobic monomers), which is obviously more compact than the complete conformations
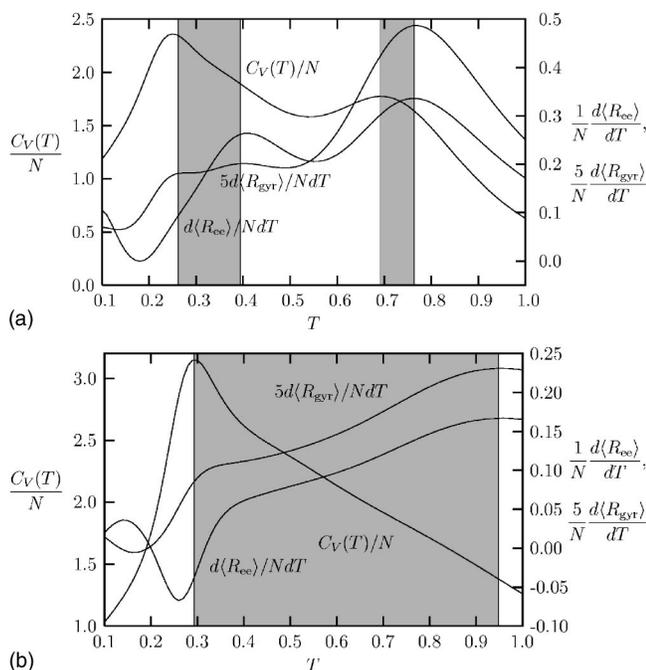
FIG. 5. Fluctuations of energy (specific heat), radius of gyration, and end-to-end distance for sequence 20.3 from simulations with (a) AB model I and (b) AB model II.

FIG. 7. Mean radius of gyration $\langle R_{\mathrm{gyr}} \rangle$ as a function of the temperature T for the sequences 20.1–20.4 (solid curves), 20.5, 20.6 (long dashes), and the homopolymer A_{20} (short-dashed curve) for (a) AB model I and (b) AB model II.

of the heteropolymers, we conclude that hydrophobic monomers weaken the compactness of low-temperature conformations. Thus, homopolymers and heteropolymers show a different "phase" behavior in the dense phase. Homopolymers fold into globular conformations which are hydrophobic cores with maximum number of hydrophobic contacts. Heteropolymers also form very compact hydrophobic cores which are, of course, smaller than that of the homopolymer due to the smaller number of hydrophobic monomers in the sequence. In total, however, heteropolymers are less compact than homopolymers because the hydrophilic monomers are pushed off the core and arrange themselves in a shell around the hydrophobic core. For model I, we also see in Fig. 7(a) a clear tendency that the mean radius of gyration and thus the compactness strongly depends on the hydrophobicity of the sequence, i.e., the number of hydrophobic monomers. The curves for sequences 20.5 and 20.6 (long-dashed curves)
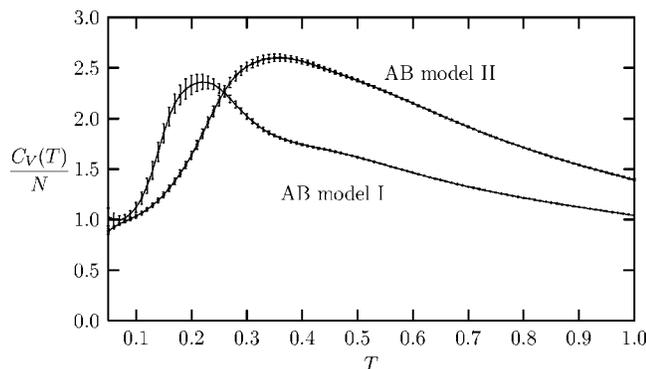

FIG. 6. Specific heats of the homopolymer A_{20} with 20 monomers for both models.
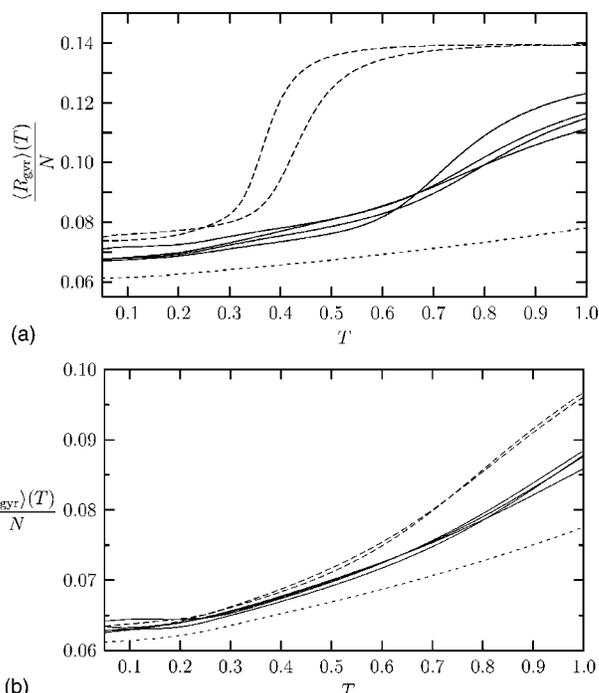
with 10 A's in the sequence can clearly be separated from the other heteropolymers in the study (with 14 hydrophobic monomers) and the homopolymer. This supports the assumption that for heteropolymers, the formation of a hydrophobic core is more favorable than the folding into an entire maximally compact conformation.

V. SUMMARY

We have investigated two coarse-grained off-lattice heteropolymer models of AB type that mainly differ in the modeling of energetic bending and torsional barriers. While the original AB model (model I) [6], which was first introduced for two-dimensional heteropolymers, treats the polymer as a stiff chain of hydrophobic (A) and hydrophilic (B) monomers without considering torsional barriers, model II favors bending, and an additional contribution containing also torsional energy is regarded [7]. Another noticeable property of model I is that contacts between A and B monomers are always suppressed, in contrast to model II.

We studied these models by means of the multicanonical Monte Carlo method. In a first test, we checked the ability of the algorithm to find lowest-energy conformations. The results were compared with minimum energy values obtained with the energy-landscape paving (ELP) algorithm by measuring the root mean square deviation and a generalized overlap parameter that in addition to torsional degrees of freedom also allows the comparison of bond angles. We found very good coincidences for minimum energies and associated conformations for all sequences under study, with

the exception of the 34mer in model II and the 55mer in both models. In the latter case, the random walk in the energy space, which is considered as the system parameter, in not sufficient to find the global energy minimum, and a more detailed study of the origin of the free energy barriers, i.e., the identification of an appropriate order parameter, is required. Nonetheless, for all sequences we obtained much lower values for the respective putative global-energy minimum than formerly quoted in the literature using an off-lattice chain-growth algorithm [16] and also considerably lower values than with the annealing contour Monte Carlo (ACMC) method [17].

Our main objective was the comparison of the conformational transitions in both models. We primarily studied energetic and conformational fluctuations of several sequences with 20 monomers and found that in model I there is a general tendency that, independent of the sequence, the folding from random coils (high temperatures) to lowest-energy conformations is a two-step process and the folding is slowed down by weakly stable intermediary conformations. This is different in model II, where traps are avoided and the heteropolymers exhibit a two-state folding behavior. This is (qualitatively) in correspondence with the observation that many short peptides seem to possess a rather smooth free-energy landscape, where only a single barrier separates unfolded states and native fold [13]. Most of the previous investigations in this regard were performed by means of a kind of knowledge-based potentials, where topological properties (e.g., the native contacts) of the folded state explicitly enter. These potentials allow then quantitative studies of the dynamics of the folding process for the specified protein. In our study, however, we were more interested in the influence of basic principles on the folding transition, and therefore a quantitative comparison with the folding kinetics of realistic proteins is not appropriate at this level of abstraction.

In order to reveal folding properties specific to heteropolymers, we also compared with a purely hydrophobic polymer. Our results for the homopolymer are in accordance with the widely accepted view of the formation of compact globular conformations below the Θ transition temperature. The globules, which are in our interpretation only compact hydrophobic cores with maximum number of hydrophobic contacts, minimize the surface exposed to the (implicit) solvent. This behavior was observed for both models which differ, for the purely hydrophobic homopolymer, only in local covalent bond properties, but not in the nonbonded interaction between the monomers. Since heteropolymers with the same chain length contain less hydrophobic monomers, the hydrophobic core is, although still very dense, smaller and the hydrophilic monomers form a shell surrounding the core and screen it from the solvent. This, in consequence, leads to a native conformation that qualitatively differs from the typical globules known from polymers. In models allowing for intermediary states (such as model I in our study), the globular "phase" is actually present for heteropolymers, too, and is dominant in a temperature interval between the hydrophobic-core phase at very low temperatures and the random coils that are present at high temperatures.

In conclusion, simple, coarse-grained off-lattice heteropolymer models without specific or knowledge-based parametrization are attractive as they allow for the study of how basic energetic contributions, bonded and nonbonded, in the model are related to and influence the conformational folding process of heteropolymers, respectively. Herein, the main focus is on the study of general properties of these systems. In perspective, such models, after refinements with respect to a few specific interatomic interactions (e.g., hydrogen bonds), will be suitable for studying heteropolymers of a few hundreds of monomers even quantitatively without the requirement of a prior knowledge of the final fold.

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997); J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).

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

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

[4] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003); Phys. Rev. E **68**, 021113 (2003).

[5] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003); J. Chem. Phys. **120**, 6779 (2004).

[6] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993); F. H. Stillinger and T. Head-Gordon, *ibid.* **52**, 2872 (1995).

[7] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J.

Chem. Phys. **107**, 273 (1997).

[8] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).

[9] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 1273 (1999).

[10] T. R. Weikl and K. A. Dill, J. Mol. Biol. **332**, 953 (2003).

[11] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).

[12] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).

[13] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).

[14] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).

[15] H. Kaya and H. S. Chan, Proteins **52**, 510 (2003).

[16] H.-P. Hsu, V. Mehra, and P. Grassberger, Phys. Rev. E **68**, 037703 (2003).

[17] F. Liang, J. Chem. Phys. **120**, 6756 (2004).

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

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

[20] U. H. E. Hansmann and Y. Okamoto, J. Comput. Chem. **14**, 1333 (1993).

[21] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).

[22] H. Arkın and T. Çelik, Eur. Phys. J. B **30**, 577 (2002).

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

[24] W. Janke, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, Proceedings of the Euro Winter School, edited by J. Grotendorst, D. Marx, and A. Muramatsu, John von Neumann Institute for Computing, NIC Series (NIC-Directors, Forschungszentrum Jülich, Jülich, 2002), Vol. 10, p. 423.

[25] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).

[26] S. K. Kearsley, Acta Crystallogr., Sect. A: Found. Crystallogr. **45**, 208 (1989).

[27] U. H. E. Hansmann, M. Masuya, and Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. **94**, 10652 (1997).

[28] B. A. Berg, H. Noguchi, and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

[29] N. A. Alves and U. H. E. Hansmann, J. Chem. Phys. **117**, 2337 (2002); J. Phys. Chem. B **107**, 10284 (2003).

[30] P. D. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).