Multicanonical Chain-Growth Algorithm

Michael Bachmann* and Wolfhard Janke†

Institut für Theoretische Physik, Universität Leipzig, Augustusplatz 10/11, D-04109 Leipzig, Germany

(Received 27 April 2003; published 14 November 2003)

We present a temperature-independent Monte Carlo method for the determination of the density of states of lattice proteins that combines the fast ground-state search strategy of the new pruned-enriched Rosenbluth chain-growth method and multicanonical reweighting for sampling the complete energy space. Since the density of states contains all energetic information of a statistical system, we can directly calculate the mean energy, specific heat, Helmholtz free energy, and entropy for all temperatures. We apply this method to lattice proteins consisting of hydrophobic and polar monomers, and for the examples of sequences considered, we identify the transitions between native, globule, and random coil states. Since no special properties of heteropolymers are involved in this algorithm, the method applies to polymer models as well.

DOI: 10.1103/PhysRevLett.91.208105 PACS numbers: 87.15.Aa, 05.10.Ln, 87.15.Cc

The simulation of protein folding is extremely challenging, since the interactions between the constituents of the macromolecule and the influence of the environment require sophisticated models. One of the most essential aspects in the description of the folding process is the formation of a compact core of hydrophobic amino acid residues (H) which is screened from water by hydrophilic or polar residues (P). This characteristic property of realistic proteins can qualitatively be studied with simple lattice models such as the HP model [1]. By taking into account the attractive interaction between hydrophobic monomers only, the energy of a lattice protein with certain conformation and sequence is calculated as follows:

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

where $\langle i, j < i-1\rangle$ symbolizes that the sum is taken only over nearest lattice neighbors being nonadjacent along the self-avoiding chain of monomers. If the ith monomer is hydrophobic, $\sigma_i = 1$, while for a polar monomer $\sigma_i = 0$.

As it is one of the main goals in off-lattice simulations to find low-lying energy states within a rough free energy landscape, good lattice folders are expected to have ground states with low degeneracy. Much work has been done on identifying designing sequences with such native states. Ground-state search strategies for lattice models range, for example, from enumeration [2,3] over hydrophobic core construction [4,5] and contact interaction [6] to chain-growth methods [7–10]. Low-lying energy states for HP sequences with up to 136 monomers were identified with these methods.

In contrast, there were only a few attempts to study the thermodynamic properties of the HP model in three dimensions [11]. The main reason is that conventional Monte Carlo methods like Metropolis sampling, but also more sophisticated methods like simulated [12] and parallel [13] tempering as well as histogram reweighting Monte Carlo algorithms such as multicanonical sampling [14] or the Wang-Landau method [15] expose problems in tackling "hidden" conformational barriers in combination with chain update moves which usually become inefficient at low temperatures, where many attempted moves are rejected due to the self-avoidance constraint. One possibility to update the conformation is to apply move sets. Widely used sets usually consist of operations that change a single bond (end flips), two bonds (corner flips), three (crankshaft) or even more bonds, and pivot rotations.

Alternatively, it is possible to let the polymer grow; i.e., the nth monomer is placed at a randomly chosen next-neighbor site of the $(n-1)$th monomer ($n \le N$ with N being the total length of the polymer). If this new site is already occupied, the entire chain would have to be discarded to obtain correct statistics. This simple chain growth is not efficient, since the number of discarded chains grows exponentially with the chain length. Rosenbluth chain growth [16] avoids occupied neighbors at the expense of a bias, since the probability of such a chain is $p_n \sim (\prod_{l=2}^{n} m_l)^{-1}$, where m_l is the number of free lattice sites to place the lth monomer. This bias is balanced out by assigning each conformation a Rosenbluth weight $W_n^{\mathrm{R}} \sim p_n^{-1}$. Chain-growth methods with population control such as PERM (pruned-enriched Rosenbluth method) [8,9] and its recent modifications nPERM$_{\mathrm{is}}^{\mathrm{ss}}$ [10] improve this procedure considerably by utilizing the counterbalance between Rosenbluth weight and Boltzmann probability. The weight factor W_n^{R} is therefore replaced by

$$W_n^{\mathrm{PERM}} = \prod_{l=2}^{n} m_l e^{-(E_l - E_{l-1})/k_B T}, \tag{2}$$

$$2 \le n \le N \qquad (E_1 = 0, W_1^{\mathrm{PERM}} = 1),$$

where E_l is the energy of the partial chain $\mathbf{X}_l =$

$(\mathbf{x}_1, \ldots, \mathbf{x}_l)$ created with Rosenbluth chain growth and T is the temperature.

To explain the main ideas, we shall confine ourselves for the moment to the original PERM formulation [8], where the sample of chains of length n is enriched by making identical copies once W_n^{PERM} is bigger than a certain threshold value $W_n^{>}$. In this case, the weight W_n^{PERM} is divided among the clones. For W_n^{PERM} being smaller than a lower bound $W_n^{<}$, the chain is pruned with probability $1/2$ and the weight of a surviving chain is doubled. The partition sum is proportional to the sum of weight factors (2) for the conformations $\mathbf{X}_{n,t}$ of length n sampled at "time" t,

$$Z_n \sim \sum_t W_n^{\mathrm{PERM}}(\mathbf{X}_{n,t}). \tag{3}$$

The PERM algorithms are very successful as ground-state searchers and the canonical distribution at a given temperature T is well reproduced over some orders of magnitude, but states that are highly suppressed at this temperature are not hit in a reasonable time. Standard reweighting techniques are applicable only in a small region around T. Thus, recording temperature-dependent quantities such as the specific heat requires simulations at different temperatures.

As the partition sum of a polymer or a heteropolymer with a fixed sequence can be expressed in terms of the density (or degeneracy) of states $g(E)$, $Z = \sum_{\{\mathbf{x}\}} e^{-\beta E(\{\mathbf{x}\})} = \sum_i g(E_i) e^{-\beta E_i}$ ($\beta \equiv 1/k_B T$), all energetic quantities such as the mean energy $\langle E \rangle(T) = -(\partial/\partial\beta)\ln Z$, specific heat $C_V(T) = (\langle E^2 \rangle - \langle E \rangle^2)/k_B T^2$, Helmholtz free energy $F(T) = -k_B T \ln Z$, and entropy $S(T) = (\langle E \rangle - F)/T$ can directly be calculated if $g(E)$ is known. These quantities are of particular interest, since they are indicators of temperature-dependent conformational transitions.

Our method allows within one simulation the direct sampling of the density of states $g(E)$ over the entire range of the energy space with probabilities ranging over many orders of magnitude, due to combining the advantages of energetic flat histogram reweighting and chain growth. The flat distribution can then be reweighted to any desired temperature. Rare, i.e., low-lying energy states are also hit, and therefore the low-temperature behavior of the polymer can be reproduced well, in particular, the low-temperature transition between compact globules and ground states of lattice proteins with low ground-state degeneracy. Using the HP model, we applied the method to lattice proteins with more than 40 monomers and different ground-state degeneracies and found for examples with low ground-state degeneracy pronounced low-temperature peaks in the specific heat indicating ground-state–globule transitions. Since our method is completely general, it is also applicable to other polymer models.

In order to achieve a flat distribution of energetic states using chain growth, we introduce into the partition sum (3) an additional weight $W_n^{\mathrm{flat}}(E_n(\mathbf{X}_n))$ that depends on the energy E_n of a given conformation $\mathbf{X}_n$:

$$Z_n \sim \sum_t W_n^{\mathrm{PERM}}(\mathbf{X}_{n,t}) W_n^{\mathrm{flat}}(E_n(\mathbf{X}_{n,t}))[W_n^{\mathrm{flat}}(E_n(\mathbf{X}_{n,t}))]^{-1}. \tag{4}$$

Since the histograms at all intermediate stages of the chain-growth process are required to be flat, the new reweighting factor is rewritten in product form and we have

$$Z_n \sim \sum_t [W_n^{\mathrm{flat}}(E_n)]^{-1} \prod_{l=2}^{n} m_l e^{-(E_l - E_{l-1})/k_B T} \frac{W_l^{\mathrm{flat}}(E_l)}{W_{l-1}^{\mathrm{flat}}(E_{l-1})} \tag{5}$$

with $W_1^{\mathrm{flat}} = 1$. The PERM weight factors (2) lead to a canonical distribution $P_n^{\mathrm{can},T}(E_n)$ which shall be deformed to a constant distribution $P_n^{\mathrm{flat},T}(E_n)$ over the entire energy space. This requires the weights W_n^{flat} to be proportional to the inverse of the canonical distribution, $W_n^{\mathrm{flat}} \sim 1/P_n^{\mathrm{can},T}(E_n)$, a condition that can obviously be satisfied only iteratively [14]. As we are mainly interested in the density of states, which is proportional to the canonical probability distribution at $\beta = 1/k_B T = 0$, $g_n(E_n) \sim P_n^{\mathrm{can},\infty}(E_n)$, it is convenient (but inessential) to choose $\beta = 0$ in the multicanonical formulation. Consequently, $W_n^{\mathrm{flat}} \sim 1/g_n(E_n)$ and $Z_n \sim \sum_t g_n(E_n(\mathbf{X}_{n,t})) W_n(\mathbf{X}_{n,t})$. Here we have introduced the combined weight

$$W_n(\mathbf{X}_n) = \prod_{l=2}^{n} m_l \frac{g_l^{-1}(E_l)}{g_{l-1}^{-1}(E_{l-1})}, \qquad W_1 = g_1 = 1, \tag{6}$$

which can also be written recursively, $W_n = W_{n-1} m_n g_n^{-1}(E_n)/g_{n-1}^{-1}(E_{n-1})$.

The most important technical part of the algorithm is the determination of the weights W_n^{flat}, since they are directly connected with the desired densities of states $g_n(E)$. As the weights are completely unknown in the beginning, we evaluate them iteratively, starting from unity, $W_n^{\mathrm{flat},(0)}(E) = 1$ $(2 \le n \le N)$ for all values of E. This means that the zeroth iteration is a pure chain-growth run at infinite temperature without reweighting.

Each time, a chain of length n with energy E is created, the corresponding histogram value $H_n(E)$ is increased by the weight W_n of the chain. This weight is used to decide about enriching the sample (if $W_n \ge W_n^{>}$), pruning ($W_n \le W_n^{<}$), or simply continuing the chain (if $W_n^{<} < W_n < W_n^{>}$). For updating the threshold values we apply similar rules as in Ref. [10], i.e., $W_n^{>} = C(Z_n^{\mathrm{flat}}/Z_1^{\mathrm{flat}})(c_n/c_1)^2$ and $W_n^{<} = 0.2 W_n^{>}$, where $Z_n^{\mathrm{flat}} = \sum_t W_{n,t}$ is an estimate for the partition function associated with the flat distribution $H_n(E)$. The number of created chains having length n is denoted by c_n. The parameter C controls the pruning-enrichment statistics.

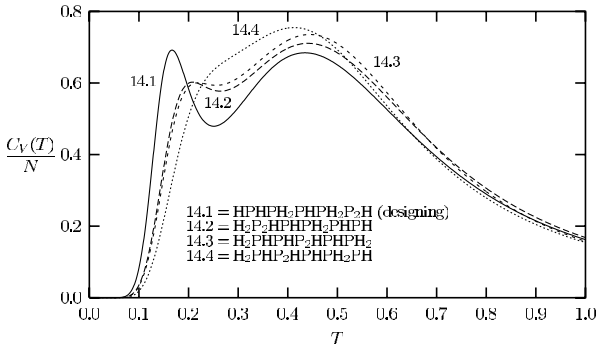

FIG. 1. Specific heat for four 14mers with different sequences from exact enumeration. Only the 14mer with the native ground state shows a pronounced low-temperature peak indicating a transition between ground states and globules. We have not included the simulation results because they are indistinguishable on the scale of this figure.

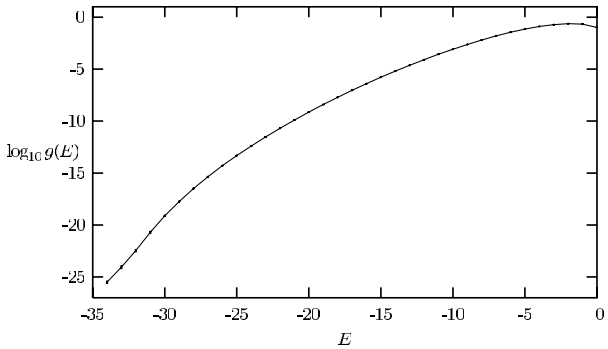

FIG. 2. Density of states of the 42mer, normalized to unity.

Once a certain number of chains with total length N has been produced, the iteration is finished and the new weights $W_n^{\mathrm{flat},(i)}(E)$ are determined by calculating $W_n^{\mathrm{flat},(i)}(E) = W_n^{\mathrm{flat},(i-1)}(E)/H_n(E)$, $2 \le n \le N$. Before the next iteration starts, the histogram is reset to $H_n(E) = 0$, and the threshold values are initialized to $W_n^{>} = \infty$ and $W_n^{<} = 0$. The weight iteration is stopped when $H_n(E)$ looks "flat," i.e., once it has approximately the same value for all energies. In our actual implementation we employed a suitably adapted multicanonical variant of nPERMis (new PERM with importance sampling) [10]. For the results presented here, 10–20 iterations were found to be sufficient to determine the multicanonical weight factors with reasonable accuracy. The number of conformations of total length N created in each iteration was chosen to be of order 10^5–10^6, except in the measuring run, where we accumulated statistics of up to 10^8 chains. We adjusted the pruning/enrichment control parameter to $C = 0.01$ such that less than 20% of the chains were pruned or enriched. On average 10 chains of total length N were generated per tour with this choice. Still more important for efficiency, in almost all started tours at least one such chain was created. The technical details of our implementation will be described elsewhere [17].

As a first example and for validation of the new algorithm, we discuss general properties of heteropolymers exemplified for 14mers. These results can still be compared with data obtained from exact enumeration. Among all 2^{14} sequences for 14mers there is only one designing sequence, i.e., a sequence to which a unique ground state belongs (up to a reflection symmetry). We compared thermodynamic properties of four 14mers with different sequences but the same hydrophobicity ($n_H = 8$) and identical lowest energy ($E_{\min} = -8$). Figure 1 shows the specific heat for the different 14mers. With a statistics of 10^8 chains in the production run of our algorithm, our curves cannot be distinguished from the exact ones at this level of resolution. A pronounced low-temperature peak that indicates the transition between ground states and compact globule states is observed only for the 14mer with the native ground state, 14.1. These results show qualitatively how the conformational transitions depend on the ground-state degeneracy of the polymer. For the sequences 14.2 and 14.3 it is twice that of the designing sequence, and 14.4 is even 4 times higher degenerate.

The next example to which we applied our multicanonical chain-growth algorithm is a 42mer with the sequence $PH_2PHPH_2PHPHP_2H_3PHPH_2PHPH_3P_2HPHPH_2PHPH_2P$ whose ground-state properties have similarities with the parallel β-helix of *pectate lyase C* [18]. The lattice model with only fourfold ground-state degeneracy has a ground-state energy of $E_{\min} = -34$. In order to investigate the low-temperature behavior of this system it is necessary that the algorithm correctly samples the low-energy states and that it also hits the ground states. The measured density of states ranges over about 25 orders of magnitude and covers the entire energy space $[-34, 0]$, as shown in Fig. 2. From the density of states it is straightforward to compute the specific heat and mean energy shown in Fig. 3 as well as the free energy and entropy [17]. Writing out the raw energies and weights from the simulation, we analyzed the data and calculated the statistical error by

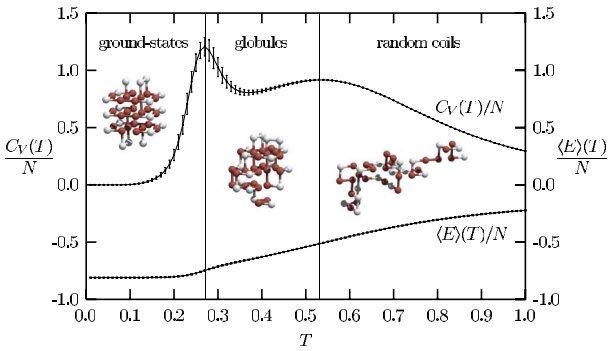

FIG. 3 (color online). Specific heat and mean energy as functions of the temperature for the 42mer. The ground-state–globule and the globule–random coil transition occurs at $T_0 \approx 0.27$ and $T_1 \approx 0.53$, respectively.

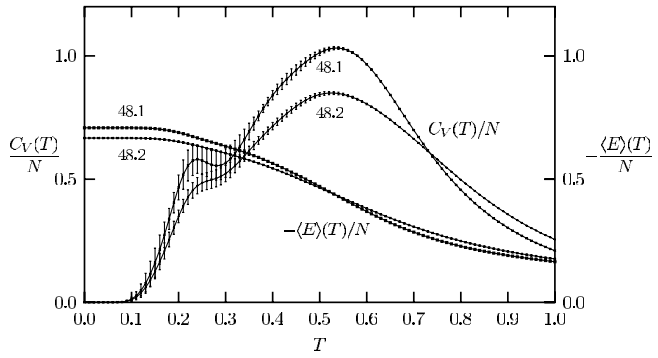

FIG. 4. Specific heat and mean energy for the two 48mers with different ground-state properties. The low-temperature conformational transition is more pronounced for the example 48.1 with lower ground-state degeneracy.

means of the jackknife blocking method. Because of the low degeneracy of the ground states, the transition between native states and globule states is very pronounced and occurs at a temperature $T_0 \approx 0.27$. The globule–random coil transition at $T_1 \approx 0.53$, on the other hand, is rather weak. This confirms the results of Ref. [11]. We also compared with results obtained from multihistogram reweighting of independent nPERMis runs at five different temperatures with overlapping canonical distributions. Consistent results were found, but the patching approach is much more cumbersome and turned out to be less accurate than our method [17].

Finally, similar to the consideration of the 14mers, we compare two 48mers with different ground-state properties. The first one, which we denote by 48.1, has the sequence $PHPH_2PH_6P_2HPHP_2HPH_2PHPHP_3HP_2H_2P_2H_2P_2HPHP_2HP$ and its ground state with the energy -34 is approximately 5000-fold degenerate. The ground state of the other 48mer (48.2) with the sequence $HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3HP_8H_2$ has a much higher degeneracy of order 10^6 and possesses the energy -32 [5,17]. As is demonstrated in Fig. 4, we also observe for these longer chains that the conformational transition between the lowest-energy states and the globules is stronger the lower the ground-state degeneracy is. Further applications to longer HP sequences with around 100 monomers also perform well [17] and gave, for instance, a new upper bound ($E_{\min} = -56$) for the ground-state energy of a designed 103mer [6,10].

In conclusion, we have developed a multicanonical chain-growth algorithm that allows the simulation of the thermodynamic properties of polymers and heteropolymers. It is based on energetic flat histogram sampling of the density of states in combination with PERM chain growth. For heteropolymers with more than 40 monomers accurate densities of states over more than 25 orders of magnitude were obtained that cover the entire energy range, thus yielding very good results for all derived energetic quantities such as mean energy, specific heat, free energy, and entropy. In particular, this enabled us to determine the low-temperature behavior of the systems with high precision and to observe pronounced low-temperature peaks of the specific heat for lattice proteins with low ground-state degeneracy indicating the ground-state–globule transition.

We thank Peter Grassberger and Hsiao-Ping Hsu for helpful discussions on the new PERM algorithms. This work is partially supported by the German-Israel-Foundation (GIF) under Contract No. I-653-181.14/1999.

*Email address: michael.bachmann@itp.uni-leipzig.de
†Electronic addresses: wolfhard.janke@itp.uni-leipzig.de; http://www.physik.uni-leipzig.de/CQT

[1] K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[2] C. Tang, Physica (Amsterdam) **288A**, 31 (2000); H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, C. Tang, and N. Wingreen, J. Chem. Phys. **116**, 352 (2002).
[3] A. Irbäck and E. Sandelin, J. Chem. Phys. **108**, 2245 (1998); A. Irbäck and C. Troein, J. Biol. Phys. **28**, 1 (2002).
[4] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993); Proc. Natl. Acad. Sci. U.S.A. **92**, 146 (1995).
[5] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shaknovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[6] L. Toma and S. Toma, Protein Sci. **5**, 147 (1996).
[7] T. C. Beutler and K. A. Dill, Protein Sci. **5**, 2037 (1996).
[8] 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).
[9] 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.
[10] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003).
[11] G. Chikenji, M. Kikuchi, and Y. Iba, Phys. Rev. Lett. **83**, 1886 (1999), and references therein.
[12] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[13] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[14] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992); W. Janke, Physica (Amsterdam) **254A**, 164 (1998).
[15] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[16] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).
[17] M. Bachmann and W. Janke, e-print cond-mat/0310707.
[18] M. D. Yoder, N. T. Keen, and F. Jurnak, Science **260**, 1503 (1993).