Versatile Approach to Access the Low Temperature Thermodynamics of Lattice Polymers and Proteins

Thomas Wüst* and David P. Landau

Center for Simulational Physics, The University of Georgia, Athens, Georgia 30602, USA

(Received 18 December 2008; published 29 April 2009)

We show that Wang-Landau sampling, combined with suitable Monte Carlo trial moves, provides a powerful method for both the ground state search and the determination of the density of states for the hydrophobic-polar (HP) protein model and the interacting self-avoiding walk (ISAW) model for homopolymers. We obtain accurate estimates of thermodynamic quantities for HP sequences with >100 monomers and for ISAWs up to >500 monomers. Our procedure possesses an intrinsic simplicity and overcomes the limitations inherent in more tailored approaches making it interesting for a broad range of protein and polymer models.

DOI: 10.1103/PhysRevLett.102.178101 PACS numbers: 87.15.ak, 05.10.Ln, 05.70.Fh, 36.20.Ey

Coarse-grained polymer and protein models play an important role in understanding physical phenomena such as, e.g., protein folding or the phase behavior of flexible macromolecules, and Monte Carlo simulation methods have become an indispensable tool for the study of such models [1]. One of the most prominent examples is the hydrophobic-polar (HP) lattice model [2], where the protein is represented as a self-avoiding chain of beads (the amino acid residues) on a lattice. The amino acids are divided into two classes—hydrophobic (H) and polar (P)—and an attractive interaction ϵ acts between nonbonded neighboring H residues mimicking the hydrophobic force ($\epsilon_{\mathrm{HH}} = -1$, $\epsilon_{\mathrm{HP,PP}} = 0$). The special case of a chain consisting entirely of H residues (homopolymer), the interacting self-avoiding walk (ISAW), is an important model for studying the statistical physics of polymers [3,4].

Despite their formal simplicity and minimalistic framework, lattice models represent a challenging testing ground for computational methods because of their complex energy landscapes, conformational constraints, and dense packings. The HP model has become a standard for assessing the efficiency of folding algorithms, and numerous—some very tailored—conformational ground state search strategies have been proposed, see, e.g., [5–8] and references therein.

More revealing than algorithms that merely search for low energy states, however, are methods which target the sampling of the entire conformation and energy space. They can provide an estimate of the density of states (DOS) $g(E)$ of energy E which, in turn, gives access to thermodynamic properties (e.g., internal energy, specific heat, entropy, or free energy) of a system at any temperature [9]. Only a few attempts have been undertaken to this end for the HP model, the most notable approaches being multi-self-overlap ensemble (MSOE) Monte Carlo simulations [10], multicanonical chain growth (MCCG) [11], and equi-energy sampling (EES) [12]. Although inventive and powerful, these methods also suffer from severe limitations: Large memory needs for keeping track of all sampled conformations (construction of microcanonical ensembles) (EES); (quasi-) statics, i.e., one bead of the chain is permanently fixed in space (MCCG); or the necessity to treat an expanded ensemble resulting in a large amount of computer time spent in sampling nonphysical space (MSOE). Such restrictions can become increasingly important for more complex biological setups such as multichain systems or protein folding in heterogeneous environments (e.g., membranes) [13].

In this Letter, we show that a generic algorithm—Wang-Landau sampling [14]—together with appropriate Monte Carlo trial moves, provides a powerful, yet flexible methodology for the simulation of HP-like lattice proteins and homopolymers that does not suffer from any of the above limitations.

The key to our approach is the combination of two ''nontraditional'' Monte Carlo trial moves, which complement one another extremely well, namely, pull moves [6] and bond-rebridging moves [15], see Fig. 1. Originally proposed for the square and simple cubic lattices only, here we extended both types of trial moves to any n-dimensional space ($n \geq 2$). (i) Pull moves [6] allow for the close-fitting motion of a polymer chain within a confining environment by ''pulling'' portions of the polymer to unoccupied neighboring sites. Pull moves are reversible and fulfill ergodicity; moreover, they provide a good balance between local and global conformational changes, as well as a ''natural'' dynamics of folding. These features are important to an algorithm that seeks to sample the entire conformational space such as Wang-Landau sampling and thus requires an efficient move for the continuous folding

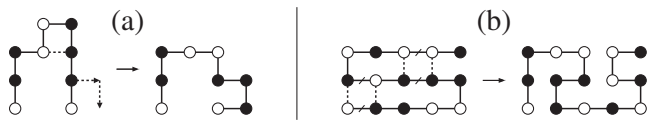

FIG. 1. Typical example of pull (a) and bond-rebridging (b) move in 2D. For details, see [6,15].

and unfolding of the polymer. (ii) Bond-rebridging moves [15]: Trial moves which displace monomers become ineffective for very compact conformations where few unoccupied neighboring sites remain available. In contrast, bond-rebridging moves allow the polymer to change its conformation even at highest densities by reordering bonds while leaving the positions of monomers unchanged. Moreover, they facilitate long-range topological changes, e.g., entanglement, which otherwise require costly unfolding and folding processes. This later feature becomes particularly important when the sampling of the DOS is split up into energy subintervals as it substantially reduces the risk of ''locking out'' conformational space [16]. During sampling, pull or bond-rebridging trial moves were selected randomly (usually with a 1:1 ratio for each type). Pull moves enabled us to sample the entire conformational space of long polymer chains which was not feasible with ''traditional moves'' only [3]. Furthermore, the combination of bond-rebridging and pull moves provided a speedup of a factor of 3 for the HP model and a factor of 10 for the ISAW as comparing with pull moves only.

Wang-Landau (WL) sampling is an efficient and robust algorithm for the computation of the DOS for diverse statistical physical systems, see [9,14] for details. To fulfill detailed balance in conjunction with pull moves, in this study, the WL transition probability from a state A to a state B has been generalized to

$$P(A \to B) = \min\left(1, \frac{g(E_A)}{g(E_B)} \frac{n_{B\to A}/n_B}{n_{A\to B}/n_A}\right). \quad (1)$$

$n_{A\to B}$ denotes the number of pull moves from A to B and n_A the total number of possible pull moves from A, $n_{B\to A}$, and n_B correspondingly (here, $n_{A\to B} = n_{B\to A}$ because of reversibility). Selecting only within the list of possible pull moves (n_A) also increases the dynamics for dense conformations as compared to a standard ''trial and error'' procedure. In order to yield accurate and reliable DOS estimates over the entire energy range (including the lowest energies), we used a very stringent parameter set for all our simulations, i.e., final modification factor $\ln(f_{\text{final}}) = 10^{-8}$ and flatness criterion $p = 0.8$; statistical errors were always calculated from 15 independent DOS estimates (by means of a jackknife analysis).

The knowledge of the exact energy range is essential in the WL algorithm for the examination of the flatness of the histogram. Often, however, energy boundaries are *a priori* unknown, (hence, the use of ground state search algorithms, e.g., for the HP model). To solve this dilemma, the following procedure proved to be most efficient: Every time a new energy level E_{new} is found, it is marked as ''visited'' and $g(E_{\text{new}})$ is set to $g_{\min}$, i.e., the minimum of g among all previously visited energy levels. The flatness of the histogram is checked for visited energy levels only. With this self-adaptive procedure, new regions of conformational space can be explored while, at the same time, the current DOS estimate is further refined.

First, we applied our procedure to various benchmark HP sequences found in the literature. Since heteropolymers with $N \leq 50$ monomers no longer represent a significant challenge and our results are in perfect agreement with previous works, we restrict our presentation to two longer sequences which turned out to be particularly demanding, namely, a 100mer in 2D (2D100b) and a 103mer in 3D (3D103); for definitions of HP sequences, see, e.g., [8]. The ground states of sequence 2D100b are believed to have an energy $E = -50$ [5,6,8,10]; however, previous attempts to obtain the DOS over the entire energy range $[-50, 0]$ within a single simulation have failed [10,12]. In contrast, with our approach, we were able to achieve this with high accuracy. Figure 2 (*top*) shows the resulting specific heat $C_V(T)/N$, depicting a peak at $T \approx 0.48$ (coil-globule transition) and a very weak shoulder at $T \approx 0.23$ (folding transition). Such two-step acquisition of the native state has been observed in studies of realistic protein models and is not restricted to lattice models. For sequence 3D103, the lowest energy found so far was -57, achieved only by fragment regrowth Monte Carlo simulations via energy-guided sequential sampling (FRESS) [8]. With our approach, we discovered an even lower state with energy -58. Moreover, we were also able to obtain the DOS in the energy range $[-57, 0]$, within a single simulation, and with very high accuracy. It was nonetheless not possible to determine the relative magnitudes of the ground state ($E = -58$) and first excited state ($E = -57$) DOS with high precision. Figure 2 (*bottom*) displays the specific heat for

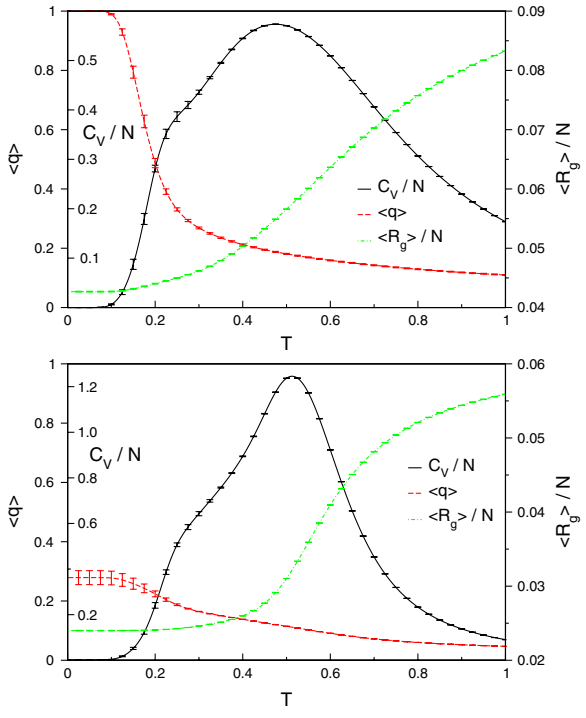

FIG. 2 (color online). Specific heat C_V/N, mean radius of gyration $\langle R_g\rangle/N$ (N, chain length), and mean Jaccard index $\langle q\rangle$ as a function of temperature (T) for HP sequence 2D100b (*top*) and 3D103 (*bottom*), respectively.

TABLE I. Energy minima found by several methods for benchmark HP sequences in 2D and 3D. The first column names the sequence (dimension and length), see [8]. In case of Wang-Landau sampling (WLS), numbers in parentheses denote that the DOS has been obtained down to this energy. Center dots mean no data available.

Seq.	WLS	EES	MCCG	MSOE	FRESS[b]	PERM[b]
2D100a	−48	−48	⋯	−47	−48	−48
2D100b	−50	−49	⋯	−50[c]	−50	−50
3D88	−72 (−69)	⋯	⋯	⋯	−72	−69
3D103	−58[a] (−57)	⋯	−56	⋯	−57	−55
3D124	−75 (−74)	⋯	⋯	⋯	−75	−71
3D136	−82 (−81)	⋯	⋯	⋯	−83	−80

[a]See, e.g., $dr_2u_2ldbdrubdblfldrbl_2f_2dr_2dbrulbr_2drf-rul_2dfurufdldflb_2ufuf_2rb_2urb_2rbluldfu_2fd_3brblbul_2brd-f_4lf_2dbld$ (encoded as sequence left[l], right[r], up[u], down[d], forward[f], backward[b]).
[b]Ground state search only (no DOS estimate).
[c]DOS not attained.

sequence 3D103, manifesting a peak at $T \approx 0.51$ and a shoulder at $T \approx 0.27$. We do not observe an additional peak in C_V at very low temperatures, contrarily to Ref. [11]. However, since only conformations with energies down to $E = -56$ were found and the estimated errors near that peak were rather large, we think that this finding was an artefact of insufficient sampling. Indeed, our C_V curves indicate that the folding transitions from unstructured globular conformations to the ground states are rather weak for both sequences—despite the difficulty in sampling their low energy regimes.

By means of multicanonical sampling given our DOS estimates, we obtained the radius of gyration R_g [3] and the Jaccard index $q = \max\{c_{s,g}/(c_{s,g} + c_s + c_g)|E_g = \min\}$ which measures the structural similarity between any conformation s and the ground states g of an HP sequence [17]. $c_{s,g}$ denotes the number of common (native) H-H contacts between s and g, and c_s, c_g are the numbers of H-H contacts found only in s and g, respectively (the maximum stems from the possible degeneracy of ground states). Figure 2 also shows the averages $\langle R_g \rangle$ and $\langle q \rangle$ for sequences 2D100b and 3D103 and illustrates the complementary information in these two quantities. While $\langle R_g \rangle$ indicates the coil-to-globule collapse, $\langle q \rangle$ identifies the folding transition to the native state and thus may serve as a suitable structural order parameter for these kind of systems. In case of sequence 3D103, the ground state ($E = -58$) was excluded from the sampling (due to the difficulty in finding this state) which results in $\langle q \rangle$ saturating at a rather low value (<0.3) for $T \to 0$. This manifests the still large structural differences between conformations with $E = -57$ and the ground state.

Table I compares various methods in finding low energy conformations and, if available, the DOS for common benchmark HP sequences. We also included results from methods which were focused on the low temperature range only, i.e., FRESS [8] and the variants of PERM (pruned-enriched Rosenbluth method) [5] and hence do not provide the entire DOS. Except for the longest sequence (3D136), we could confirm all minimum energy states found previously. The superiority of FRESS for this sequence is the result of various ''efficiency enhancements'' towards low energy states (see [8]) which become obviously the more effective the longer the chain length. However, they do not permit anymore a correct sampling, let alone an estimation of the DOS.

As a second test of performance, we applied our method to the interacting self-avoiding walk (ISAW) representing a homopolymer with nearest-neighbor attraction ($\epsilon = -1$) on the square (sq, 2D) and simple cubic (sc, 3D) lattice. Unraveling the ''phase transition'' behavior of flexible macromolecules in the thermodynamic limit ($N \to \infty$) by means of simple (lattice) models—such as, e.g., the ISAW, the bond-fluctuation model or systems in the continuum—has been a challenge for decades [4,18]. Although the θ point (coil-globule transition) could be investigated well for polymer chains with $N \geq 10000$ monomers, our understanding of the ISAW at very low temperatures remains elusive. Because of the very dense packings resulting for this model, accurate estimates of thermodynamic quantities below T_θ are difficult to obtain. In the most recent computational studies, only chains with $N \leq 125$ in 3D (multicanonical chain growth [19]) and $N \leq 300$ in 2D (adaptive WL sampling with reptation, but without the lowest energy states [16]) could be investigated.

With our generic approach, we were able to obtain accurate DOS estimates for ISAWs up to chain lengths $N = 400$ (2D) and $N = 512$ (3D) over the *entire* energy range (including ground states), and we could then determine reliable thermodynamic quantities even at lowest temperatures ($T \to 0$), see Fig. 3. The possibility to compare the specific heat $C_V(T)$ for various system sizes up to these chain lengths allowed us to draw interesting conclusions which apply for the ISAW on both the sq and sc lattice: At high T, the collapse transition (θ point) indicates a clear phase transition manifested by cooperative structural rearrangements from the coil to the globular state and $C_V(T_\theta) \to \infty$ for $N \to \infty$. At very low T, a pronounced peak appears due to various ground state excitations (here, the ground states form either regular squares or cubes). These excitations are induced by *local* rearrangements at the surface, and therefore, the magnitude of the peak decreases systematically with chain length (2D) or becomes constant to within statistical error bars (3D). The breaking up of the ground state structure bears similarity to surface roughening on crystal facets, i.e., the formation of kinks and edges at the surface of a compact core without vacancies (indeed, bulk vacancies appear at much higher T only). At intermediate temperatures, metastable (and chain length dependent) phases emerge, but they gradually diminish for $N \to \infty$. Most notably, the ISAW on the sq/sc lattice does not undergo a true crystallization transition as observed for other lattice and off-lattice polymer models

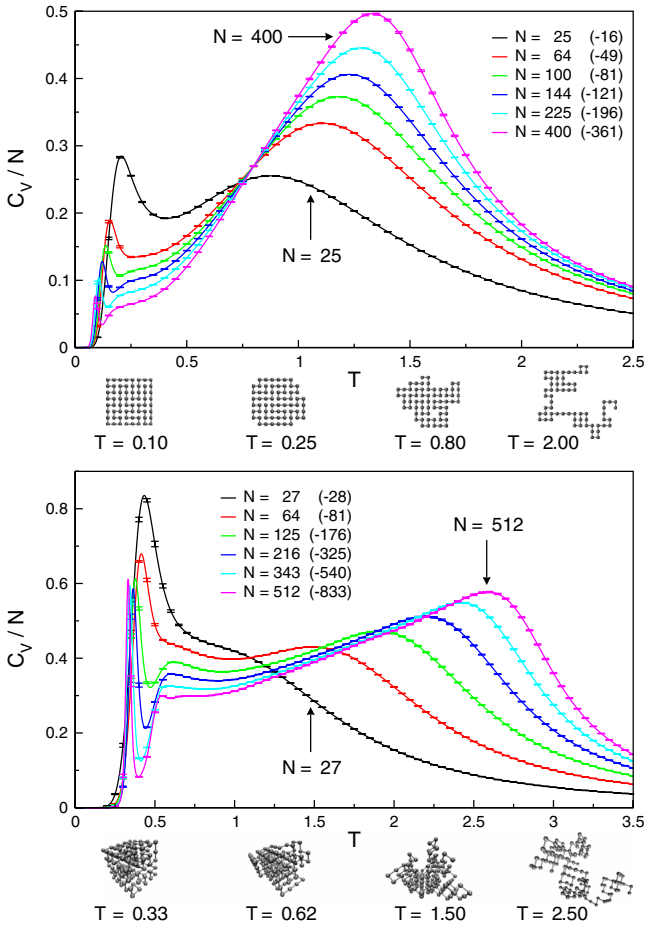

FIG. 3 (color online). Specific heat C_V/N as a function of temperature (T) for ISAWs of various chain lengths N on square (*top*) and simple cubic (*bottom*) lattice. Numbers in parentheses denote corresponding energy minima. *Bottom rows*: Representative structures at specific temperatures for $N = 64$ (2D) and $N = 125$ (3D).

[4,18]. Once in the globular phase, the rigidity of the model does not permit a further cooperative effect (i.e., symmetry breaking) which would be necessary for such a transition. Whereas a variation of chain length ($N \neq$ "magic" number) has an influence on the magnitude and the position of the excitation peak at low T, the overall thermodynamic scenario remains the same for sufficiently large N. Note that it was essential to have data for chains that were longer than other methods could treat in order to ascertain the low T behavior of the ISAW in the thermodynamic limit.

In summary, we have shown that Wang-Landau sampling with suitable Monte Carlo trial moves (pull and bond-rebridging moves combined) offers a powerful solution for studying the thermodynamics of lattice homo- and heteropolymers even in the very demanding low temperature ranges of such models. A major advantage of our method is that it remains rather simple and flexible beside its proven performance which has not been achieved earlier, by more elaborate attempts [10–12,19]. These features make it readily applicable to the study of complex biological phenomena such as, e.g., protein aggregation or protein insertion into a membrane [13]. Since both trial moves are usable for lattice and off-lattice models [20], other systems with conformational constraints should also benefit from our self-adaptive WL procedure.

We thank K. Binder and W. Paul as well as C. Gervais and D. T. Seaton for fruitful discussions. This work was supported in part by NSF Grant No. DMR-0810223.

*twuest@physast.uga.edu

[1] K. A. Dill *et al.*, Protein Sci. **4**, 561 (1995); A. Kolinski and J. Skolnick, Polymer **45**, 511 (2004).
[2] K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[3] A. D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, 1995), p. 47.
[4] K. Binder and W. Paul, Macromolecules **41**, 4537 (2008).
[5] H. Frauenkron *et al.*, Phys. Rev. Lett. **80**, 3149 (1998); H.-P. Hsu *et al.*, Phys. Rev. E **68**, 021113 (2003).
[6] N. Lesh *et al.*, in RECOMB (Berlin, Germany, 2003), p. 188.
[7] R. Backofen and S. Will, Constraints **11**, 5 (2006).
[8] J. Zhang *et al.*, J. Chem. Phys. **126**, 225101 (2007).
[9] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, UK, 2005), 2nd ed.
[10] Y. Iba *et al.*, J. Phys. Soc. Jpn. **67**, 3327 (1998); G. Chikenji *et al.*, Phys. Rev. Lett. **83**, 1886 (1999).
[11] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003); J. Chem. Phys. **120**, 6779 (2004); T. Prellberg and J. Krawczyk, Phys. Rev. Lett. **92**, 120602 (2004).
[12] S. C. Kou *et al.*, J. Chem. Phys. **124**, 244903 (2006).
[13] See, e.g., P. M. Harrison *et al.*, J. Mol. Biol. **286**, 593 (1999); R. Bonaccini and F. Seno, Phys. Rev. E **60**, 7290 (1999); L. Zhang *et al.*, Biophys. Chem. **133**, 71 (2008).
[14] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001); Phys. Rev. E **64**, 056101 (2001).
[15] J. M. Deutsch, J. Chem. Phys. **106**, 8849 (1997).
[16] A. G. Cunha-Netto *et al.*, Phys. Rev. E **78**, 055701(R) (2008).
[17] R. Fraser and J. I. Glasgow, in ICANNGA (Warsaw, Poland, 2007), Vol. I, p. 758.
[18] D. F. Parsons and D. R. M. Williams, J. Chem. Phys. **124**, 221103 (2006); W. Paul *et al.*, Phys. Rev. E **75**, 060801(R) (2007); D. Seaton *et al.*, Comput. Phys. Commun. **180**, 587 (2009).
[19] T. Vogel *et al.*, Phys. Rev. E **76**, 061803 (2007).
[20] P. V. K. Pant and D. N. Theodorou, Macromolecules **28**, 7224 (1995).