Monte Carlo simulation of proteins through a random walk in energy space

Nitin Rathore and Juan J. de Pablo[a)]
Department of Chemical Engineering, University of Wisconsin–Madison, Madison, Wisconsin 53706

(Received 5 November 2001; accepted 29 January 2002)

A Monte Carlo algorithm that performs a random walk in energy space has been used to study random coil–helix and random coil–beta sheet transitions in model proteins. This method permits estimation of the density of states of a protein via a random walk on the energy surface, thereby allowing the system to escape from local free-energy minima with relative ease. A cubic lattice model and a knowledge based force field are employed for these simulations. It is shown that, for a given amino acid sequence, the method is able to fold long polypeptides reproducibly. Its results compare favorably with those of annealing and parallel tempering simulations, which have been used before in the same context. This method is used to examine the effect of amino acid sequence and chain length on the folding of several designer polypeptides. © *2002 American Institute of Physics.* [DOI: 10.1063/1.1463059]

I. INTRODUCTION

The protein folding problem has challenged researchers for several decades. Today, as the genomic puzzle is being elucidated, scientists are gradually turning their attention to proteomics, in the hope of predicting the structure and function of proteins from knowledge of their sequence. Predicting a protein's three-dimensional structure, exploring its energy landscape, understanding the kinetics of the folding process and computing its thermodynamic properties are all problems of current interest.

In recent years, considerable progress has been achieved in our understanding of protein folding. The energy landscape theory has provided a theoretical framework in which to work; important insights into the stability and dynamics of complex systems have emerged from that formalism (as recently reviewed in Ref. 1). The energy landscape approach has led to a renewed interest in determining the precise nature of a protein's energy surface, as well as, that of transitions between different basins of that surface. For a protein, that landscape is now believed to consist of a multitude of local minima separated by high energy barriers.

Given the complexity of proteins, molecular simulations have been used extensively to examine their structure and function.[1] Unfortunately, conventional canonical molecular dynamics and Monte Carlo simulations are ill-suited to sample their energy landscape; with few exceptions, simulations have only been able to provide partial answers to the questions that arise in the study of proteins.

Several advanced techniques have been proposed in attempts to improve the simulation of protein and polymer structure. Some of the more recent methods have relied on the idea of generalized ensembles such as multicanonical ensembles and simulated tempering.[2–8] Multicanonical methods are particularly attractive in that they rely on the idea of artificially eliminating the energy barriers, thereby circumventing some of the problems associated with traditional sampling techniques (e.g., trapping in local free energy minima). Furthermore, they also permit efficient calculation of thermodynamic quantities with high accuracy. These techniques, however, have the disadvantage of being computationally demanding and tedious. They require an iterative calculation of weight factors which are not known *a priori*. And, while algorithms to compute multicanonical weight factors have continually been modified,[9–10] significant efforts must still be devoted to their calculation.

In this work, a Monte Carlo method proposed by Wang and Landau[11] is implemented for the study of protein folding. Its results are compared to those of simulated annealing and parallel tempering calculations. Like other multicanonical algorithms, this method seeks to overcome the problems associated with local free energy barriers. However, in this method temperature plays no role in the sampling and a direct estimate of the density of states is generated in a self consistent manner.

Without loss of generality, a lattice model is used in this work to represent proteins. The model is sufficiently simple to permit efficient calculations (needed for precise characterization of the thermodynamics of the folding process) and it is sufficiently detailed to capture many of the physical interactions that give rise to complex three-dimensional structures comprising α-helices and β-sheets.

II. METHODS

A. Model

The lattice model employed here is based on the side-chain-only (SICHO) model by Kolinski *et al.*[12–15] Each amino acid is represented by the centroid of its side chain; a protein is modeled as a chain connecting these virtual particles on a cubic lattice, with the lattice spacing corresponding to 1.45 Å in real proteins. The chain vectors representing virtual bonds between interaction centers are of variable length, ranging from $9^{1/2}$ to $30^{1/2}$ lattice units. The model is reminiscent of the bond-fluctuation model,[16] but has a higher coordination number; there are in all 646 possible bond vec-

a)Electronic mail: depablo@engr.wisc.edu

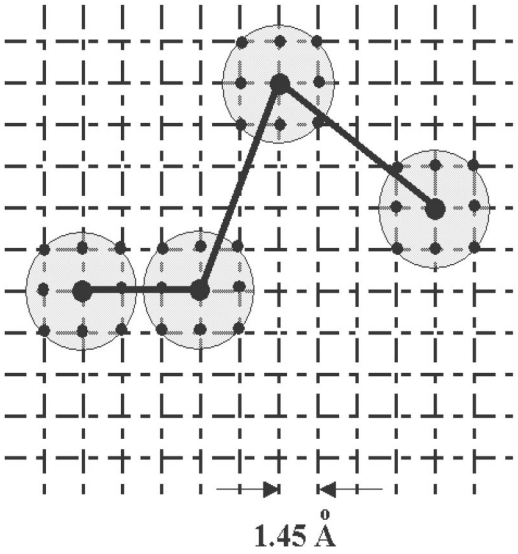

FIG. 1. Schematic illustration of the protein chain representation. The solid small circles correspond to the lattice points occupied by an amino acid; the transparent circles represents amino acids, the solid lines represent the virtual bonds connecting these amino acids, and the grid represents the lattice.

tors. A hard-core excluded volume cluster, consisting of 19 lattice points (see Fig. 1), is associated with each interaction site or amino acid. These 19 points are the center, six nearest neighbors at a unit distance from the center, and 12 second nearest neighbors at a distance $2^{1/2}$ from the center. To describe bulkier residues, this hard core interaction is supplemented by a soft repulsive sphere having a radius whose magnitude depends on the two amino acids involved in pairwise interactions.

The knowledge-based force field proposed by Kolinski *et al.*[12,14,15] includes a chain stiffness potential, a secondary structure bias, short-range interactions, hydrogen-bond interactions, and long-range interactions. The parameters describing these interactions were originally derived by analyzing several geometric characteristics of known protein structures, which were then translated into a lattice discretized form. Different values are reported for the prefactors of these different energy components in the two citations (e.g., coefficient of E_{hbond} is 1.25 in Ref. 12, whereas it is 0.875 in Ref. 14, similarly the prefactors for $E_{\text{short range}}$ differ in the two references). In this work, we therefore chose to adjust these prefactors on the basis of our own structure prediction results for small globular proteins.

The effective potential energy function used in our simulations has the form

$$E_{\text{total}} = E_{\text{stiff}} + E_{\text{struct}} + E_{\text{map}} + 1.25E_{\text{hbond}} + 0.5E_{\text{shortrange}} + 1.25E_{\text{longrange}}, \quad (1)$$

where E_{stiff} is the chain stiffness energy, E_{struct} and E_{map} represent the secondary structure bias, E_{hbond} denotes the energy of the hydrogen bond network, and $E_{\text{shortrange}}$ and $E_{\text{longrange}}$ are the short-range and long-range contributions to the energy, respectively. Details of the calculation of the various energy components and the necessary parameters can be found in Refs. 14 and 15.

Three artificial polypeptide sequences of varying lengths are considered in this work. The first two sequences, SeqA and SeqB, were used by Ilkowski *et al.*[12] in the original development of the lattice model employed here. The third sequence, SeqC or polyalanine, was used in the literature for the study of the helix–coil transition using a multicanonical algorithm.[7] These three sequences can be written as

(1) SeqA: $(-\text{Ala}-\text{Leu}-\text{Ser}-\text{Ser}-\text{Ala}-\text{Ala}-\text{Ser}-)_n$,
(2) SeqB: $(-\text{Val}-\text{Ser}-)_n$,
(3) SeqC: $(-\text{Ala}-)_n$.

As discussed in Ref. 13, SeqA has the characteristics of a helical structure; a repeat period of 7-residues, with well defined hydrophobic (alanine and leucine) and polar (serine) residues. Alanine and serine are often found in the helical fragments of proteins. SeqC, too, is therefore expected to form a helix. On the other hand, valine and threonine residues often appear in beta sheets; SeqB is therefore used to study β-sheet structures. The values of chain length n, considered in this work range from 10 to 56.

B. Simulation method

Three types of Monte Carlo methods are considered in this work: simulated annealing, parallel tempering, and the Monte Carlo method proposed by Wang and Landau.[11] In simulated annealing, conventional, canonical-ensemble simulations are started at a high temperature, from a totally random initial configuration. The system is then gradually cooled in steps, allowing it to equilibrate at each temperature. The final structure obtained at the lowest temperature is assumed to be the stable, folded conformation of the protein.

In parallel tempering, N noninteracting copies or replicas of the protein molecule are simulated in N boxes, each at a different temperature. In addition to the standard Monte Carlo moves in each box, the conformations in different replicas are swapped at regular intervals. Trial swaps are accepted with probability

$$P_{\text{acc}} = \min[1, \exp(\Delta\beta_{ij}\Delta E_{ij})], \quad (2)$$

where ΔE_{ij} is the difference in energy between conformations in boxes i and j, and where $\Delta\beta_{ij}$ is the difference between their inverse temperatures. The exchange between the structures in different replicas facilitates relaxation of structures that might otherwise be trapped in local energy minima. This feature is absent from standard annealing simulations, and parallel tempering is therefore believed to be a more effective technique for locating the global free energy minimum of the protein.

In the annealing and parallel-tempering methods described above, a traditional Monte Carlo simulation is conducted in each replica or simulation box. At any given temperature, configurations are therefore generated with a probability $P(E)$ proportional to the Boltzmann weight, i.e.,

$$P(E) \propto g(E) \exp(-E/k_B T), \quad (3)$$

where $g(E)$ denotes the density of states corresponding to energy level E. Recently, Wang and Landau[11] have proposed a novel simulation technique which, when applied to the

Ising model, permits accurate estimation of the density of states of that model. This method has elements of multicanonical sampling in that the weight factor is the reciprocal of the density of states. Multicanonical techniques have been used before to study protein folding.[7,10] Our work differs from previous studies in that the Wang and Landau's technique is employed to construct a random walk in energy space. In order to avoid confusion with earlier multicanonical implementations, we refer to the method employed here as the random walk algorithm (RWA). The goal of this method is to generate a random walk in energy space with probability proportional to the reciprocal density of states, i.e.,

$$p(E) \propto \frac{1}{g(E)}. \tag{4}$$

If $g(E)$ was known with sufficient accuracy, a RWA simulation would lead to flat energy histograms. The density of states however, is not known *a priori*. In RWA, it is generated "on the fly" as the simulation proceeds. At the beginning of the simulation, $g(E)$ is assumed to be unity for all energy levels E. Trial Monte Carlo moves are accepted with probability,

$$P_{\mathrm{acc}}(E_1 \rightarrow E_2) = \min\left[1, \frac{g(E_1)}{g(E_2)}\right], \tag{5}$$

where $E_1$and E_2 are the energy of the system before and after a trial move. After each trial move, the corresponding density of states is updated by multiplying the current, existing value by a convergence factor f which is greater than unity $(f>1)$, i.e.,

$$g(E) \rightarrow g(E) \times f. \tag{6}$$

If the move is rejected, the instantaneous energy of the system remains unchanged at E_1, and therefore $g(E_1)$ is modified by the convergence factor. If the move is accepted, $g(E_2)$ is modified. We have used an initial convergence factor of $e^1 \simeq 2.718\,28$ in all our simulations.

Every time that $g(E)$ is modified, a histogram of energies $H(E)$ is also updated. The $g(E)$ refinement process outlined above is continued until $H(E)$ becomes sufficiently flat. In this work, we consider a histogram to be flat if $H(E)$ for all possible E is not less than 80% of the average energy histogram, $\langle H(E) \rangle$. Once this condition is satisfied, the convergence factor is reduced by an arbitrary amount. Here we follow Wang and Landau's recommendation, and we set $f_{\mathrm{new}} = \sqrt{f_{\mathrm{old}}}$. The energy histograms are then reset to zero $(H(E)=0)$, and a new simulation cycle is started, continuing until the new histogram $H(E)$ is flat again. The process is repeated until f is smaller than some specified value, e.g., $f_{\mathrm{final}} = \exp(10^{-8})$ in our case.

Throughout the course of the simulation, the condition of detailed balance is not satisfied. Only towards the end of a calculation, when $f \rightarrow 1$, is detailed balance approached. Thermodynamic averages are therefore calculated using only information generated during the last iteration step of the process described above.

The main product of a simulation is the density of states over a specified energy range, which is determined to within a multiplicative constant. In Wang and Landau's original formulation, the energy range of interest is decomposed into a set of smaller energy windows, in which independent RWA simulations are conducted; $g(E)$ data corresponding to different overlapping energy windows can subsequently be combined to yield the density of states over a wider energy range. Thermodynamic quantities such as the free energy $F(T)$, internal energy $U(T)$, entropy $S(T)$ and specific heat capacity $C(T)$ are then determined according to

$$F(T) = -k_B T \ln\left(\sum g(E) e^{-\beta E}\right), \tag{7}$$

$$U(T) = \langle E \rangle_T = \frac{\sum E g(E) e^{-\beta E}}{\sum g(E) e^{-\beta E}}, \tag{8}$$

$$S(T) = \frac{U(T) - F(T)}{T}, \tag{9}$$

$$C(T) = \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{k_B T^2}, \tag{10}$$

For the specific case of the Ising model, Wang and Landau have suggested dividing the energy range of interest into small overlapping windows to facilitate convergence. For proteins, however, it is advisable to use sufficiently large energy windows, so that drastic moves needed to restructure the protein conformation can be implemented, and their likelihood of being rejecting is not too high. In fact, our experience indicates that window sizes below some critical size may lead to erroneous results in the form of incorrect transition temperatures. If a window is too narrow, the system can be trapped in a low probability configuration, and be deprived of a mechanism to escape that minima, without violating the bounds imposed by the window size. This also emphasizes the need for drastic Monte Carlo moves that enhance sampling and increase the efficiency of walks in the energy space.

All of the simulations described above (annealing, parallel-tempering, and RWA) have been conducted using two kinds of trial moves. In the first of these, a randomly chosen interaction site is displaced by a random amount of up to $5^{1/2}$ lattice units, taking into consideration the fact that bond length and excluded volume constraints must not be violated. These moves are accepted with probability given by[17]

$$P_{\mathrm{acc}} = \min[1, \exp(-\beta \Delta E)]. \tag{11}$$

As discussed above, local moves are insufficient to generate drastic configurational rearrangements of a protein's structure. A second type of nonlocal trial move, namely a pivot move, is therefore introduced.[18] In a pivot move, an interaction site is randomly assigned to act as a pivot point for either the amino or the carboxyl end segment of the protein (with equal probability); these segments are then rotated clockwise or anticlockwise by 90° about this pivot site and about a randomly chosen axis. The direction cosine method

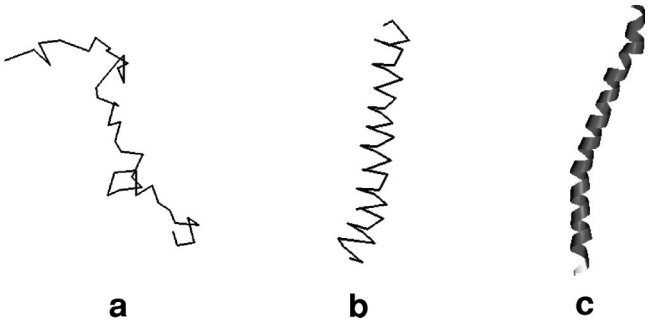

FIG. 2. Folding of SeqA from (a) initial unfolded random coil on a lattice to (b) the final folded helical conformation on lattice. (c) Shows the complete regenerated protein using MMTSB (Ref. 19). (Drawn with Rasmol and WebLab Viewer Lite.)

is used to generate the postmove coordinates, and the move is accepted with probability given by Eq. (11).

III. RESULTS AND DISCUSSION

All of the methods employed in this work were successful at folding short SeqA, SeqB, and SeqC sequences into the correct structure. Figures 2 and 3 show the initial and folded structures obtained from our simulations for one such sequence of both the secondary structure class: 42-residue SeqA (α-helix) and 42 residue SeqB (β-sheet), respectively.

In Fig. 4, the average conformational energy obtained from annealing and tempering simulations is compared to that obtained from RWA simulations. The annealing and tempering simulations provide thermodynamic information only at discrete temperatures. Histogram reweighting techniques can subsequently be used to extract more information from individual simulation runs. In contrast, the RWA method provides a direct and precise estimate of the energy distribution and other thermodynamic quantities over the entire temperature range of interest from a single calculation. The results of all methods are consistent with each other. Note, however, that RWA simulations require a fraction of the computational effort of annealing or parallel-tempering simulations.

Figure 5 shows specific-heat curves for SeqA and SeqB peptides of two different lengths. The two sequences exhibit a different folding behavior. For both chain lengths, the β-sheets (SeqB) exhibit a lower transition temperature and specific heat than the α-helices (SeqA). This is different from the findings of Skolnick *et al.*[12] who reported that β-sheets have a higher transition temperature and much larger specific

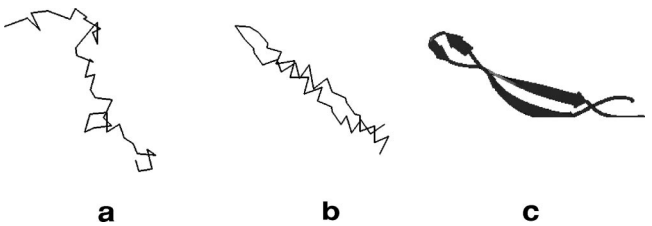

FIG. 3. Folding of SeqB from (a) initial unfolded random coil on a lattice to (b) the final folded beta sheet conformation on lattice. (c) Shows the complete regenerated protein using MMTSB (Ref. 19). (Drawn with Rasmol and WebLab Viewer Lite.)

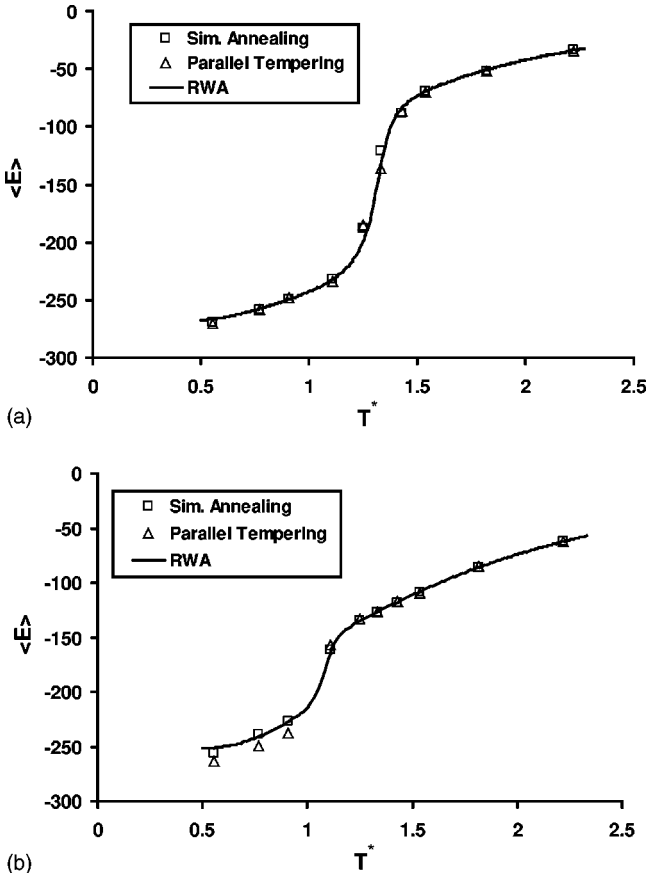

FIG. 4. Average potential energy $\langle E \rangle$ of (a) SeqA and (b) SeqB obtained through different algorithms as a function of dimensionless temperature T^*. Chain length is 42 and $\langle E \rangle$ is in $k_B T$ units.

heats than α-helices. Note, however, that in our force field we have chosen not to incorporate contributions such as the multibody potential[14] and the centrosymmetric potential;[12] also the prefactors are different; this could be the cause of the discrepancies between our results and theirs.

The absence of an explicit solvent treatment may also result in incorrect folding behavior. Polypeptides having the same amino acid sequence have been shown to form both α-helix and β-sheet structures, depending on the electrostatic environment.[6] We too find that, in some of our RWA simulations of SeqB, though the β-sheet structure dominates the folded conformation regimes, some of the folded configurations (energy minima) do exhibit a helical structure. This finding suggests that minor perturbations of the force field (or environment) could balance the distribution of structures in favor of helices. A more accurate characterization of these transitions would merit a higher resolution model. Here we merely point out that the RWA method is able to reveal a level of conformational detail that is not afforded by other simulation techniques.

Figure 6 shows the effect of chain length on the three designer sequences. The peak in the specific heat as a function of temperature is identified with the folding transition. All of our results for transition temperatures are summarized in Table I. The trend for SeqA and SeqB at different lengths are in agreement with the literature.[12] As the peptide chain length is increased, the transition temperature also increases.

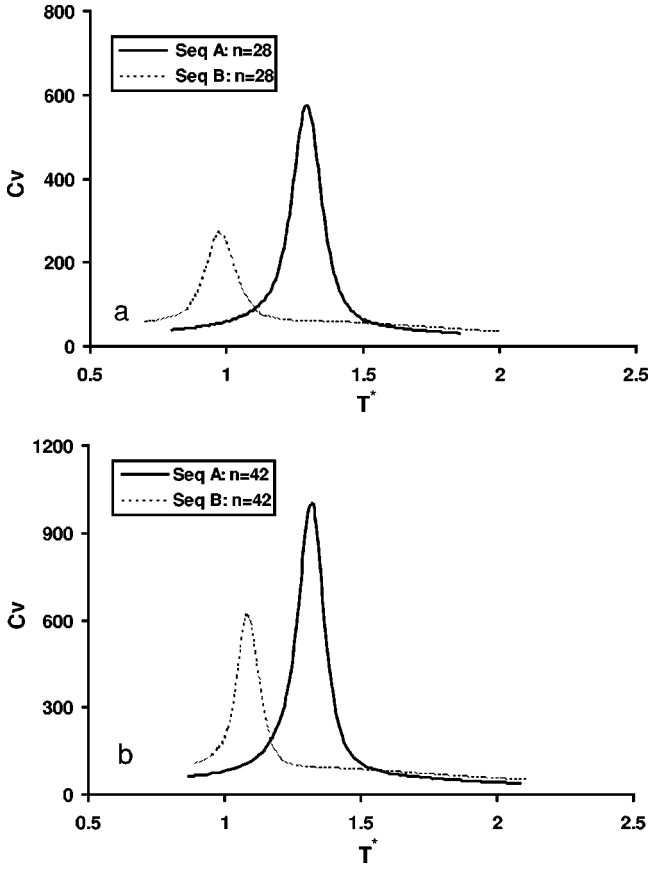

FIG. 5. Effect of amino acid sequence on the specific heat plots for SeqA and SeqB for two different chain lengths. Specific heat C_v as a function of dimensionless temperature T^* for (a) $n=28$ and (b) $n=42$.

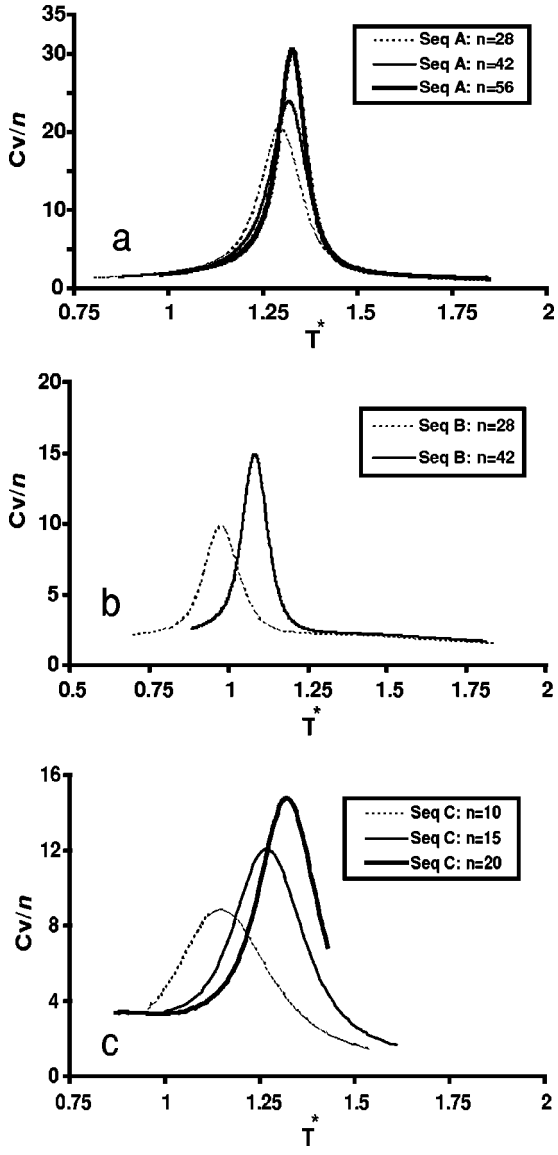

FIG. 6. Effect of chain length on the specific heat plots for (a) SeqA and (b) SeqB and (c) SeqC. Specific heat per residue C_v/n is shown as a function of dimensionless temperature T^*. The peaks get more pronounced and shift to right with increasing chain length.

This increase appears to be more pronounced for β-type polypeptides [Fig. 6(b)].

The RWA method provides a sufficiently high resolution in the helix-forming SeqA peptide [Fig. 6(a)] to identify a small but noticeable shift in the peaks with increasing chain length. This was reported to be beyond the resolution of regular canonical simulations.[12] For the case of SeqC, polyalanine, Hansmann and Okamoto[7] used an all-atom representation and a multicanonical Monte Carlo technique to investigate the folding transition. The results of our RWA simulations using a simpler, lattice model are in good qualitative agreement with their reported trends. As is evident from Fig. 6(c), the peak height and the transition temperature increase with the number of residues. This is in accordance to what is expected for a thermodynamic phase transition. Based on the results SeqA and SeqC simulations, we can also conclude that the magnitude of the shifts in the peaks and in the peak heights with chain length depends on the amino acid sequence, even for the same secondary structure. The peaks are expected to be more pronounced for a peptide comprising good helix formers (e.g., SeqA), than for a sequence having a lesser propensity towards a helical structure.

The sequences discussed above have been used in the literature to establish the validity of force fields and the efficiency of simulation techniques. We now extend our analysis of protein folding to several real helical proteins for which the folded structure is available. Figure 7 shows results for one such protein, namely C-peptide. This peptide, also studied by Okamoto,[6] comprises the residues 1–13 of ribonuclease A. The density of states $g(E)$ over a wide energy range as obtained from a single production run of a

TABLE I. Transition temperatures for various proteins as determined by the random walk algorithm (RWA).

Sequence	Chain length	Transition temperature
SeqA	28	1.294
SeqA	42	1.319
SeqA	56	1.327
SeqB	28	0.973
SeqB	42	1.082
SeqC	10	1.150
SeqC	15	1.268
SeqC	20	1.324
C-Peptide	13	0.850
1A11	25	0.824

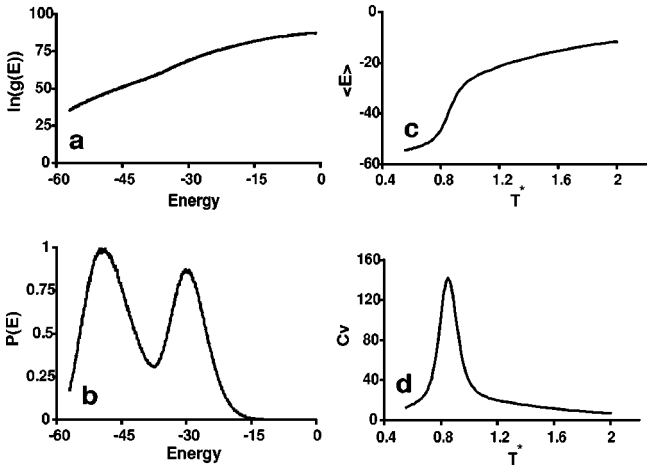

FIG. 7. Folding behavior of C-peptide: (a) The normalized density of states vs energy calculated using the RWA method; (b) the normalized energy distribution at the transition temperature; (c) the average energy $\langle E \rangle$ as a function of temperature; and (d) the specific heat C_v/n as a function of temperature.

RWA Monte Carlo simulation is shown in Fig. 7(a). The secondary structure obtained at low temperatures is consistent with the experimental structure. The density of states can be used to compute the energy distribution at any temperature of interest. Figure 7(b) presents such a normalized distribution precisely at the transition temperature. Figures 7(c) and 7(d) show the average potential energy and specific heat as a function of temperature.

A similar analysis for other helical proteins (e.g., 1A11, Protein Data Bank accession number) yields results of comparable accuracy. The results are summarized in Table 1.

IV. CONCLUSION

A Monte Carlo technique based on sampling according to the density of states has been implemented to study the folding transition of several proteins. For model polypeptide sequences, the results of this method are consistent with those of other simulation techniques. In the case of realistic protein sequences, the method is able to generate folded structures in agreement with experimental data. In addition to providing the folded structure of a protein, this method offers the added advantage of providing high-accuracy thermodynamic information over a wide range of interest, thereby offering the possibility of studying folding transitions with an unprecedented level of detail.

For the particular case of a lattice model and a knowledge based force field, the RWA method has been used to examine the effect of amino acid sequence and peptide chain length on the folding behavior of helix–coil and β-sheet–coil transitions. Within the limitations of the lattice model, we are able to consider a broad and diverse range of folding behavior. A more detailed study of these transitions using more refined force fields is now under way.

ACKNOWLEDGMENT

This work was supported by NSF(CTS-9901430).

[1] J.-E. Shea and C. L. Brooks III, Ann. Rev. Phys. Chem. **52**, 499 (2001).
[2] F. A. Escobedo and J. J. de Pablo, J. Chem. Phys. **105**, 4391 (1996).
[3] Q. L. Yan and J. J. de Pablo, J. Chem. Phys. **113**, 1276 (2000).
[4] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **54**, 5863 (1996).
[5] U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).
[6] Y. Okamoto, Int. J. Mod. Phys. C **10**, 1571 (1999).
[7] Y. Okamoto and U. H. E. Hansmann, J. Phys. Chem. **99**, 11276 (1995).
[8] D. Gront, A. Kolinski, and J. Skolnick, J. Chem. Phys. **113**, 5065 (2000).
[9] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).
[10] F. Yasar, T. Celik, B. A. Berg, and H. Meirovitch, J. Comput. Chem. **21**, 1251 (2000).
[11] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[12] B. Ilkowski, J. Skolnick, and A. Kolinski, Macromol. Theory Simul. **9**, 523 (2000).
[13] A. Kolinski, L. Jaroszewki, P. Rotkiewicz, and J. Skolnick, J. Phys. Chem. B **102**, 4628 (1998).
[14] A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, Proteins **37**, 592 (1999).
[15] A. Kolinski and J. Skolnick, Proteins **32**, 475 (1998).
[16] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).
[17] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[18] J.-M. Shin and W. S. Oh, J. Phys. Chem. B **102**, 6405 (1998).
[19] MMTSB (Multiscale Modeling Tools in Structural Biology) is a web based NIH research resource for the development and integration of modeling tools to explore multiresolution models in structural biology. The package is available at http://mmtsb.scripps.edu