Density of states simulations of proteins

Nitin Rathore, Thomas A. Knotts IV, and Juan J. de Pablo[a)]
Department of Chemical Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706

(Received 22 October 2002; accepted 6 December 2002)

A modified version of a recently introduced algorithm that calculates density of states by performing a random walk in energy space has been proposed and implemented to study protein folding in a continuum. A united atom representation and the CHARMM19 [B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983)] force field are employed for these simulations. This method permits estimation of the density of states of a protein via a random walk in the energy space, thereby allowing the system to escape from local free-energy minima with relative ease. Unlike the earlier formulation that showed slow convergence for continuum simulations, this methodology is designed to achieve better sampling and faster convergence. The modified method is used to examine folding transitions of two peptides: deca-alanine and Met-enkephalin. Protein folding both with and without an implicit solvent (solvent accessible surface area model) has been studied to validate the usefulness of the proposed algorithm.
© *2003 American Institute of Physics.* [DOI: 10.1063/1.1542598]

I. INTRODUCTION

Canonical simulations of folding processes in model proteins are challenging due to the rough potential energy landscape of these systems. Several advanced techniques have been proposed in attempts to improve simulations by smoothing out such landscapes.[1–7] These include parallel tempering, umbrella sampling and generalized ensemble techniques.[8] Multicanonical methods are based on non-Boltzmann probability distributions and are particularly attractive in that they rely on the idea of artificially eliminating energy barriers, thereby circumventing some of the problems associated with traditional sampling techniques (e.g., trapping in local free energy minima). They also permit efficient calculation of thermodynamic quantities with high accuracy. These techniques, however, have the disadvantage of being computationally demanding and tedious. Their use requires an iterative calculation of weight factors which are not known *a priori*.

The central quantity of interest in all these simulations is the density of states, $\Omega(U)$. Once the density of states is known, energy states can be visited with uniform probability regardless of their location on the energy landscape. Recently, a different class of algorithms[9,10] has emerged with the potential of providing a direct estimate of the density of states in a self-consistent manner. Like more established multicanonical algorithms, this method seeks to overcome the problems associated with local free energy barriers. In recent work,[11] we have shown how the method of Wang and Landau[9] could be used for the study of helix–coil and beta sheet–coil transitions of designer peptides on a lattice. In this work we examine, whether similar ideas can be used to study proteins in a continuum.

More specifically, a modified random walk algorithm is implemented in this work for study of protein folding in a continuum using a united atom representation and the CHARMM19 potential function.[12] Protein simulations in a continuum (using an atomistic representation) require a random walk in a much larger energy space with a higher degeneracy than that on a lattice. Efficient sampling and convergence are both difficult to achieve with the Wang–Landau algorithm in its original formulation. A method is proposed here to circumvent these problems by merging the random walk scheme with a parallel tempering methodology. The resulting technique achieves fast convergence and is shown to be capable of efficient sampling by its ability to continually alternate between the folded and unfolded energy regimes of model peptides.

We begin with a brief description of the atomistic model, the force field employed in this work and the two peptides, deca-alanine, and Met-enkephalin, that are used to benchmark the proposed algorithm. We then discuss the simulation scheme in detail. This method is used to study folding transitions of the two peptides in different environments and the results are presented in the form of thermodynamic quantities as functions of temperature.

II. METHODS

A. Model

The CHARMM19 force field is used with a united atom representation where the nonpolar hydrogen atoms are combined with the heavy atoms to which they are bound. For both *in vacuo* and implicit solvent simulations, we use the EEF1 model parameters,[13] where the partial charges on the amino acids are modified to neutralize the side chains and the patched termini. The interactions between atoms are described by the following potential energy function:

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

$$
\begin{aligned}
V_{\text{total}}=&\sum_{\text{bonds}} K_r(r-r_{\text{eq}})^2+\sum_{\text{angles}} K_\theta(\theta-\theta_{\text{eq}})^2\\
&+\sum_{\text{dihedrals}} K_\phi[1+\cos(n\phi-\delta)]\\
&+\sum_{\text{impropers}} K_\omega(\omega-\omega_{\text{eq}})^2\\
&+\sum_{\text{LJ}} \epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\\
&+\sum_{\text{Coulombic}} \frac{1}{4\pi\epsilon(r)\epsilon_0}\frac{q_iq_j}{r_{ij}}.
\end{aligned}
$$

For the nonbonded energy evaluation a 1–3 exclusion principle is used. In addition to this, 1–4 Coulombic interactions are scaled down by a factor of 0.4. This is consistent with the original parametrization of CHARMM19. A cutoff of 12 Å is used for both the electrostatic and van der Waals terms. Since the long range cutoff is sufficiently large, a simple cut shift potential scheme is used for nonbonded interactions.

An implicit solvent model based on the solvent accessible surface area (SASA) is employed with solvation parameters as proposed by Ferrara *et al.*[14] Electrostatic screening effects are approximated by a distance dependent dielectric (DDE) function and a set of partial charges with neutralized side chains. The model assumes that the mean solvation energy is proportional to the SASA of the solute. For a solute having M atoms with Cartesian coordinates $\mathbf{r}$, the solvation term is given by

$$V_{\text{solvation}}=\sum_{i=1}^{M}\sigma_iA_i(\mathbf{r}), \tag{1}$$

where σ_i and $A_i(\mathbf{r})$ are the atomic solvation parameters and SASA of atom i, respectively. The computations of atomic solvation parameter and the SASA is performed as indicated in Ref. 14. The SASA model, however, only accounts for the free energy cost of burying a charged residue in the interior of a protein. The solvent screening effect is approximated by using a distance-dependent dielectric (DDE) function, $\epsilon(r)=2r$, as suggested in Ref. 14. This is a simplified way of accounting for the solvent polarization effects and is consistent with the formulation of SASA model parameters.

Two polypeptide sequences are considered in this work. The first sequence, deca-alanine, is assumed to form an α-helix because alanine has a high propensity to appear in helical fragments of proteins. It is a common peptide used in the literature for the study of the helix–coil transition using simulations.[11,15] The second peptide, Met-enkephalin, is a brain neuropeptide with five amino acids whose sequence is Tyr-Gly-Gly-Phe-Met. It has been studied in detail through experiments[16] and has also served as benchmark for testing simulations,[17–20] where the ground state conformation has been reported to be a β-turn. Both these peptides have electrically neutral side chains. The patched residues at the N and C termini were also neutralized using the EEF1 model parameters.[13]

B. Simulation method

The random-walk algorithm, as originally proposed by Wang and Landau, was previously used to study protein folding transitions on a lattice, where the conformational space is limited and the energy range of interest is small.[11] In a continuum, however, traversing in the conformational space is slow and convergence deteriorates rapidly with increasing system size. A modified version of the algorithm is presented here for simulations of proteins in a continuum.

We begin with a brief description of the earlier formalism. The goal of the method is to perform a random walk in energy space with probability proportional to the reciprocal density of states, i.e.,

$$p(U)\propto\frac{1}{\Omega(U)}. \tag{2}$$

If $\Omega(U)$ was known with sufficient accuracy, a random walk would lead to flat energy histograms. The density of states however, is not known *a priori*. In the Wang–Landau method, it is generated ‘‘on the fly’’ as the simulation proceeds. At the beginning of the simulation, $\Omega(U)$ is assumed to be unity for all energy levels U. Trial Monte Carlo moves are accepted with probability

$$P_{\text{acc}}(U_1\rightarrow U_2)=\min\left[1,\frac{\Omega(U_1)}{\Omega(U_2)}\right], \tag{3}$$

where U_1 and U_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 that is greater than unity $(f>1)$, i.e., $\Omega(U)\rightarrow\Omega(U)f$. Every time that $\Omega(U)$ is modified, a histogram of energies $H(U)$ is also updated. This $\Omega(U)$ refinement process is continued until $H(U)$ becomes sufficiently flat. Once this condition is satisfied, the convergence factor is reduced by an arbitrary amount. Here we follow Wang and Landau's recommendation and set $f_{\text{new}}=\sqrt{f_{\text{old}}}$. The energy histogram is then reset to zero $(H(U)=0)$, and a new simulation cycle is started, continuing until the new histogram $H(U)$ is flat again. The process is repeated until f is smaller than some specified value, e.g., $f_{\text{final}}=\exp(10^{-7})$. Throughout the course of the simulation, the condition of detailed balance is not satisfied. Only towards the end of a calculation, when $f\rightarrow1$, is detailed balance approached.

The Monte Carlo algorithm proposed here comprises of two types of trial moves. The first move consists of displacements performed using a molecular dynamics scheme. Each MC step includes m molecular dynamics steps, with a time step longer than that required for standard MD simulations. At the beginning of each MC step, the velocities are assigned at random, based on a Gaussian distribution corresponding to the temperature of the simulation box. Then m NVE molecular dynamics moves are performed with the standard velocity verlet integrator. We point out that temperature in our simulations is only used to assign kinetic energy to the system at the beginning of each MC step. A random walk is still performed in potential energy space, but since the kinetic energy is now used to propose trial moves, the acceptance criteria given by Eq. (3) are modified according to

$$P_{\mathrm{acc}}(1\rightarrow 2)=\min\left[1,\exp(-\beta\Delta K)\frac{\Omega(U_1)}{\Omega(U_2)}\right], \tag{4}$$

where ΔK is the change in the total kinetic energy when the system moves from state 1 to state 2.

These moves performed through molecular dynamics may be 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. In this move, an interaction site p (one of the heavy atoms) is randomly chosen as a pivot point for either amino or carboxyl end segments of the protein (with equal probability). A rotation angle θ is then selected such that $0<\theta<\theta_{\max}$, where $\theta_{\max}$ is some maximum rotation angle size. One of the segments is then rotated clockwise or counterclockwise by an angle θ about the pivot site p and about a randomly chosen coordinate axis. The direction cosine method[21] is used to generate the postmove coordinates, and the move is accepted with probability given by Eq. (3).

The main product of the simulation is the density of states over a specified energy range, that is determined to within a multiplicative constant. In earlier work,[9] it was suggested that the energy range of interest can be decomposed into a set of smaller energy windows to facilitate convergence; $\Omega(U)$ data corresponding to different overlapping energy windows can subsequently be combined to yield the density of states over a wider energy range. In our previous work on lattice proteins, however, we opted for the use of sufficiently large energy windows, so that drastic moves needed to restructure the protein conformation could be implemented. If a window is too narrow, the system can be trapped in a low probability configuration, and be deprived of a mechanism to escape local minima without violating the bounds imposed by the window size. For the density of states simulations in a continuum, the energy range of interest and the number of possible energy states are much larger than on a lattice. The convergence of these simulations deteriorates rapidly as the energy window size is increased, particularly in the energy range corresponding to the transition regime. In that sense it is advantageous to go to a smaller window size. Our method gets around this apparent paradox by merging the density of states implementation with a parallel tempering formalism; that is, N noninteracting replicas of the protein molecule are simulated in N boxes, each at a different temperature. Each simulation box represents an energy window, and the energy ranges in these boxes are assigned so that the windows in the adjacent boxes overlap. As before, the purpose of the box temperature is just to assign random velocities to all the atomic sites at the beginning of each MC move. To achieve a higher acceptance rate the box temperature is set approximately to the middle of the potential energy range corresponding to each replica. Also, as we move toward the higher energy end, the severity of the drastic (pivot) moves should be progressively increased by increasing the parameter $\theta_{\max}$ associated with each simulation box.

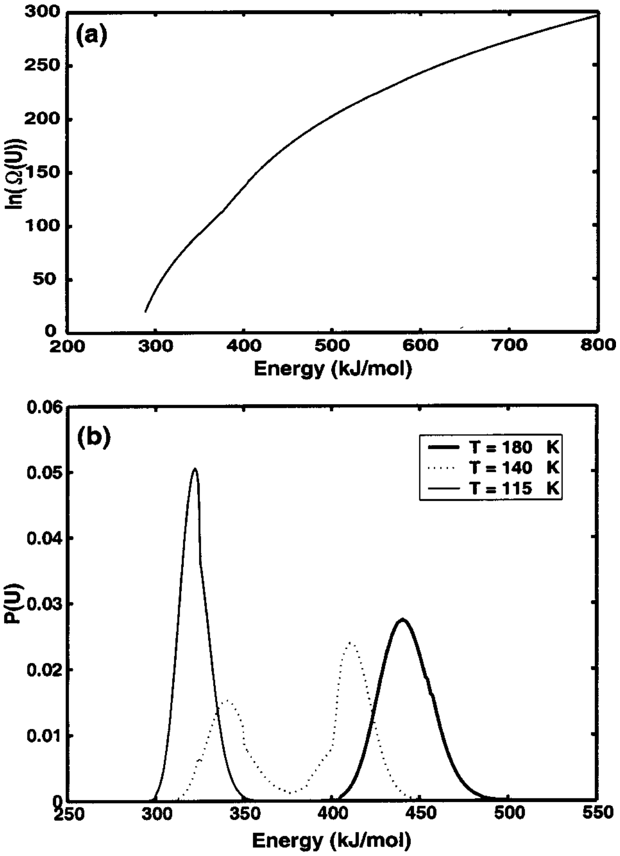

FIG. 1. Folding behavior of Met-enkephalin in implicit solvent (SASA): (a) Normalized density of states vs energy, (b) the normalized energy distribution below, at and above the transition temperature (≈140 K).

The density of states simulation, as outlined above, is performed independently in these boxes. However, in addition to the hybrid MD-MC and pivot moves in each box, the conformations in different replicas are now swapped at regular intervals. A swap move between the replicas in box i and j is accepted only if both the potential energies, U_i and U_j, lie in the overlap region of the two corresponding windows. A small window size in the folded regime and a low $\theta_{\max}$ value will ensure that we always maintain a copy of the near-folded conformation. Window sizes on the other extreme and the corresponding $\theta_{\max}$ should be large enough to facilitate major rearrangements of protein structure. The windows in the transition regime should be sufficiently small to achieve faster convergence. Since the configurations in different boxes are periodically swapped, the scheme ensures that systems in smaller windows do not get trapped in a low probability configuration due to the bounds imposed by the window size. If the random walk is performed in smaller energy windows with no swap moves, convergence can still be achieved but the sampling is insufficient, resulting in an incorrect estimate of $\Omega(U)$ near the ends of the energy range. This is evidenced by the non-superimposable nature of density of states data obtained in the overlap range. By introducing the periodic swaps, we alleviate this problem and the resulting $\Omega(U)$ can now be merged by superimposition of the overlapping energy ranges.

Once $\Omega(U)$ is known, thermodynamic quantities such as

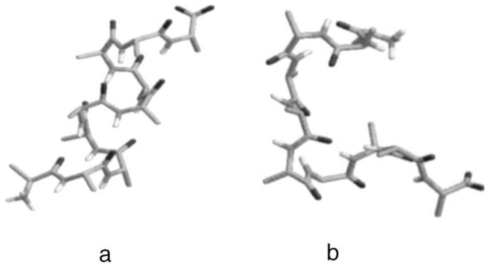

FIG. 2. United atom representation of structures of deca-alanine corresponding to the (a) folded energy regime ($T<T_{\mathrm{trans}}$) and (b) unfolded energy regime ($T>T_{\mathrm{trans}}$).

free energy $F(T)$, internal energy $U(T)$, entropy $S(T)$, and specific heat capacity $C(T)$ are determined according to

$$F(T)=-k_BT\ \ln\left(\sum\ \Omega(U)e^{-\beta U}\right), \tag{5}$$

$$U(T)=\langle U\rangle_T=\frac{\Sigma U\Omega(U)e^{-\beta U}}{\Sigma\Omega(U)e^{-\beta U}}, \tag{6}$$

$$S(T)=\frac{U(T)-F(T)}{T}, \tag{7}$$

$$C(T)=\frac{\langle U^2\rangle_T-\langle U\rangle_T^2}{k_BT^2}. \tag{8}$$

III. RESULTS AND DISCUSSION

Figure 1(a) shows the calculated density of states for Met-enkephalin in the presence of an implicit solvent. From the density of states, the energy distribution can be computed at any temperature of interest through

$$P(U)=\frac{\Omega(U)e^{-\beta U}}{\Sigma_i\Omega(U_i)e^{-\beta U_i}}. \tag{9}$$

Figure 1(b) shows such energy distributions for Met-enkephalin for temperatures above the transition temperature, near the transition temperature ($T_{\mathrm{trans}}\approx 140$ K) and below the transition temperature. Similar results were obtained for the density of states computation and the energy distribution for deca-alanine. As can be observed, the energy distribution is nearly Gaussian at temperatures sufficiently above and below T_{trans}, implying the existence of a well defined class of conformations. In Figs. 2 and 3, two representative conformations, corresponding to the folded (a) and unfolded (b) energy regimes are shown, for deca-alanine and Met-enkephalin, respectively. At T_{trans}, the energy distribution is double peaked, showing the existence of two classes of conformations.

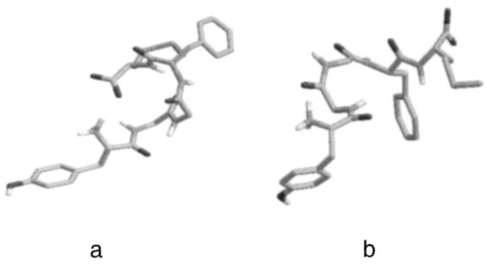

FIG. 3. United atom representation of structures of Met-enkephalin corresponding to the (a) folded energy regime ($T<T_{\mathrm{trans}}$) and (b) unfolded energy regime ($T>T_{\mathrm{trans}}$).

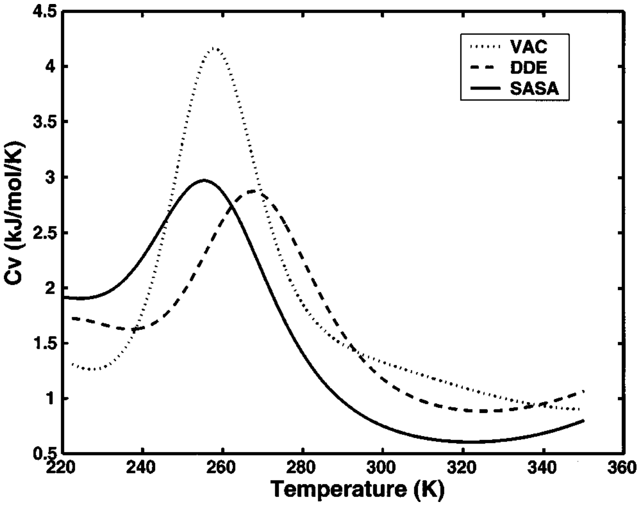

FIG. 4. Specific heat, $C_v(T)$, as a function of temperature for deca-alanine.

Figures 4 and 5 show the specific heat of deca-alanine and Met-enkephalin, respectively. Both molecules are studied in three different environments: in vacuum (VAC), in a distance dependent dielectric (DDE) and in an implicit solvent (SASA). All the specific heat plots have a distinct single peak which is associated with a transition from an unfolded to a folded state. We refer to the temperature corresponding to the maximum in the specific heat as the transition temperature for that protein. The figures show that the manner in which we account for the solvent (distance dependent dielectric or with an additional accessible surface area dependent

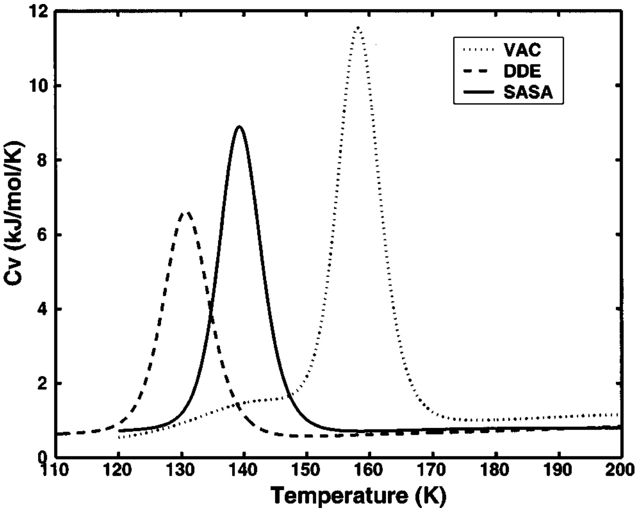

FIG. 5. Specific heat, $C_v(T)$, as a function of temperature for Met-enkephalin.

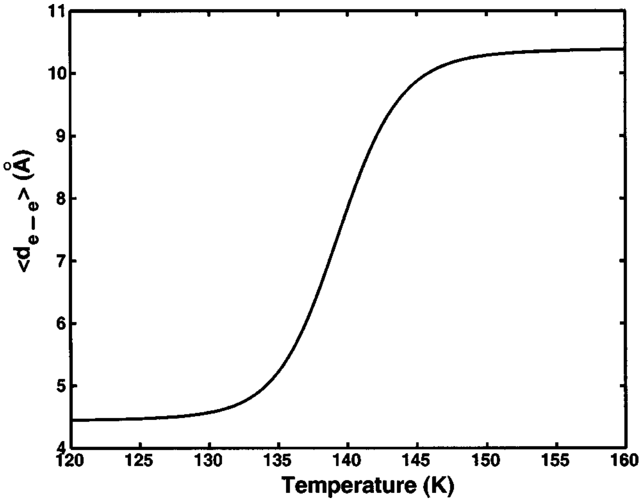

FIG. 6. Average end-to-end distance $\langle d_{e-e} \rangle$ as a function of temperature for Met-enkephalin in an implicit solvent (SASA).

solvation term) shifts the transition temperatures relative to the *in vacuo* simulations. The shift, however, is not consistent for different molecules. For Met-enkephalin (Fig. 5), solvation (DDE or SASA) effects tend to shift the transition towards lower temperatures, whereas no such pattern is evident for deca-alanine (Fig. 4). Peng and Hansmann[22] have conducted a thorough comparison of different solvation models and the effect of these solvation parameters on the folding of polyalanine with the ECEPP/2 force field. They conclude that the transition temperatures may shift towards lower or higher values depending on the choice of solvation parameters. It was also reported that some of these solvation models may even give rise to misleading results such as the complete absence of transition. With this in mind, our findings about the folding transitions for deca-alanine with a different model and potential function are not surprising.

In Fig. 6, the average end-to-end distance $\langle d_{e-e} \rangle_T$ is shown as a function of temperature; $\langle d_{e-e} \rangle_T$ here is defined as the average distance between the nitrogen atom of the N-terminus and the oxygen atom of the C-terminus. It is evident that in all the three environments, vacuum (VAC), distance dependent dielectric (DDE), and implicit solvent (SASA), the peptide is extended at high temperatures and compact at low temperatures. Our data are in good agreement with earlier simulation studies of Met-enkephalin.[20]

In Figs. 7 and 8 we display the different energy components as a function of temperature for the two molecules in the three different environments. The average covalent energy $\langle E_{\text{covalent}} \rangle$ here refers to the sum of the following energy components: $\langle E_{\text{bond}} \rangle$, $\langle E_{\text{bend}} \rangle$, $\langle E_{\text{tors}} \rangle$, and $\langle E_{\text{impr}} \rangle$, which are computed based on expressions given in Eq. (1).

To estimate the statistical error in our calculations of the folding transition temperature, four independent simulations were conducted for the case of Met-enkephalin in vacuum. All these density-of-states simulations were performed starting from the same initial estimate of $\Omega(U)$ and an intermediate value for the initial modification factor $[f_{\text{initial}} > \exp(10^{-4})]$. The simulations were performed with the same code but using different strings of random numbers. The resulting four independent estimates of $\Omega(U)$ were used to compute thermodynamic quantities. Our calculations indicate that errors in $\Omega(U)$ and in the resulting thermodynamic quantities are small. For the specific case of Met-enkephalin in vacuum, we find that the transition temperature is $T_{\text{trans}} = 157.9 \pm 2.1$ K; the average total internal energy at T_{trans} is $\langle U_{\text{total}} \rangle = 186.4 \pm 1.5$ kJ/mol and the specific heat at T_{trans} is $C_v = 12.9 \pm 0.9$ kJ/mol/K (where all the uncertainties correspond to one standard deviation).

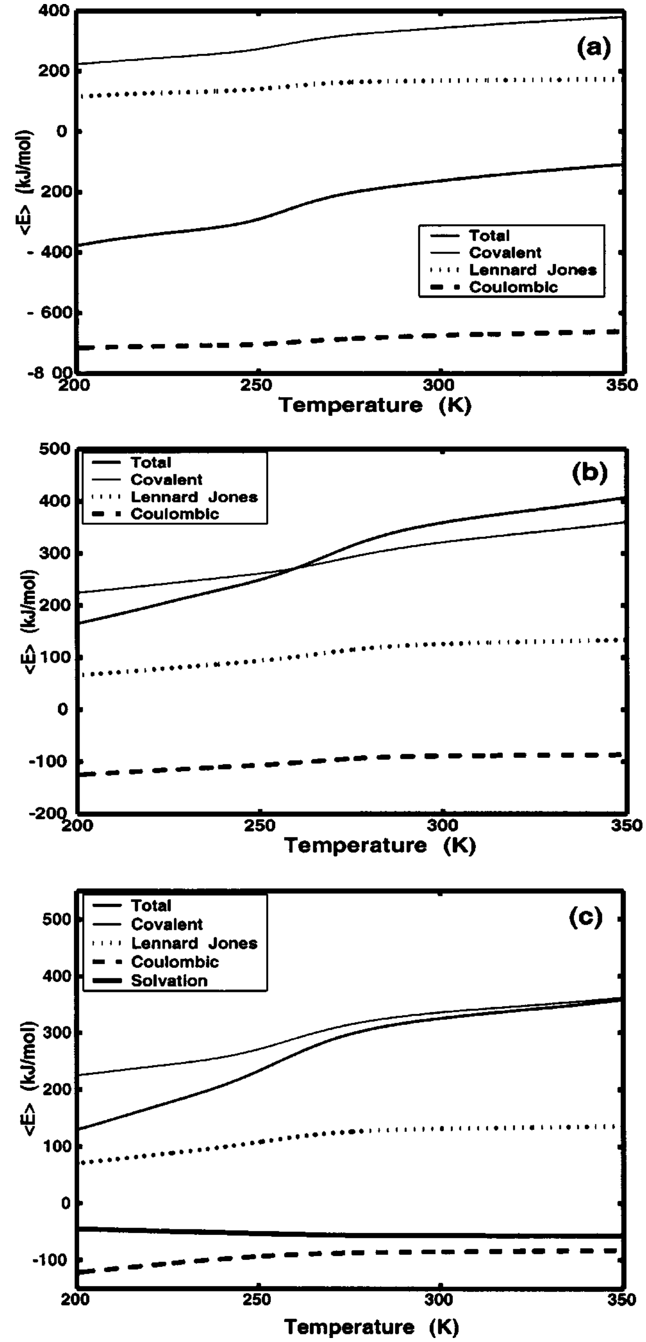

FIG. 7. Temperature dependence of the average energies (total, covalent, Lennard-Jones, Coulombic, and solvation) for deca-alanine in three different environments: (a) vacuum (VAC), (b) distance dependent dielectric (DDE), and (c) implicit solvent (SASA).

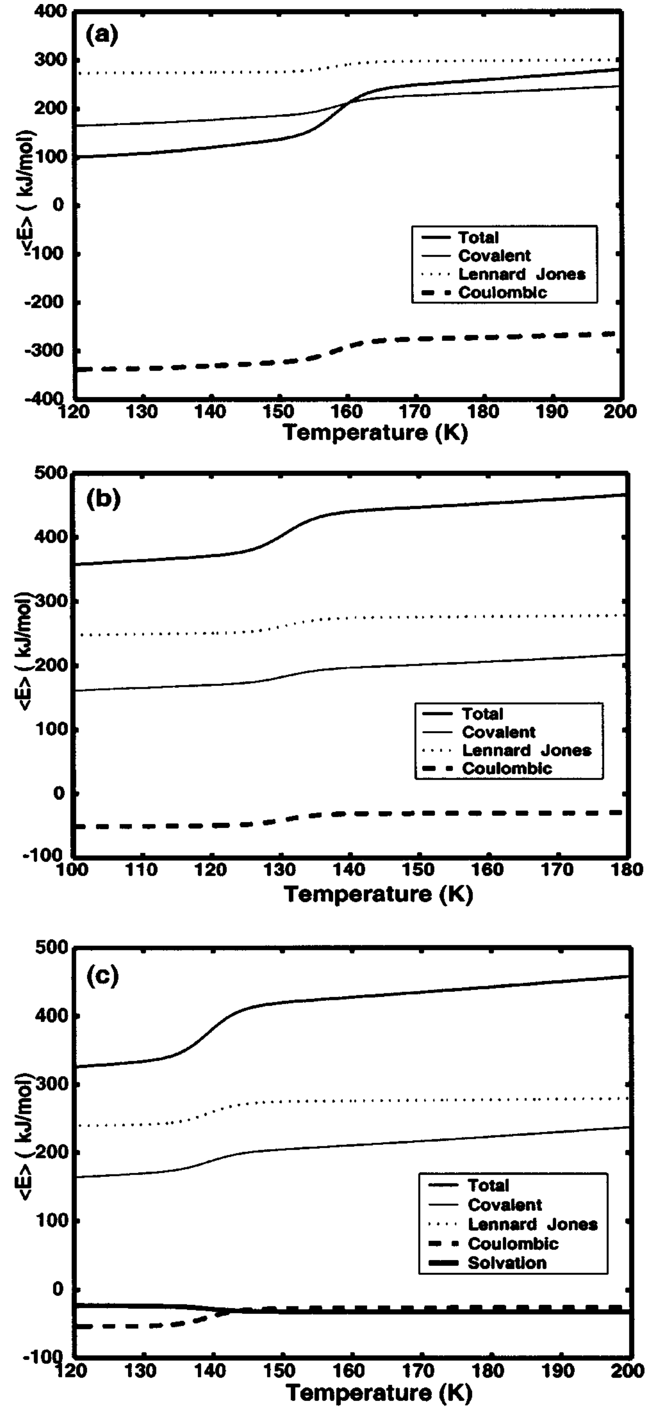

FIG. 8. Temperature dependence of the average energies (total, covalent, Lennard-Jones, Coulombic, and solvation) for Met-enkephalin in three different environments: (a) vacuum (VAC), (b) distance dependent dielectric (DDE), and (c) implicit solvent (SASA).

The method proposed here provides sufficiently high resolution to look into protein folding transitions. Solvation effects seem to have a significant effect on the transition temperature of the peptides. This is evident from Figs. 4 and 5 and also previous studies.[22] The SASA implicit solvent model employed in this work also has its limitations. The solvation parameters do not have a temperature dependence and the effective dielectric constant between partial changes does not depend on the environment. A more precise treatment of dielectric effects, either with an improved implicit solvent model or with explicit water molecules should lead to more realistic results. However, here we merely point out that our algorithm is able to reveal conformational details of higher resolution and accuracy than is possible by more traditional simulation techniques.

IV. CONCLUSION

A modified random walk technique based on sampling according to the density of states has been proposed and implemented to study folding transitions of short peptides in a continuum. The usefulness of the method is illustrated on two small peptides: deca-alanine and Met-enkephalin, for which we also compute a number of thermodynamic quantities as a function of temperature. We have also looked into the effect of solvent treatment [through a distance dependent dielectric and an implicit solvent model (SASA)] on the helix–coil and β turn–coil transitions. A more detailed study of these transitions using all atom representation for proteins and explicit solvent molecules is now under way.

ACKNOWLEDGMENTS

This work was supported by NSF (CTS-0210588 and EEC-0085560). We would also like to acknowledge Professor Qiang Cui and Dr. Qiliang Yan for helpful discussions.

[1] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).
[2] F. Yasar, T. Celik, B. A. Berg, and H. Meirovitch, J. Comput. Chem. **21**, 1251 (2000).
[3] Q. L. Yan and J. J. de Pablo, J. Chem. Phys. **113**, 1276 (2000).
[4] D. Gront, A. Kolinski, and J. Skolnick, J. Chem. Phys. **113**, 5065 (2000).
[5] Y. Okamoto, Int. J. Mod. Phys. C **10**, 1571 (1999).
[6] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **54**, 5863 (1996).
[7] F. A. Escobedo and J. J. de Pablo, J. Chem. Phys. **105**, 4391 (1996).
[8] B. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).
[9] F. Wang and D. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[10] F. Wang and D. P. Landau, Phys. Rev. E **64**, 056101 (2001).
[11] N. Rathore and J. J. de Pablo, J. Chem. Phys. **116**, 7225 (2002).
[12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[13] T. Lazaridis and M. Karplus, Proteins **35**, 133 (1999).
[14] P. Ferrara, J. Apostolakiz, and A. Caflisch, Proteins **46**, 24 (2002).
[15] Y. Okamoto and U. H. E. Hansmann, J. Phys. Chem. **99**, 11276 (1995).
[16] J. F. Griffin, D. A. Langs, G. D. Smith, T. L. Blundell, I. J. Tickle, and S. Bedarkar, Proc. Natl. Acad. Sci. U.S.A. **83**, 3272 (1986).
[17] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[18] U. H. E. Hansmann, M. Masuya, and Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. **94**, 10652 (1997).
[19] U. H. E. Hansmann, Eur. Phys. J. B **12**, 607 (1999).
[20] U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger, Chem. Phys. Lett. **259**, 321 (1996).
[21] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison–Wesley, Reading, 1980), Chap. 4.
[22] Y. Peng and U. H. E. Hansmann, Biophys. J. **82**, 3269 (2002).