Molecular Simulation, Vol. 33, Nos. 6–8, 15 May–15 July 2007, 577–582

Taylor & Francis
Taylor & Francis Group

Protein folding with the adaptive tempering Monte Carlo method

XIAO DONG†, DMITRI KLIMOV†‡ and ESTELA BLAISTEN-BAROJAS†*

†Computational Materials Science Center, College of Science, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA
‡Department of Bioinformatics and Computational Biology, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

(Received June 2006; in final form July 2006)

Characterization of the folding transition in a model protein was achieved with the recent multicanonical tempering method implemented with Monte Carlo the adaptive tempering Monte Carlo (ATMC) (X. Dong and E. Blaisten-Barojas. Adaptive tempering Monte Carlo method. *J. Comput. Theor. Nanosci.*, **3**, p. 118 2006). The folding transition temperature was successfully determined and a spread of states was observed around the interface between native and folded regions. Energy states collected from all tempering events in a series of parallel runs were used in the calculation of the free energy, internal energy, order parameter and radius of gyration as a function of temperature through the weighted histogram method. Not only the calculated thermodynamic properties are in good agreement with results from Langevin dynamics simulations (D. K. Klimov and D. Thirumalai. Native topology determines forced-induced pathways in global proteins. *Proc. Natl. Acad. Sci. USA*. **97**, p. 7254 2000), but also this multicanonical approach is noticeably more efficient because of the adaptive manner in which the system visits states near a transition in the interface between two phases. Additionally, the ATMC is advantageous for protein simulation over regular single canonical ensemble methods because it accelerates the hopping between local energy minima on the potential energy surface.

Keywords: Protein folding; Tempering; Monte Carlo; Multicanonical; ATMC

1. Introduction

Characterization of the thermodynamics and kinetics of the protein folding transition remains a challenging problem in protein physics due to numerous local minima of the protein potential energy landscape (PES) [1–3]. A similar difficult problem arises in finding the global minimum of nanoclusters, where the multitude of geometries associated with isomers have energies very close to each other [1]. In this recent publication, the adaptive tempering Monte Carlo (ATMC) method was developed and used for obtaining the global minimum of systems with rough PES. This new methodology excels by allowing for adaptive excursion-inquiries of the PES consistent with local changes in temperature, each temperature identifying a different canonical ensemble and thus setting the transitioning between ensembles smoothly. This strategy results in rapid discovery of complex topological paths on the PES leading towards the global minimum. The ATMC belongs to the family of multicanonical methods, also called simulated tempering [4,5]. The strategy of the ATMC is based on sequential swapping between canonical ensembles. This is advantageous when compared to parallel tempering methods such as the replica exchange parallel algorithm [6] because no synchronization is required between the processors associated to each of the tempering events. The sequential ATMC has been successful in finding the most ordered state for atomic nanoclusters described both classically and quantum mechanically [7].

In this work, we extend the application of the ATMC to the characterization of polypeptide folding using a continuum minimal model representation. Additionally, we describe and use our new parallel version of the ATMC, which utilizes several processors to speed-up calculations without disturbing the advantageous sequential canonical sampling of the PES. It is demonstrated along this work that ATMC sampling allows for the prediction of the folding transition temperature and several thermodynamics properties of the model protein.

*Corresponding author. Email: blakisten@gmu.edu

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020600930532

The methodology is successful, thus providing a good new approach for the study of protein folding in other proteins. This paper is organized as follows: the protein model Hamiltonian is described in Section 1, a brief description of the ATMC is given in Section 2. Section 3 contains the results and the paper is concluded in Section 4.

2. Description of the protein model

Each amino acid in the protein is represented by a bead (figure 1) giving a coarse-grained representation of the polypeptide [2]. The beads are connected by pseudo chemical bonds to form a linear polypeptide chain. The amino acid sequence in such polypeptide is composed of three types of residues: the hydrophobic (B), the hydrophilic (L) and the neutral (N). The conformational potential energy of this model contains the bond length potential V_{BL}, bond angle potential V_{BA}, dihedral angle potential V_{DIH} and non-bonded potential V_{NON}:

$$E_{\mathrm{p}}(\{\vec{r}_i\}) = V_{\mathrm{BL}} + V_{\mathrm{BA}} + V_{\mathrm{DIH}} + V_{\mathrm{NON}} \tag{1}$$

where $\{\vec{r}_i\}$ represents the coordinates characterizing a peptide conformation, $i = 1, 2, \ldots, N_r$, and N_r is the number of beads in the polypeptide chain. V_{BL} has the following expression:

$$V_{\mathrm{BL}} = \sum_{i=1}^{N_r-1} \frac{k_r}{2}(|\vec{r}_{i+1} - \vec{r}_i| - a)^2 \tag{2}$$

where $k_r = 100\epsilon_{\mathrm{h}}/a^2$, $a = 3.8\,\text{Å}$ is the average bond length between two beads and $\epsilon_{\mathrm{h}} \approx 1.25\,\mathrm{kcal/mol}$ is the average strength of the hydrophobic interactions. The bond angle potential V_{BA} is taken to be

$$V_{\mathrm{BA}} = \sum_{i=1}^{N_r-2} \frac{k_\theta}{2}(\theta_i - \theta_0)^2 \tag{3}$$

where $k_\theta = 20\epsilon_{\mathrm{h}}/(\mathrm{rad})^2$ and $\theta_0 = 1.8326\,\mathrm{rad} = 105°$ is the average bond angle between three successive beads $i, i+1, i+2$. The dihedral angle ϕ_i is the angle between two planes, each of which is defined by beads i, $i+1$, $i+2$ and by beads $i+1$, $i+2$, $i+3$, respectively. Repulsive interactions between overlapping orbitals and steric overlap between atoms are contributing factors to a dihedral angle potential V_{DIH}, which describes the energy variation due to bond rotations;

$$V_{\mathrm{DIH}} = \sum_{i=1}^{N_r-3} [A_i(1 + \cos\phi_i) + B_i(1 + \cos 3\phi_i)] \tag{4}$$

where $A_i = 0$ and $B_i = 0.2\epsilon_{\mathrm{h}}$ when two or more than two of the four beads are neutral, otherwise $A_i = 1.2\epsilon_{\mathrm{h}}$ and $B_i = 1.2\epsilon_{\mathrm{h}}$. This choice of V_{DIH} facilitates chain flexibility and turn formation within the chain regions containing large numbers of N residues. In other regions, the polypeptide backbone has a propensity to adopt β-strand conformation. The non-bonded potential V_{NON} describes the pair interactions between the residues that are not covalently bonded:

$$V_{\mathrm{NON}} = \sum_{i=1}^{N_r-3} \sum_{j=i+3}^{N_r} V_{ij}(r) \tag{5}$$

where $r = |\vec{r}_i - \vec{r}_j|$ is the distance between beads i and j. Three cases arise depending on the combinations of the types of i and j in equation (5):

$$V_{L\alpha} = 4\epsilon_L\left[\left(\frac{a}{r}\right)^{12} + \left(\frac{a}{r}\right)^{6}\right], \quad \alpha = L \text{ or } B \tag{6a}$$

$$V_{N\alpha} = 4\epsilon_{\mathrm{h}}\left(\frac{a}{r}\right)^{12}, \quad \alpha = N, L, \text{ or } B \tag{6b}$$

$$V_{BB} = 4\lambda\epsilon_{\mathrm{h}}\left[\left(\frac{a}{r}\right)^{12} - \left(\frac{a}{r}\right)^{6}\right] \tag{6c}$$

where $\epsilon_L = 2/3\epsilon_{\mathrm{h}}$ and λ is a dimensionless parameter that introduces a Gaussian distribution in the strength of the hydrophobic interactions. The data reported in this work are in reduced units of ϵ_{h} for energy and a for length. The reducing unit of mass for the beads is $m_0 \approx 210^{-22}\,\mathrm{g}$.

Due to the proper distribution of B, L and N residues and λ factors, the potential function E_{p} encodes the formation of four strands β-barrel as the conformation with the lowest energy (native state) shown in figure 1. In this figure, hydrophilic, hydrophobic and neutral residues are colored in black, grey and white, respectively. A quasi order parameter Q for monitoring the structural changes in the simulations is introduced with the following definition: for the native state, all the inter-bead distances that appear in equations (6a–c) are calculated and those that fall within a cut-off distance of 1.8 are counted. In the native state, the value of Q is 106. For other states, Q changes value and decays to zero. In previous studies, several properties for this polypeptide were computed with Langevin dynamics and the WHAM multiple histogram analysis [2].

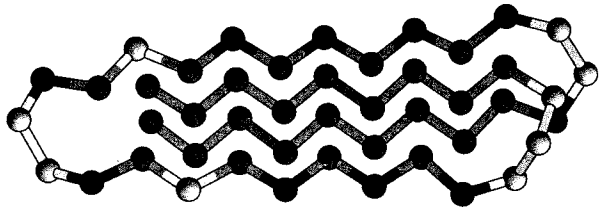

Figure 1. The native state of the polypeptide. Beads represent amionoacids. Colors identify neutral (white), hydrophobic (gray) and hydrophilic (black) residues.

3. The adaptive tempering method

The ATMC belongs to the family of expanded ensembles methods [4,5]. In these approaches, a multitude of different Gibbs ensembles are connected by different techniques. The ATMC links many canonical ensembles

by a super Markov chain which simulates how the temperature hops between two contiguous canonical ensembles with sub-partition functions Z_A and Z_B. The total partition function Z of the expanded ensemble is the sum of n of such sub-partition functions:

$$Z = \sum_{i=1}^{n} Z_i \exp(-\eta_i) \tag{7}$$

where Z_i is the partition function associated with the ith canonical ensemble and η_i is the corresponding weight.

In the ATMC method [1], the system accesses a multitude of NVT_i (canonical) ensembles and each T_i characterizes a different canonical ensemble. Each canonical ensemble is simulated with the standard Metropolis Monte Carlo (MC) algorithm for a fixed number, M_{fixed}, of steps on the PES. The canonical ensembles are connected along the simulation such that their characteristic temperature T is allowed to hop to either $T + \Delta T$ or $T - \Delta T$ ($\Delta T > 0$) with probabilities:

$$\pi_+ = \frac{\exp[-(E - \langle E\rangle)(1/k_B(T + \Delta T) - (1/k_B T))]}{W} \tag{8a}$$

$$\pi_- = \frac{\exp[-(E - \langle E\rangle)(1/k_B(T - \Delta T) - (1/k_B T))]}{W} \tag{8b}$$

Here E is one of the M_{fixed} instantaneous energy points of the PES corresponding to one state building the canonical ensemble with temperature T, $\langle E\rangle$ is the Metropolis-MC average energy of the M_{fixed} states and W is a normalization factor. After some algebraic rearrangements, a new canonical ensemble adapted to the location of the system in the PES is determined by characterizing its new temperature (N and V continue to be the same) as:

$$T_{\text{adapt}} = \frac{(T^2 - \Delta T^2)}{\Delta T} \tag{9}$$

where ΔT is:

$$\Delta T = \frac{T}{1 - \delta E/(\ln(a) k_B T)} \tag{10}$$

Here δE is the standard deviation of the energy about the average $\langle E\rangle$ at temperature Tover the M_{fixed} NVT-MC trials. One parameter $\ln(a)$ is introduced, in terms of which T_{adapt} is readily obtained. A large $\ln(a)$ results in a fast excursion to low temperatures whereas a small $\ln(a)$ leads the system to a larger number of temperature changes. Full details about this method have been published elsewhere [1,7].

In what follows a *tempering event* refers to the process of hopping between temperatures. Therefore, along the simulation the tempering events are numbered sequentially indicating how the system visits the range of temperatures spanned by the simulation. After each tempering event, the system evolves for another M_{fixed} NVT-MC steps at the new chosen temperature (either $T + \Delta T$ or $T - \Delta T$). All simulations were started from different high temperature random conformations of the protein. The simulation is stopped when the temperature is close to zero. The potential energy E is calculated consistent with the model described in Section 2.

All simulations in this work were done with the parallel implementation of the ATMC. This approach allows starting the ATMC from several protein configurations, each processor handling the tempering from a different initial condition. The parallel process permits simultaneous excursions on the PES, which accelerates the calculations with an almost perfect speedup. Segments of data are appended when a temperature is repeated, ensuring coordination between the processors.

It is important to compare the ATMC method with the temperature replica exchange method (TREM), which is widely used in biomolecular simulations and often referred as the parallel tempering method [6]. Several advantages over TREM methodologies are offered by ATMC: (i) ATMC can be implemented either in parallel or serial computations. In particular, ATMC is well suited for grid computing using stand-alone computers connected via low-speed networks. In contrast to TREM, we should emphasize that ATMC does not require dedicated multiprocessor clusters, in which all replicas are to be simulated concurrently; (ii) the temperature in ATMC is a dynamically adjusted variable and the predetermined grid of temperature values required in TREM is not needed; (iii) the ATMC method can perform two tasks simultaneously—a collection of conformational states for the weighted histogram analysis method (WHAM) [8] computations and a search for the ground state of the system. Although a rigorous comparison of the efficiency of conformational sampling provided by TREM and ATMC is beyond the scope of this paper, it is clear that ATMC can be used as a precursor for TREM simulations designed to provide a rough mapping of system's equilibrium behavior and to identify the global minimum corresponding to the ground state. Based on this mapping the distribution of temperature values in TREM can be better established. Therefore, we believe that ATMC can be a useful addition to the algorithmic toolkit available for computational biophysicists.

4. Protein folding process

The optimum values for the two ATMC parameters are $M_{\text{fixed}} = 1000$ and $\ln(a) = -1$. These parameters were determined after a few trials to ensure that the ground state is not reached too fast and are used throughout in this work. Figure 2 shows results from one of the parallel simulations illustrating the evolution of the temperature (a), the potential energy (b) and the order parameter Q (c) as a function of tempering event. There are 5079 tempering events in this case, showing that the polypeptide explores a multitude of conformations. A correlation plot between the potential energy and the temperature is shown in figure 3(a) and the correlation between the order parameter and temperature is shown in figure 3(b). In these two figures, each dot corresponds to one of the 5079 tempering events. Figure 3(a) is a collection of eight simulations showing that

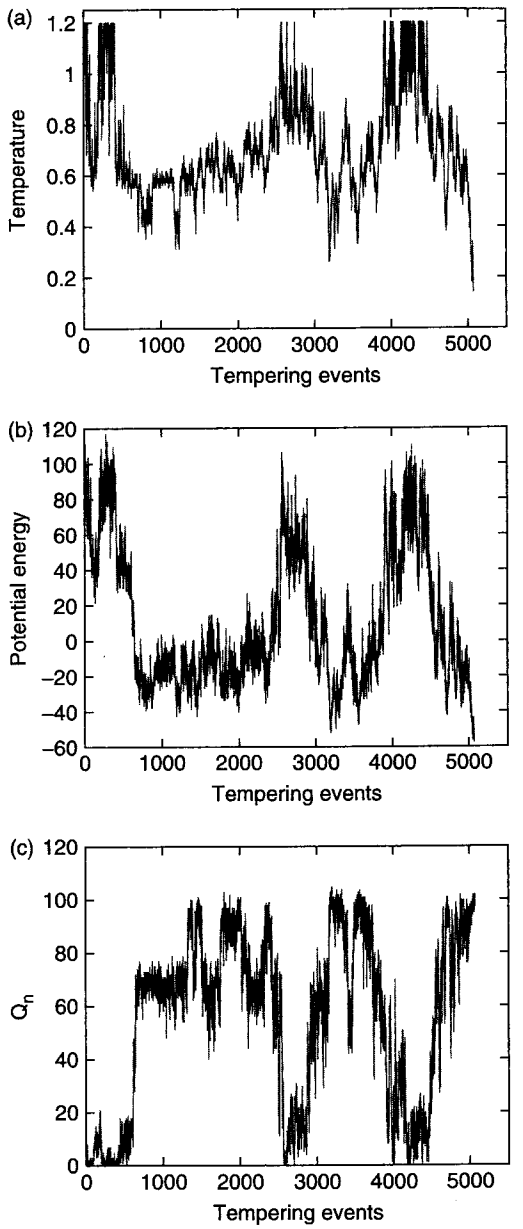

Figure 2. Properties evolution along the tempering simulation: (a) temperature; (b) potential energy; (c) order parameter; and (d) radius of gyration. Reduced units are used throughout.

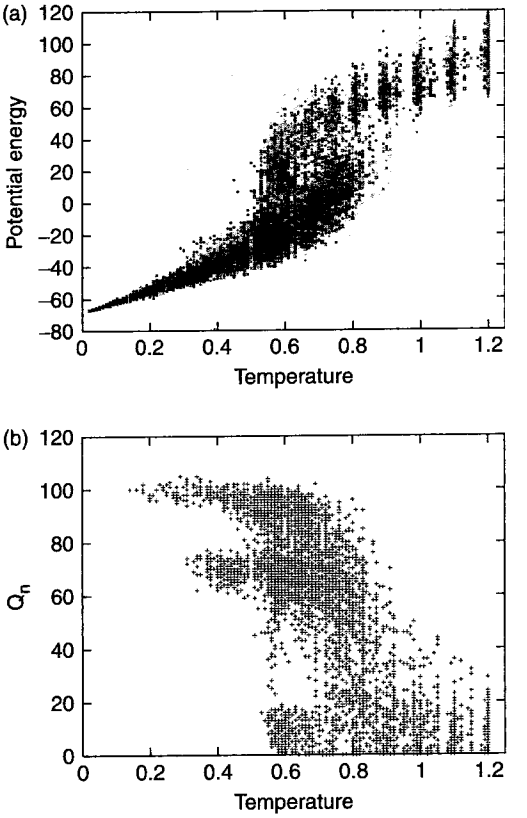

Figure 3. Correlation plots: (a) between temperature and total potential energy from eight parallel runs; (b) between temperature and order parameter for the simulation shown in figure 2. Reduced units are used throughout.

folding and unfolding of the protein occurs around the transition temperature of 0.79 [2]. Indeed, figure 3(a) shows a transition region associated with fluctuating states around the interface between the protein native and unfolded regions. The interface region spans from $T = 0.5$ to 0.9.

Details of the folding mechanism are clarified in figure 3(b) where the order parameter shows the formation of intermediate partially folded structures. These partially folded structures appear in all simulations. It is apparent that the interface region around the folding transition temperature is due to excursions between the native structure and partially unfolded structures with $Q \approx 70$. Further analysis of the sequence of events indicates that when this partially folded protein was tempered to a slightly higher temperature, the subsequent temperature changes emulated an annealing towards lower temperatures that produced full folding into the native state. This scheme of annealing–tempering-annealing occurred several times populating the interface region.

Simulation with the ATMC is fast for collecting thermal data because each thread (processor) performing one simulation contributes to the mutual sampling of the PES. These data can be used to study several thermodynamic functions based on multiple histogram analysis. The WHAM [8] has been a standard tool for predicting thermodynamic properties from independent *NVT* simulations. To illustrate the use of ATMC data with WHAM, histograms of the energies at all tempering events of the ATMC simulations were built containing 71,339 states. For comparison, six independent *NVT*-MC simulations at temperatures $T = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1$ were performed, which accumulated histograms with two million states for each temperature. Comparison of the two types of histograms of the potential energy is shown in figure 4. The smooth curve identifies the sampling done with regular *NVT*-MC and the jagged curve corresponds to the sampling from the ATMC. The major features of the histograms are similar considering that sampling with the *NVT*-MC was not done at temperatures below 0.4 and above 1.1. In general, figure 4 shows that ATMC samples diverse regions of the PES.

Next, the energy histograms from these two approaches (*NVT*-MC and ATMC) were used in the WHAM

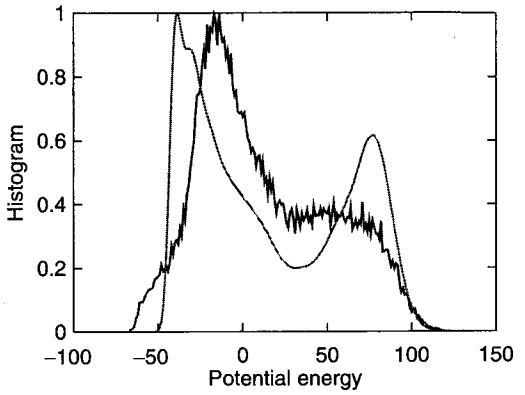

Figure 4. Histogram of the energy states accessed by the *NVT*-MC at $T = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1$ (smooth line) and by the ATMC (jagged line).

calculations to estimate the temperature dependence of various thermodynamic quantities. The essence of the WHAM method consists in adding a pre-factor weight to the Boltzmann factor in the canonical ensemble. Then the average value of a quantity such as E at a fixed temperature T' for a single histogram $h(E;T)$ collected at various temperatures T is:

$$\langle E(T')\rangle = \frac{\sum_E E\Omega(E)\exp(-\beta' E)}{\sum_E \Omega(E)\exp(-\beta' E)} \tag{11}$$

$$\Omega(E) = h(E;T)Z\exp(\beta E) \tag{12}$$

where $\beta = (k_B T)^{-1}$ and Z is the canonical partition function. Then WHAM assumes that many independent simulations are carried out at R predefined temperatures and a consolidated histogram with M bins is constructed by merging data from all the simulations. It is assumed that the histograms of E at each of the predetermined temperatures are independent. Consequently, the β-scaled free energy $f = \beta A$ and the thermal average of a property such as Q are given at a temperature i by:

$$\exp(-f_i) = \exp(-\beta_i A_i) = Z_i \tag{13a}$$

$$Z_i = \sum_{k=1}^{M} \frac{h_k \exp(-\beta_i E_k)}{\sum_{m=1}^{R} n_m \exp(\beta_m A_m - \beta_m E_k)} \tag{13b}$$

$$Q_i = \sum_{k=1}^{M} \frac{q_k \exp(-\beta_i E_k)/Z_i}{\sum_{m=1}^{R} n_m \exp(\beta_m A_m - \beta_m E_k)} \tag{13c}$$

where q_k is the kth bin of the histogram for the quantity Q. Equation (13b) is solved iteratively to calculate the free energy and thereof equation (13c) is used to calculate the thermodynamic properties. One has to be aware that the WHAM requires input from many independent simulations. There have been studies [9] addressing the application of WHAM with data obtained from simulated tempering and parallel tempering and the problem arising due to the fact that tempering data might not be independent. Results in this work contribute to elucidate this point.

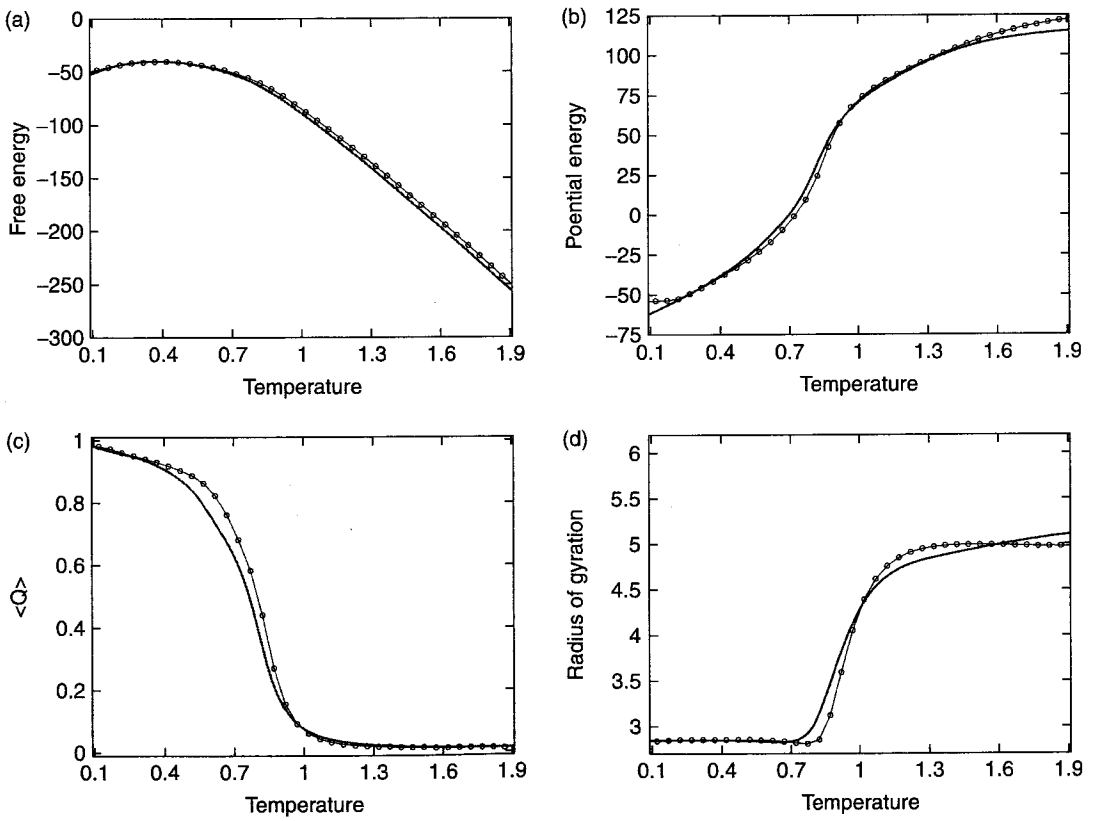

Figure 5. Temperature dependence of the free energy (a); potential energy (b); order parameter (c); and radius of gyration (d) estimated with WHAM using the states from *NVT*-MC simulations (circles) and the states from ATMC (continuous line). Here, the radius of gyration is defined as the arithmetic average of the squared distances of the beads from the center of mass.

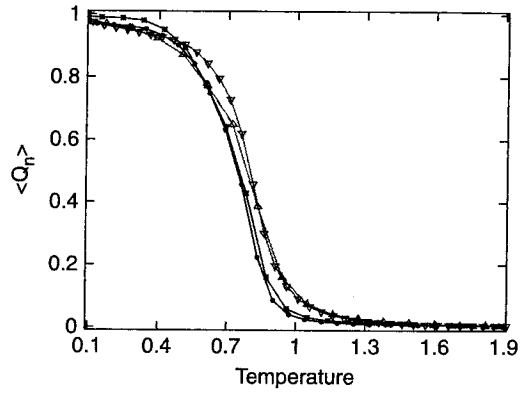

Figure 6. The four partial order parameters Q_n as a function of temperature calculated with WHAM using data from ATMC. Triangles-line are for the two terminal strands and dotted-lines are for the two central strands.

Figure 5 shows the comparison of the free energy, potential energy, order parameter and radius of gyration calculated with WHAM using the histograms built with the states from *NVT*-MC (dots) and ATMC (continuous line) simulations. The free energies from the two calculations are almost identical. It is to be noted that the free energy temperature profile obtained from WHAM in figure 5 is also consistent with that obtained previously from Langevin dynamics [2]. Because the range of temperatures sampled with the *NVT*-MC simulations does not contain temperatures below 0.4 or above 1.1, WHAM calculated properties are valid within these temperatures. This limitation is visible in figure 5(b–d). Previous studies identified the folding transition temperature as 0.79 [2]. The ATMC sampling used to obtain the quantities represented in figure 5 is consistent with previous results [2].

The order parameter Q could be divided into four similarly defined partial order parameters Q_n, each of them associated with one of the nth strands in the native conformation. Figure 6 illustrates the temperature dependence of these four partial order parameters calculated within the WHAM approach and with the ATMC sampling. It is evident that around the folding transition temperature, all the strands fold cooperatively at the temperature close to the folding transition temperature. This result demonstrates all-or-none characteristics of this model.

The excellent agreement of ATMC data with the standard *NVT*-MC WHAM approach indicates that the states from all the tempering events are independent. There is a much richer set of histograms from the PATMC than from the *NVT*-MC. In our case, the former gave rise to 120 histograms, whereas the latter was based on 6 histograms. This confirms that ATMC is very efficient for sampling a wide range of energy states.

5. Conclusion

In this paper, we demonstrated the use of the ATMC allowed for characterizing the folding transition in a continuum minimal representation model protein. The folding transition temperature was successfully identified and the collection of tempering events suggests that the folding transition is weakly first order. It was also demonstrated that performing several parallel simulations enriches the sequential tempering data accumulated in each ATMC run. Data gathered from the ATMC parallel runs are useful for further estimates of the protein thermodynamic quantities based on the weighted multi histogram method.

The parallel new implementation of the ATMC is an appropriate approach for systems that undergo complex phenomena, such as protein folding, where the PES needs to be sampled extensively. The result of our study shows that ATMC represents a new efficient method for sampling conformational space in biomolecules and alerts the molecular simulation audience of its potential success in protein folding searches. The general nature of ATMC algorithm makes it applicable to more detailed models of proteins with ragged energy landscape and also opens a new avenue for adaptive tempering implementations of isothermal molecular dynamics, including Langevin dynamics.

Acknowledgement

XD acknowledges financial support form the School of Computational Sciences. All computations were done in the Supercomputing Center of the College of Science at George Mason University.

References

[1] X. Dong, E. Blaisten-Barojas. Adaptive tempering Monte Carlo method. *J. Comput. Theor. Nanosci.*, **3**, 118 (2006).
[2] D.K. Klimov, D. Thirumalai. Native topology determines force-induced pathways in global proteins. *Proc. Natl. Acad. Sci. USA*, **97**, 7254 (2000).
[3] J.M. Onuchic, P.G. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70 (2004) D. Thurumalai and D.K. Klimov. Deciphering the time scales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.* **9**, 197 (1999).
[4] E. Marinari, G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, **19**, 451 (1992).
[5] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: method of expanded ensembles. *J. Chem. Phys.*, **96**, 1776 (1992).
[6] Y. Sugita, Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**, 141 (1999).
[7] X. Dong, S. Gatica, E. Blaisten-Barojas. Tight-binding calcium clusters from adaptive tempering Monte Carlo simulation. *Comput. Lett.*, **1**, 152 (2005).
[8] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, **13**, 1011 (1992).
[9] J.D. Chodera, W.C. Swope, J.W. Pitera, Ch. Seok, K.A. Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theor. Comput.*, in press.