Efficient scheme for optimization of parallel tempering Monte Carlo method

2007 J. Phys.: Condens. Matter 19 036225

(http://iopscience.iop.org/0953-8984/19/3/036225)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 036225 (9pp) doi:10.1088/0953-8984/19/3/036225

Efficient scheme for optimization of parallel tempering Monte Carlo method

Dominik Gront and Andrzej Kolinski

Warsaw University, Faculty of Chemistry, Pasteura 1, 02-093 Warsaw, Poland

E-mail: dgront@chem.uw.edu.pl

Received 4 October 2006, in final form 12 December 2006
Published 5 January 2007
Online at stacks.iop.org/JPhysCM/19/036225

Abstract

The parallel tempering (PT) Monte Carlo sampling scheme has already been applied to studying different systems, including spin glasses and biomolecules. In this work we examine the efficiency of PT simulations and propose an iterative procedure for the optimal selection of the replicas' temperatures. The method returns a set of temperatures for a PT simulation for which the overlap of the distribution of states (referred to as an overlap ratio) measured for every pair of adjacent replicas remains constant. The computational procedure is tested for two distinct simplified molecular models of polypeptides. The method is based on the most fundamental thermodynamic properties and therefore it could be applied to virtually any system governed by the canonical ensemble.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

By means of molecular dynamics or Monte Carlo sampling it is possible to generate an ensemble of states corresponding to a Boltzmann distribution at a given temperature. However, this is insufficient to obtain a full thermodynamic description of a model system. To derive thermodynamic properties, one has to follow one of the two possible methods: sampling in the multicanonical ensemble [1, 2] or using a re-weighing technique. In a multicanonical simulation the modelled system performs a random walk in energy space that allows sampling of multiple local minima. During such a simulation, the density of states as a function of energy has a uniform distribution. In order to achieve such a distribution the conventional Boltzmann weights must be replaced by different, non-Boltzmann weights. From these weights one can retrieve a full thermodynamic description of a given system. The weights, however, unlike in the canonical ensemble, are not known *a priori*. Usually they are determined in an iterative procedure and a substantial part of the simulation time must be spent on this task. This is a significant limitation of the multicanonical method. Another very efficient method for conformational sampling in systems with rugged energy landscapes that has

been proposed recently is parallel tempering (abbreviated in this work as PT, also known as replica-exchange Monte Carlo or Metropolis-coupled chain), where multiple copies of a model system are sampled at various temperatures. The method was discovered independently by computer scientists working for the fifth-generation computer project [3], statisticians [4] and physicists [5, 6]. The PT algorithm has been applied successfully to simplified [7] and realistic peptide models [6]. PT could be combined with molecular dynamics [8] or with multicanonical Monte Carlo [9]. Densities of states obtained for several different temperatures may be used to compute the full density of states of a given system. This goal may be achieved by means of the multihistogram method [10–12]. However, the convergence of the multihistogram method depends on mutual overlap of two distributions obtained for two adjacent temperatures. The set of temperatures for PT simulation is not known *a priori*. Several strategies for its selection have been proposed in the literature [13–16]. They follow two basic assumptions: constant overlap between distributions of states [15, 13] and maximum flow in the replica space [16, 17]. So far, the PT temperature sets were optimized for the purpose of efficient energy minimization. In this work we show how to derive an optimal set of temperatures that allows for accurate calculations of the models' thermodynamics. Our approach is more general, since it is based on the observed distribution of states rather than on its Gaussian approximation (typical for the previous work).

2. Materials and methods

2.1. Protein models

In order to test the method presented in this work, we perform folding simulations using two distinct simplified protein models of various levels of structural details. The first one is a lattice Go polypeptide model [18] of 56 amino acid 2gb1 protein. The Go model provides a funnel-like energy landscape of a sampled system. Such energy functions have been successful in describing the configurational ensembles for a protein during folding toward its native basin. This simplified energy landscape makes conformational sampling very efficient. Moreover, the choice of the Go model ensures that the native structure corresponds to the global minimum of energy. Those features allow for a very accurate assessment of our technique. Protein conformations for this model are restricted to a very flexible hybrid (310) lattice [19]. The possible orientations of the virtual Cα–Cα bonds belong to the set of 90 vectors of the following type: (3, 1, 1), (3, 1, 0), (3, 0, 0), (2, 1, 1) and (2, 2, 0). Assuming that the lattice unit is equal to 1.22 Å, protein data bank (PDB) structures of real polypeptides could be fitted to the lattice with average coordinate root-mean square deviations (cRMSD) from the crystallographic structures in the range 0.6–0.7 Å. The only interaction controlling the sampling is a well-type potential for the $N \times N$ distances between the alpha carbons in the N-amino acids' native structure (see figure 1) and hard core repulsions for the united atoms centred on the alpha carbons. The side chains in this model are ignored. The hybrid lattice model has been applied extensively in studies of various aspects of protein folding, dynamics and thermodynamics [20].

The second model is based on the CAlpha, Beta, Side chains (CABS) [21] representation and force field. The model was used successfully in studies of protein dynamics [30], thermodynamics and the prediction of the structures of proteins [29] and protein assemblies [32]. Polypeptide chains in the CABS representation are restricted to a high-resolution lattice with up to four united atoms per residue. The accuracy of representation is of the range of 0.35 Å (cRMSD for the alpha carbon trace). The force field of the CABS model consists of a set of knowledge-based potentials derived from the statistical analysis of structural regularities seen in already solved protein structures. Here we studied the folding

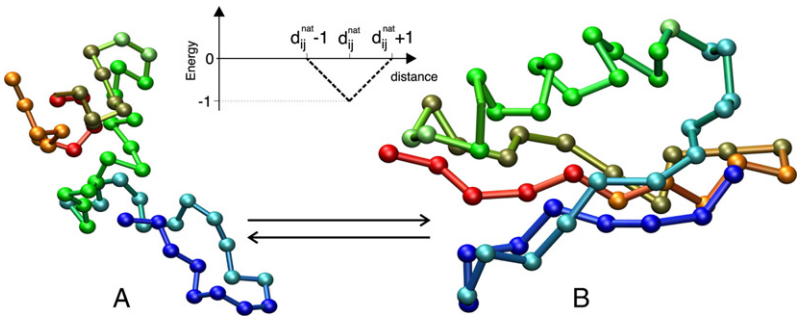

Figure 1. Two example conformations: partially folded (A) and a near-native (B) of 2gb1 protein in a hybrid representation, observed during a Monte Carlo simulation at the transition temperature. The inset explains the Go-type potential used in simulation. The distance is given in Å, and d_{ij}^{nat} is the distance between ith and jth α-carbons in the native structure.

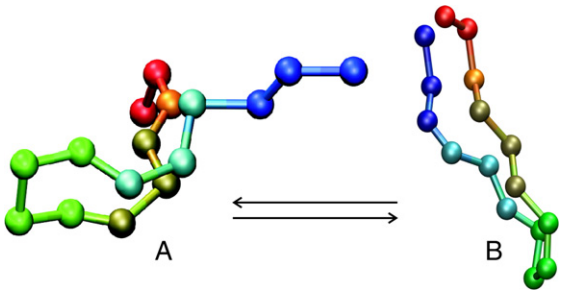

Figure 2. Cα trace of two example conformations of the GB1P peptide: partially folded (A) and near-native (B). Both conformations correspond to the same folding temperature.

transition of a protein fragment. Experiments have shown that a C terminus, 16 residues long, fragment (GEWTYDDATKTFTVTE) from the 2gb1 protein (usually denoted as GB1P) forms a β-hairpin on its own [22]. The peptide has been extensively studied experimentally [22, 23] and its kinetic and thermodynamic properties are well characterized. The GB1P peptide has also been studied by means of computer simulations [24, 31]. For a recent summary of these studies, see [25]. Figure 2 shows two representative conformations of the GB1P peptide at the folding temperature.

2.2. Parallel tempering Monte Carlo scheme

Protein models are simulated by means of parallel tempering Monte Carlo. This method relies on simultaneous simulations of $\mathcal{K}$ non-interacting replicas (copies) of a system. Each replica is placed in a separate box at a different temperature T_i. In addition to standard Monte Carlo or molecular dynamics moves that affect only one copy, PT also allows (with a certain probability) the exchange of conformations between two adjacent copies. The resulting random walk in temperature enables conformations to cross energy barriers and move out of local minima, enhancing sampling of the low-energy structures. Therefore PT is known as one of the most efficient Monte Carlo-based minimization schemes [7]. Parallel tempering can also be considered as a single process with a combined probability distribution described as a product of several Boltzmann distributions that describes separate replicas:

$$P(x_1, x_2, \ldots, x_{\mathcal{K}}) = \prod_{i=1}^{\mathcal{K}} P_{T_i}(E) = \prod_{i=1}^{\mathcal{K}} \frac{\Omega(E(x_i)) \exp(-E(x_i)/T_i)}{Z_i}. \tag{1}$$

During a simulation, all $P_{T_i}(E)$ distributions may be estimated in the form of histograms. From these histograms, a density of states $\Omega(E)$ may be estimated by means of a multihistogram reweighing technique.

3. Results

We propose a new scheme for optimal temperature selection, which can be summarized as follows:

- An initial set of temperatures is used for a PT run in which histograms for $P_{T_i}(E)$ in each temperature T_i are collected.
- The obtained statistics $P_{T_i}(E)$ are used to estimate the density of states $\Omega(E)$. From the density of states, the distribution of states can be retrieved as a two-dimensional function of energy and temperature—at this stage, temperature is considered to be a continuous variable (see figure 3(a)).
- From the estimation of the distribution of states, a set of temperatures is derived to keep an overlap ratio $r_{\text{ove}}(T_i, T_j)$ between two adjacent temperatures T_i and T_j constant (see figures 3(b) and 3(c)). Obviously, the number of temperatures in the set depends greatly on the r_{ove} value—the higher the value of r_{ove}, the greater the number of temperatures (see figure 3(d)).
- The whole procedure is iterated in order to achieve convergence of the density of states.

In this work we use C_v as a convergence criterion (see figure 4). Figure 3(a) presents the $P_T(E)$ dependence as a two-dimensional histogram with probabilities depicted on a colour scale. Using the $P_T(E)$ function, one can compute an overlap ratio $r_{\text{ove}}(T_i, T_j)$ as a function of any two temperatures. This function is plotted in figure 3(c). Now it is possible to select a set of $\mathcal{K}$ temperatures, $\mathcal{T}_\mathcal{K}$, keeping a $r_{\text{ove}}(T_i, T_{i+1})$ value constant for every $i = 1, 2, 3, \ldots, \mathcal{K}-1$. Obviously, temperatures selected in the region of a phase transition lie closer to each other than in other regions. The resulting set of temperatures is used for PT simulation in the next iteration.

3.1. Convergence of iterative temperature selection for the 2gb1 reduced model

In order to check whether the convergence of the process is achieved, virtually any thermodynamic property can be used. In the present work we employed heat capacity C_v, computed from the density of states. Figure 4 presents C_v computed in three subsequent iterations for the first model (56 residue 2gb1 domain). For these calculations we used five replicas covering the temperature range [0.3, 0.7]. We used 400 000 MC swaps per replica in each iteration of our procedure. The C_v curves (shown in figure 4) from particular iterations are compared with the 'true' C_v curve. To compute this reference curve we spent much more computer time (20 times more) than in each of the 'regular' runs.

Initially, the temperatures were spaced uniformly: 0.3, 0.4, 0.5, 0.6 and 0.7. After a PT simulation, we computed a new temperature set, starting from 0.3 for different trial values of r_{ove}. As a result of a simulation, we selected the set of temperatures that consist of five values, and the highest temperature is the closest to 0.7. The exact value of r_{ove} for which these criteria are fulfilled was 0.23. In the third iteration the resulting temperatures were 0.3, 0.363, 0.437, 0.513 and 0.685. Clearly, the system rapidly reached its ground state and the full thermodynamics of the folding transition has been characterized.

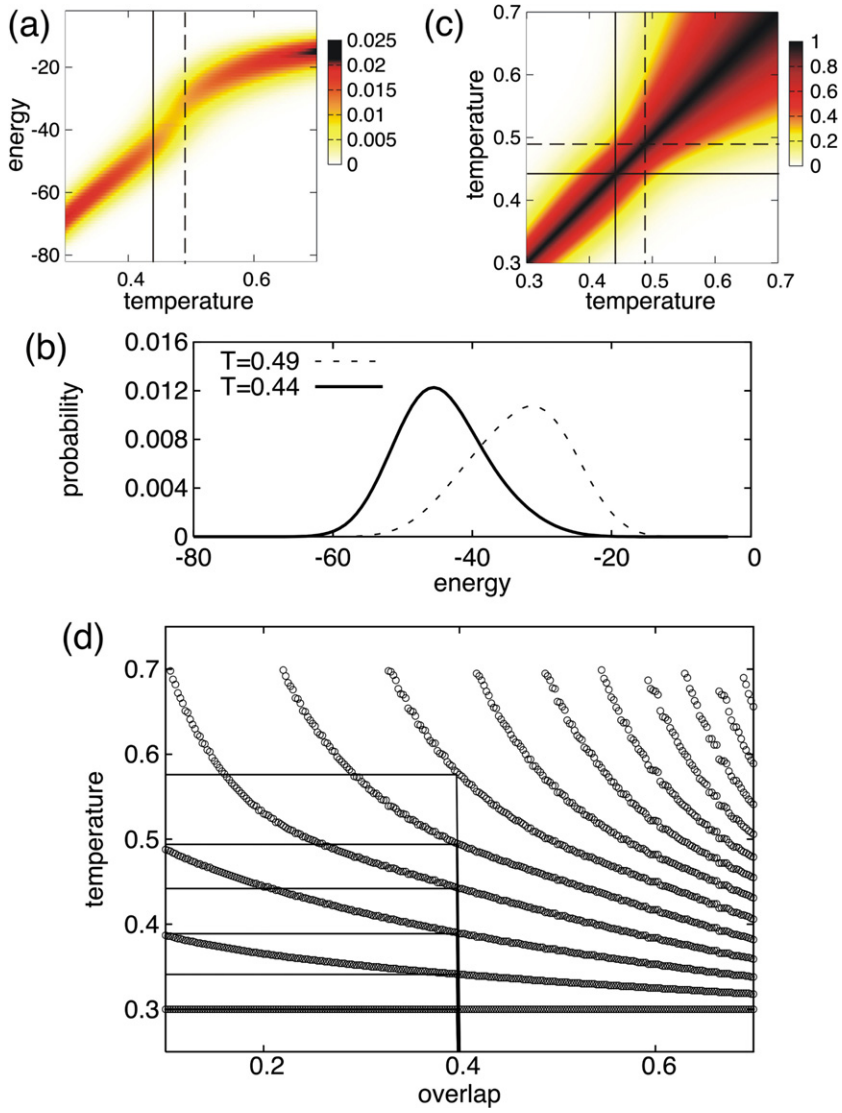

Figure 3. Schematic overview of the algorithm for optimal selection of replica temperatures. (a) From a preliminary PT run, a $P_T(E)$ probability distribution is estimated. The probability of observing the state of a given energy (horizontal axis) at a certain temperature (vertical axis) is denoted on a colour scale. The two example temperatures, 0.44 and 0.49, are marked with a solid line and a dashed line, respectively. The first one lies below and the other one above the transition temperature. (b) Density of states computed for the two temperatures marked in the plot (a). As temperature approaches the transition temperature, the two distributions become clearly non-Gaussian. The overlap for the two distribution is 0.4. (c) Two-dimensional histogram shows overlap (in the range of [0.0, 1.0]) for two distributions of states computed for any two different temperatures (both in the range [0.3, 0.7]). The two temperatures shown in the plots (a) and (b) are also marked in this plot. The intersection of the solid and dashed lines denotes an overlap for temperatures 0.44 and 0.49 of 40% (0.4). One can start from a certain temperature and walk through the plot (c), keeping a path of constant colour, i.e. constant overlap ratio. This way, a set of temperatures can be selected. (d) Optimal temperature sets computed as a function of overlap ratio. For example, for an overlap of 0.4, the following temperatures can be selected: 0.3, 0.34, 0.39, 0.44, 0.49 and 0.58.

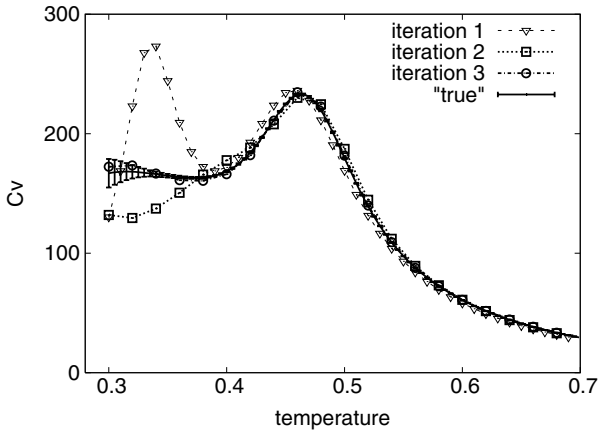

Figure 4. The convergence of our algorithm quantified by the heat capacity (C_v) as a function of temperature for the CABS model of 2gb1 protein.

Table 1. The convergence of temperatures in subsequent temperatures.

Replica no.	1	2	3	4	5	6	7	r_{ove}
Iteration 1	1.40	1.57	1.73	1.90	2.07	2.23	2.40	—
Iteration 2	1.40	1.63	1.70	1.79	1.91	2.08	2.39	0.650
Iteration 3	1.40	1.56	1.64	1.75	1.88	2.06	2.38	0.625
Iteration 4	1.40	1.59	1.66	1.76	1.89	2.07	2.38	0.635

3.2. Thermodynamics of a β-hairpin folding

For this case, besides the novel combination of multihistogram with PT sampling described in the previous sections, we also improve the efficiency of the thermodynamics calculations by applying replica multiplexing, first introduced by Rhee and Pande [26] and applied in large-scale simulations by Czaplewski *et al* [27]. To enhance sampling, the replicas are multiplexed with a number of independent simulations running at each temperature. This way, one can obtain several uncorrelated trajectories for each temperature and efficiently perform a simulation on multiple processors.

In order to make our test more difficult, we used only seven replicas. A relatively short simulation (100 000 MC swaps) with 12 copies for each temperature (i.e. the total number of replicas was $12 \times 7 = 84$) has been executed on 12 processors. In the multihistogram step of this procedure, we treat each replica independently, thus introducing 84 different histograms of energy observations. Moreover, the autocorrelation time that is used in the multihistogram procedure for error estimation [28] was computed separately for each replica. The convergence of temperatures is shown in table 1. We started from temperatures distributed uniformly between 1.4 and 2.4 (in dimensionless reduced units of energy) and kept the two flanking temperatures constant. The value of the overlap ratio was slightly different in the three iterations and close to 0.65. During the process of optimization of the temperature set, the five temperatures in the middle of the range started moving toward the transition temperature, which is equal to 1.9 for this model. Figure 5 shows the probability of energy distribution $P_T(E)$ computed for the whole range of temperatures. The figure reveals the two-state behaviour of the GB1P hairpin, consistent with the experimental findings and previous computational studies.

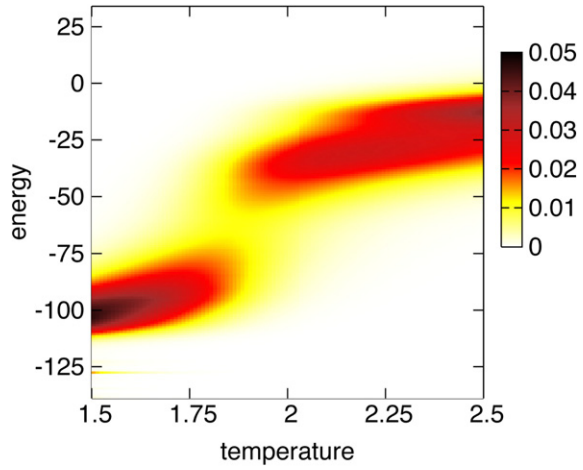

Figure 5. Probability distribution of energy for various temperatures computed from β-hairpin simulations. For the sake of clarity, the distribution has been normalized separately for each temperature. The plot illustrates the all-or-none cooperative transition at $T = 1.9$.

3.3. Comparison with other approaches

The simplest way to obtain a set of temperatures for PT simulations is to spread them equally within a given temperature range ($T_{i+1} - T_i = \text{const}$) or to use a geometric progression, i.e. $T_{i+1}/T_i = \text{const}$. In a more elaborate approach proposed by Rathore and co-workers [15], the energy distribution $P_T(E)$ is approximated by the normal distribution. The two parameters of the distribution (mean and standard deviation) are estimated from preliminary simulations. Similarly to the approach presented in this work, Rathore and co-workers proposed selecting temperatures in such a way that the value of the overlap ratio should be constant for any pair of two adjacent temperatures. In their method, the overlap ratio between two normal distributions is obtained analytically.

In this work we compare the four distinct schemes for selection of the temperatures: linear set, geometric progression, the Rathore *et al* method, and our approach. We conduct the assessment on GB1P peptide in the CABS representation. To evaluate the accuracy, we compute the $C_v(T)$ curve, spending the same amount of CPU time for each of the four methods. The results are shown in figure 6. To compute the 'true' C_v curve (solid line) we use 100 times more MC steps, as was done for the 'regular' runs. Clearly, the C_v curve computed by our approach stays in good agreement with the 'true' curve. The method proposed by Rathore *et al* is the second-best method. The good performance of the method proposed in this work is easy to understand. Figure 7 presents the density of states computed for three different values of temperature: 1.7, 1.9 (the transition temperature) and 2.1. Clearly, the distributions are far from the Gaussian distribution.

4. Conclusions and caveats

In this work we proposed a very efficient iterative procedure for the optimal allocation of temperatures for the PT sampling scheme. In contrast to other methods, we compute the optimal temperature set from estimations of the density of states for the sampled systems. The resulting sets of temperatures were employed in the next iterations for improving the thermodynamic description of the systems under consideration. It was also demonstrated that

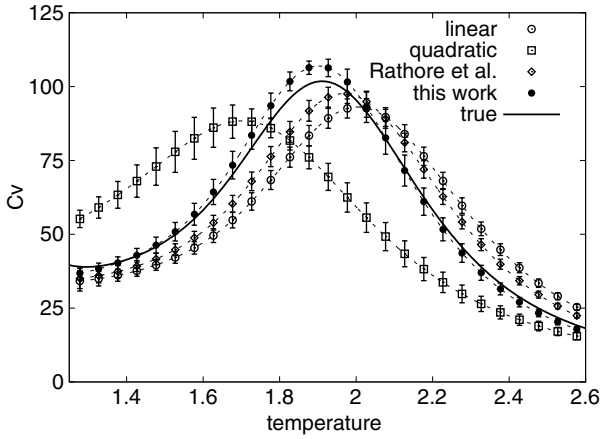

Figure 6. Heat capacity for GB1P system computed in four experiments. The experiments differ in the set of replica temperatures used in simulation.

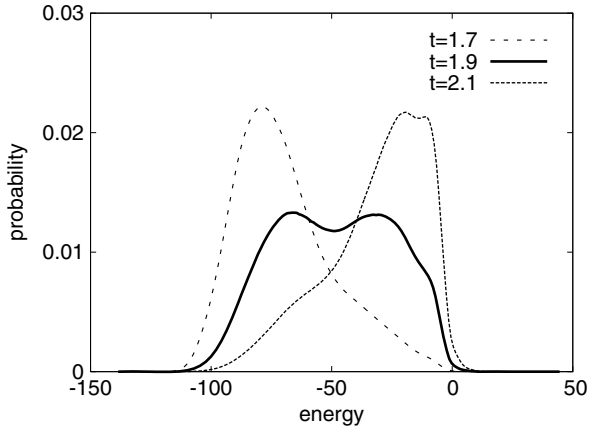

Figure 7. Distribution of states $P_T(E)$ calculated for three different values of temperature.

one can combine a multiplexing scheme with our approach in a straightforward fashion. In such cases, the autocorrelation time is computed separately for each replica, even if they feature the same temperature. For some systems the time spent on equilibration can be very long. In this case, our method becomes computationally expensive because, once a new iteration is started, one has to wait until a Boltzmann distribution is achieved. Indeed, introducing non-Boltzmann histograms (i.e. gathered without equilibration) can cause severe errors and make our procedure unstable. Fortunately, this problem can be easily avoided if system configurations are stored during simulations along with their energies. In this case, one can use $\Omega(E(x_i))$ obtained in the last iteration to pick randomly the starting configurations for the next iteration, according to Boltzmann weights computed for the new temperature.

In summary, an efficient and straightforward strategy for the simulation of macromolecular systems is provided and evaluated. The computational test for two different reduced models of polypeptides shows generality of the proposed method, which can be used for the effective search for the lowest-energy structures (the prediction of protein native-like structures) and for detailed studies of folding thermodynamics.

Acknowledgments

This work was supported by the Polish Ministry of Scientific Research and Information Technology (PBZ-KBN-088/P04/2003). The computational part of this work was performed using the computer cluster at the Computing Center of the Department of Chemistry, University of Warsaw, Poland.

References

[1] Berg B A and Neuhaus T 1992 *Phys. Rev. Lett.* **68** 9
[2] Lee J 1993 *Phys. Rev. Lett.* **71** 211
[3] Kimura K and Taki K 1991 *IMACS-91* vol 13, pp 827–8
[4] Geyer C J 1991 *Computing Science and Statistics: Proc. 23rd Symp. on the Interface Interface Foundation (Fairfax Station)* pp 156–63
[5] Swendsen R H and Wang J S 1986 *Phys. Rev. Lett.* **57** 2607
[6] Hansmann U H E 1997 *Chem. Phys. Lett.* **281** 140
[7] Gront D, Kolinski A and Skolnick J 2000 *J. Chem. Phys.* **113** 5065
[8] Sugita Y and Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141
[9] Sugita Y and Okamoto Y 2000 *Chem. Phys. Lett.* **329** 261
[10] Ferrenberg A M and Swendsen R H 1989 *Phys. Rev. Lett.* **63** 1195
[11] Ferrenberg A M and Swendsen R H 1988 *Phys. Rev. Lett.* **61** 2635
[12] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 *J. Comput. Chem.* **13** 1011
[13] Kofke D A 2002 *J. Chem. Phys.* **117** 6911
[14] Predescu C, Predescu M and Ciobanu C V 2005 *J. Phys. Chem.* B **109** 4189
[15] Rathore N, Chopra M and de Pablo J J 2005 *J. Chem. Phys.* **122** 024111
[16] Katzgraber H G, Trebst S, Huse D A and Troyer M 2006 *J. Stat. Mech.* P03018
[17] Trebst S, Troyer M and Hansmann U H E 2006 *J. Chem. Phys.* **124** 174903
[18] Taketomi H, Ueda Y and Go N 1975 *Int. J. Pept. Protein Res.* **7** 445
[19] Kolinski A and Skolnick J 1994 *Proteins* **18** 338
[20] Sikorski A and Romiszowski P 2003 *Biopolymers* **69** 391
[21] Kolinski A 2004 *Acta. Biochim. Pol.* **51** 349
[22] Blanco F J, Rivas G and Serrano L 1994 *Nat. Struct. Mol. Biol.* **1** 584
[23] Munoz V, Thompson P A, Hofrichter J and Eaton W A 1997 *Nature* **390** 196
[24] Kolinski A, Ilkowski B and Skolnick J 1999 *Biophys. J.* **77** 2942
[25] Hughes R M and Waters M L 2006 *Curr. Opin. Struct. Biol.* **16** 514
[26] Rhee Y M and Pande V S 2003 *Biophys. J.* **84** 775
[27] Czaplewski C, Kalinowski C, Oldziej S, Liwo A and Scheraga H A 2006 *From Computational Biophysics to Systems Biology* (*NIC series, Jülich*) pp 63–6
[28] Ferrenberg A M, Landau D P and Swendsen R H 1995 *Phys. Rev.* E **51** 5092
[29] Kolinski A and Bujnicki J M 2005 *Proteins* **61** 84
[30] Kmiecik S, Kurcinski M, Rutkowska A, Gront D and Kolinski A 2006 *Acta Biochim. Pol.* **53** 131
[31] Irback A, Samuelsson B, Sjunnesson F and Wallin S 2003 *Biophys. J.* **85** 1466
[32] Kurcinski M and Kolinski A 2006 *J. Steroid Biochem. Mol. Biol.* at press