Multi-Self-Overlap Ensemble for Protein Folding: Ground State Search and Thermodynamics

George Chikenji,[1] Macoto Kikuchi,[1] and Yukito Iba[2]
[1]*Department of Physics, Osaka University, Toyonaka 560-0043, Japan*
[2]*The Institute of Statistical Mathematics, Minatoku, Tokyo 106-8569, Japan*
(Received 23 February 1999)

Long chains of the *HP* lattice protein model are studied by the multi-self-overlap ensemble Monte Carlo method, which was developed recently by Iba, Chikenji, and Kikuchi. This method successfully finds the lowest energy states reported before for sequences of the chain length $N = 42-100$ in two and three dimensions. Moreover, the method realizes the lowest energy state that was ever found in a case of $N = 100$. Finite-temperature properties of these sequences are also investigated by this method. Two successive transitions are observed between the native and random coil states. Thermodynamic analysis suggests that the ground state degeneracy is relevant to the order of the transitions.

PACS numbers: 87.15.By, 02.70.Lq, 87.15.Cc, 87.10.+e

Protein folding [1] is one of the most interesting problems in biological science. To understand folding mechanism, it is useful to study simplified models. Among them, lattice protein models [2] have been playing important roles in theoretical studies of protein folding. For relatively short chains, exact enumeration of all the conformations are possible. In dealing with longer chains, however, we encounter difficulty, because dynamical Monte Carlo methods are not always efficient for long lattice proteins [3]. Extended ensemble Monte Carlo methods, such as the multicanonical ensemble [4,5], the simulated tempering [6], and the exchange Monte Carlo [7], have been applied also to simple lattice polymers [8]. These methods are, however, still not powerful enough for treating long lattice proteins, because of low degeneracy of the ground states (a necessary feature for realistic protein models) as well as rough energy landscapes caused by heterogeneity of the interactions and self-avoiding conditions. Then, what makes the Monte Carlo scheme more efficient? A key to the solution of the above question is "relaxing the self-avoidingness." In fact, a work by Shakhnovich *et al.* [9] suggested that faster dynamics is attained by relaxing the self-avoidingness. Inspired by this observation, recently we proposed a new Monte Carlo method called *multi-self-overlap ensemble (MSOE)* [10] by generalizing the idea of the multicanonical ensemble. In this method, the self-avoiding condition is systematically weakened. While the resultant ensemble contains some portions of nonphysical self-overlapping configurations, the correct canonical ensemble of the physical conformations can be reconstructed through the histogram reweighting procedure.

In a previous paper [10], we described the idea behind the algorithm and made some test calculations; we observed a considerably fast relaxation for a simple lattice heteropolymer compared with the conventional multicanonical ensemble method. In this Letter, we present the challenge to find low energy states of long chains ($42 \leq N \leq 100$) of the *HP* lattice protein model [11] in two and three dimensions using MSOE, and demonstrate that the low energy states are found successfully. We also investigate thermodynamic properties of protein-like sequences by MSOE. Two successive transitions between the native and random coil states are observed, and we study their nature in detail. These results are compared with those of poorly designed sequences. It should be noted that the principle of MSOE is independent of specific choices of the interactions between monomers. Although we restrict ourselves to the *HP* model in the following, extensions to models with other interactions, such as the one with Miyazawa-Jernigan contact matrix [12], are straightforward.

Let us explain the MSOE algorithm briefly. First, we define the degree of violation of self-avoidance V as $V = \sum_{i \in G}(n_i - 1)^2$, where n_i is the number of monomers on a lattice point i, and G denotes a set of lattice points which are occupied by at least one monomer. For self-avoiding conformations, $V = 0$. We assume that the definition of the original energy function E can be extended to conformations with self-overlaps in a reasonable manner. Then, we determine the weight factor $P_g \propto \exp[-g(E, V)]$ through preliminary runs as in the case of the conventional multicanonical ensemble [4], so that the bivariate histogram of (E, V) is sufficiently flat in a prescribed range $E_{\min} \leq E \leq E_{\max}$, $0 \leq V \leq V_{\max}$. We set $E_{\min}$ to a value which is definitely lower than the ground state energy, and set $V_{\max}$ to a value $\sim N/10-N/5$. We actually use the entropic sampling method [5] in this paper. After the weight factor $g(E, V)$ is determined, a measurement run is performed. The canonical averages $\langle A \rangle$ of an observable $A(\Gamma)$ at temperature T is calculated according to the histogram reweighting formula,

$$\langle A \rangle_T = \frac{\sum_i' A(\Gamma_i) P_g^{-1}(\Gamma_i) \exp[-E(\Gamma_i)/T]}{\sum_i' P_g^{-1}(\Gamma_i) \exp[-E(\Gamma_i)/T]}, \qquad (1)$$

where Γ_i represents a conformation at the ith Monte Carlo step and the summations are taken only over the self-avoiding conformations.

 0031-9007/99/83(9)/1886(4)$15.00 © 1999 The American Physical Society

How does MSOE works? In Fig. 1, observed transitions during a MSOE simulation are plotted on the (E, V) plane. (This example is taken from a simulation for the 2D $N = 64$ sequence shown in Table I, which we will discuss below.) We can clearly see *bridges* (or *paths*) between self-avoiding low energy states where self-overlapping states are used as *stepping stones*. That is, while no direct transition is seen between $(E, 0)$ and $(E', 0)$, there are paths that utilize the states with nonzero V as intermediate states. The existence of such paths are a key feature of MSOE, which effectively facilitates the relaxation. The idea behind MSOE can also be extended to off-lattice models. For example, one can consider the hard core repulsive part of the energy as an off-lattice counterpart of the self-avoiding conditions. Other types of bivariate extension of the multicanonical algorithm for off-lattice protein models have also been discussed by Higo *et al.* [13]. They, however, did not incorporate unphysical conformations for the purpose of attaining fast relaxations. MSOE also shares some ideas in common with the *ghost* polymer procedure by Wilding and Müller for dense polymer systems [14].

Let us discuss the results of simulations for the *HP* protein model [11], in which a protein consists of a self-avoiding chain on a lattice with two types of amino acids: *H* (hydrophobic) and *P* (polar). The energy E of a chain conformation is determined only by the number of *H*-*H* contacts h as $E = -|\epsilon|h$, where $|\epsilon|$ is a positive constant (we measure the energy in the unit of $|\epsilon|$ hereafter). In MSOE, we use the same definition of energy as the original one also for self-overlapping conformations. In principle, MSOE can be implemented with arbitrary dynamics. We use the following elementary moves in this paper: (1) jackknife move, (2) one bead flip, and (3) pivot operator. Frequency of the pivot operator trial is taken as $1/5$ of that of the other two moves. A description of the move (1) is given in Ref. [10] and the moves (2) and (3) are illustrated in Ref. [15]. All CPU times quoted below refer to PCs with 500 MHz DEC 21164A chip.

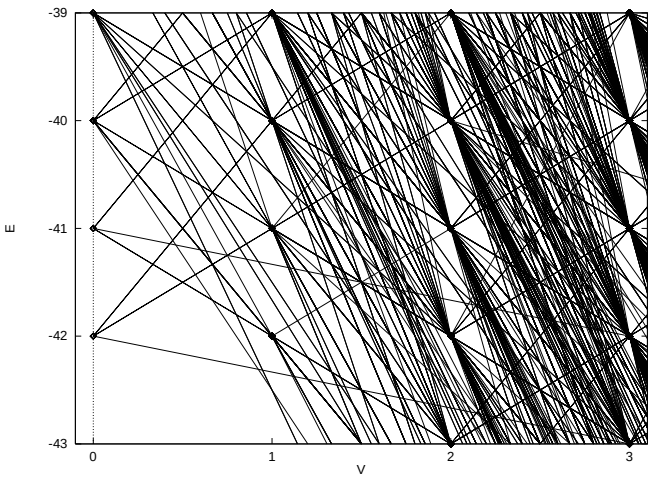

FIG. 1. Observed transitions on the (E, V) plane during a MSOE run of the $N = 64$ *HP* sequence in Table I. Solid lines connecting two states represent the transitions between them. Only a small part of the entire (E, V) plane near the ground state is shown.

First we consider a two dimensional *HP* lattice protein with the chain length $N = 64$ in Table I, which has been treated by several methods so far [16–19]. For this sequence, the ground state energy is believed to be $E = -42$. One of the ground state conformations is found in Fig. 5 of Ref. [19]; it has a highly regular shape where the *H* core is surrounded by *P* monomers forming a helixlike secondary structure. All other ground state conformations differ only in structures within *H* core. We calculated the weight factor within 30 hours with $E_{\min} = -60$ and $V_{\max} = 10$. Resultant distribution of (E, V) is sufficiently flat including the state $(-42, 0)$, and we can calculate thermodynamic quantities of the sequence at *any temperature* using it. The values of the lowest energy found by several methods are listed in Table I. *Core directed chain growth (CG) method* [18] has given the ground states. It is, however, specialized in the search of the ground state, and thus cannot be used for thermodynamics. Bastolla *et al.* also have found these states ($E = -42$) by the *pruned-enriched Rosenbluth method* [19] with an assumption that the ground states do not contain any nonbonded *HP* neighbor pair (the lowest energy obtained without this assumption was $E = -40$). Such an additional assumption will cause uncontrollable biases in calculations of finite-temperature properties. Moreover, as we will see later, there are examples in which low energy states do not satisfy this assumption. Then, as far as we know, MSOE is currently the only method which can calculate finite-temperature properties as well as the low energy states of this sequence. We calculated the specific heat and the gyration radius. We found a peak at $T \sim 0.55$ and a shoulder at $T \sim 0.4$ in the specific heat curve. The shoulder at $T = 0.4$ corresponds to the transition between the ground states and compact globule states accompanied by a rapid decrease of entropy. This transition is first-order-like since a bimodal energy distribution is observed. The peak at $T = 0.55$, on the other hand, is the transition between compact globule states and the random coil. In this case the energy distribution is unimodal; that is, the transition is second-order-like.

Consider another sequence of $N = 100$ shown in the second row of Table I. Bastolla *et al.* [19] have found conformations with $E = -49$ ($E = -48$ without the above-mentioned assumption on the ground state conformations). On the other hand, we found conformations with $E = -50$ within 50 hours by MSOE taking $E_{\min} = -65$ and $V_{\max} = 10$. Figure 2 shows a typical conformation with $E = -50$. This conformation contains some nonbonded *HP* neighbor pairs. Thus, it can never be found when these pairs are not allowed. Unfortunately,

TABLE I. The values of the lowest energy reported by several authors for three sequences of 2D *HP* model.

N	Sequence	$E_{\min}$	Ref.
64	$H_{12}PHPHP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HPHPH_{12}$	−37	[16]
		−40	[17]
		−42	[18]
		−42	[19]
		−42	present study
100	$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$	−46	[25]
		−49	[19]
		−50	present study
100	$P_6HPH_2P_5H_3PH_5PH_2P_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$	−44	[25]
		−47	[19]
		−47	present study

we have not yet obtained the weight factor that gives a sufficiently flat distribution of (E, V) for calculating thermodynamic properties.

Next, we discuss a three dimensional example. A sequence of $N = 42$ given by Yue and Dill [20] has the ground states with degeneracy of only 4-fold. Moreover, its ground state conformations resemble the parallel β-helix structure found for a real protein *pectate lyase C* [21]. The ground state energy has been calculated exactly as $E = -34$ by the constrained hydrophobic core construction method [22]. As far as we know, no attempts have been reported so far to calculate thermodynamic properties of this sequence. We applied MSOE to this sequence. By taking $E_{\min} = -50$ and $V_{\max} = 8$, we successfully obtained the appropriate weight factor within 50 hours. Temperature dependence of the specific heat, the entropy, and the gyration radius are shown in Fig. 3. Two peaks are seen in the specific heat curve. The peak at $T \sim 0.28$ corresponds to the transition between the ground states and compact globule states (first-order-like). This transition is accompanied by a rapid decrease of entropy. Another peak (at $T \sim 0.52$) corresponds to one between the compact globule states and the random coil (second-order-like). Again, energy distributions near the transition temperature are bimodal and unimodal, respectively. Let us see the transition at $T \sim 0.52$ carefully. The gyration radius is considerably small already at $T \sim 0.52$. We also found (although not shown in the figure) that the fluctuation of the gyration radius has a peak at temperature significantly higher than $T \sim 0.52$. These observations imply that the coil-globule transition takes place in two steps.

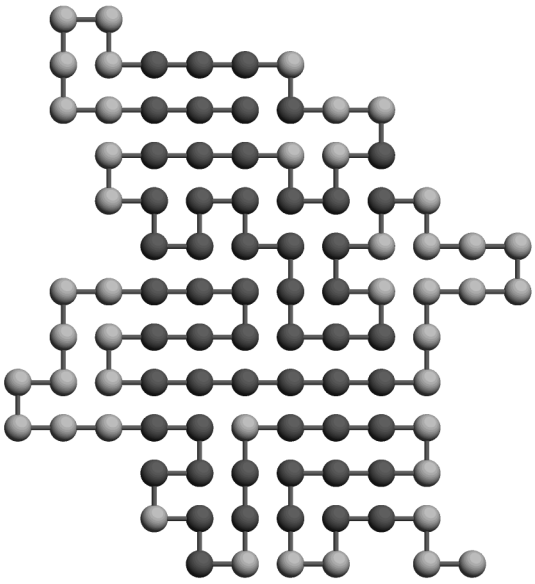

FIG. 2. A typical conformation with $E = -50$ of the second sequence in Table I. This value of the energy is reached for the first time by MSOE.

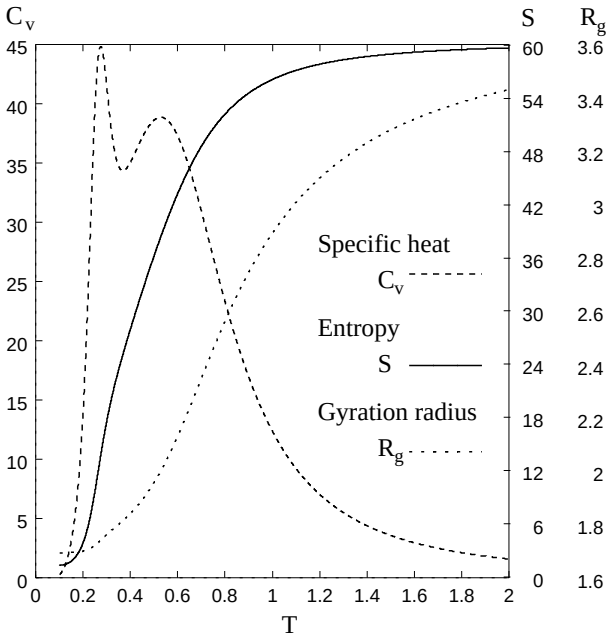

FIG. 3. Temperature dependence of the specific heat, the entropy, and the gyration radius for the β-helix sequence.

We made MSOE simulations for two more *HP* sequences: another sequence of $N = 100$ on the square lattice given in the third row of Table I and the $N = 48$ sequence on the cubic lattice given as the sequence No. 9 in Ref. [3]. For both sequences, we easily obtained appropriate weight factors down to the lowest energy states ($E = -47$ and -34, respectively), and found that no first-order-like transition takes place. Thus, not all the *HP* sequences exhibit first-order-like transition. The lowest energy states of these sequences are highly degenerated. Thus they are more like random heteropolymers rather than proteins. This result is consistent with the observation by Pande *et al.* that poorly designed sequences with highly degenerate ground states have less sharp transitions [23]. On the other hand, both the 2D $N = 64$ sequence and the β-helix sequence, which have low degeneracy of the ground states, exhibit first-order-like native-globule transitions. The above results imply that low degeneracy of the ground states is required for first-order-like native-globule transitions to take place in long *HP* chains. These two sequences are well-designed (or proteinlike) from the viewpoint of the native structure and the degeneracy of the native states as well as the nature of the transitions to the native states. It, however, has not been explored whether they are fast folders. Thermodynamic properties of fast folding *long* lattice proteins are important subjects left for future studies.

We also tested the conventional multicanonical algorithm for all the sequences studied above, but none of the lowest energy states found by MSOE could be reached within 4–5 days. Thus, for such long chains as treated in the present study, the conventional multicanonical algorithm is of no practical use.

In summary, we applied the multi-self-overlap ensemble method for long chains of the *HP* lattice protein model in two and three dimensions. For all five sequences we treated, MSOE successfully reached the lowest energy states reported so far. Especially for the second sequence in Table I, we found the lowest energy states which any other methods have failed to find. We thus confirmed that MSOE is a powerful tool for searching the ground states of lattice protein models. Moreover, we explored the native-globule transitions at finite temperature by MSOE, and found that the degeneracy of the ground states is relevant for the order of the transition. MSOE is a quite general method so that it can be applied to any types of interactions. It can calculate correct thermodynamic properties within a moderate CPU time (although not fastest in some cases actually) as well as low energy states. Especially, once the proper weight factors are determined, we can calculate thermodynamic properties in *an arbitrary wide range of temperature* in a single run of MSOE.

Note added.—After submitting this paper, we learned that Irbäck also treated the $N = 64$ sequence in Table I by a long run of the simulated tempering method [25]. A few independent visits to the ground states were recorded and no thermodynamic properties were discussed there. We thank A. Irbäck for sending us the reprint.

[1] See, e.g., *Protein Folding,* edited by T. E. Creighton (Freeman, New York, 1992).

[2] H. Taketomi *et al.,* Int. J. Pept. Protein Res. **7**, 445 (1975); J. D. Bryngelson *et al.,* Proteins **21**, 167 (1995); K. A. Dill *et al.,* Protein Sci. **4**, 561 (1995); D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996); E. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997); V. S. Pande *et al.,* Curr. Opin. Struct. Biol. **8**, 68 (1998).

[3] K. Yue *et al.,* Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).

[4] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).

[5] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).

[6] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[7] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996); C. J. Geyer, *Computing Science and Statistics,* edited by E. M. Keramides (Interface Foundation, Fairfax Station, VA, 1991), p. 156.

[8] M. C. Tesi *et al.,* J. Stat. Phys. **82**, 155 (1996); N. Urakami and M. Takasu, J. Phys. Soc. Jpn. **65**, 2694 (1996); T. Vrobova and S. G. Whittington, J. Phys. A **31**, 3989 (1998).

[9] E. Shakhnovich *et al.,* Phys. Rev. Lett. **67**, 1665 (1991); J. Mol. Biol **235**, 1614 (1994).

[10] Y. Iba, G. Chikenji, and M. Kikuchi, J. Phys. Soc. Jpn. **67**, 3327 (1998).

[11] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[12] S. Miyazawa and R. J. Jernigan, Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996).

[13] J. Higo *et al.,* J. Comput. Chem. **18**, 2086 (1997).

[14] N. B. Wilding and M. Müller, J. Chem. Phys. **101**, 4324 (1994).

[15] A. D. Sokal, in *Monte Carlo and Molecular Dynamics Simulation in Polymer Science,* edited by K. Binder (Oxford Unversity Press, New York, Oxford, 1995), p. 47.

[16] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).

[17] L. Toma and S. Toma, Protein Sci. **5**, 147 (1996).

[18] T. C. Beutler and K. A. Dill, Protein Sci. **5**, 2037 (1996).

[19] U. Bastolla *et al.,* Proteins **32**, 52 (1998).

[20] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 146 (1995).

[21] M. D. Yonder, N. T. Keen, and F. Jurnak, Science **260**, 1503 (1993).

[22] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993).

[23] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **107**, 5118 (1997).

[24] R. Ramakrishnan, B. Ramachandran, and J. F. Pekney, J. Chem. Phys. **106**, 2418 (1997).

[25] A. Irbäck, in *Monte Carlo Approach to Biopolymers and Protein Folding,* edited by P. Grassberger *et al.* (World Scientific, Singapore, 1998), p. 98.