Wang-Landau sampling in face-centered-cubic hydrophobic-hydrophilic lattice model proteins

Jingfa Liu,[1,2,3] Beibei Song,[1,2] Yonglei Yao,[1,2] Yu Xue,[1,2] Wenjie Liu,[1,2] and Zhaoxia Liu[3]

[1]*Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044, China*

[2]*School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China*

[3]*Network Information Center, Nanjing University of Information Science & Technology, Nanjing 210044, China*

(Received 8 May 2014; published 15 October 2014)

Finding the global minimum-energy structure is one of the main problems of protein structure prediction. The face-centered-cubic (fcc) hydrophobic-hydrophilic (HP) lattice model can reach high approximation ratios of real protein structures, so the fcc lattice model is a good choice to predict the protein structures. The lacking of an effective global optimization method is the key obstacle in solving this problem. The Wang-Landau sampling method is especially useful for complex systems with a rough energy landscape and has been successfully applied to solving many optimization problems. We apply the improved Wang-Landau (IWL) sampling method, which incorporates the generation of an initial conformation based on the greedy strategy and the neighborhood strategy based on pull moves into the Wang-Landau sampling method to predict the protein structures on the fcc HP lattice model. Unlike conventional Monte Carlo simulations that generate a probability distribution at a given temperature, the Wang-Landau sampling method can estimate the density of states accurately via a random walk, which produces a flat histogram in energy space. We test 12 general benchmark instances on both two-dimensional and three-dimensional (3D) fcc HP lattice models. The lowest energies by the IWL sampling method are as good as or better than those of other methods in the literature for all instances. We then test five sets of larger-scale instances, denoted by the S, R, F90, F180, and CASP target instances on the 3D fcc HP lattice model. The numerical results show that our algorithm performs better than the other five methods in the literature on both the lowest energies and the average lowest energies in all runs. The IWL sampling method turns out to be a powerful tool to study the structure prediction of the fcc HP lattice model proteins.

DOI: 10.1103/PhysRevE.90.042715 PACS number(s): 87.15.Cc, 87.15.ak, 05.10.Ln

I. INTRODUCTION

To predict the structures of the proteins, mostly coarse-grained simplified models [1–3] have been used by researchers. However, even for the simplest hydrophobic-hydrophilic (HP) lattice model [1,2], where two types of amino acids, hydrophobic and hydrophilic (or polar), are considered, the protein structure prediction (PSP) has been proven to be NP-complete [4]. The greatest difficulty lies in the huge search space, as well as the complexity of the energy surface, which contains many local minima and a few global minima. In this paper we focus on the face-centered-cubic (fcc) HP lattice model [3], which can yield very good approximations of real protein structures [5].

The deterministic approaches are not helpful in identifying minimum-energy conformations, so finding the nondeterministic heuristic approaches that can extract minimum-energy conformations efficiently is of great importance. Some outstanding heuristic approaches, such as the simple genetic algorithm (SGA) [6]; the hybrid genetic algorithm (HGA) [6], which combines generalized short pull moves and improved crossover and mutation operations; the hybrid genetic algorithm with twin removal (HGATR) [7–9]; the genetic algorithm with elite-based reproduction strategy (ERSGA) [10]; the hybrid of hill climbing and genetic algorithm (HHGA) [10] based on ERSGA; the tabu search (TS) [8]; the evolutionary algorithm (EA) with lattice rotation for crossover and K-site moves for mutation [11]; the tabu-based local search (LST) method [12,13]; the tabu-based spiral search (SST) algorithm [13]; the memory-based local search (LSM) method [13]; the hybrid search technique that embeds SST within an enhanced population-based genetic algorithm (SSTHGA) [14]; the memetic algorithm (MA) [15]; and the large neighborhood search [16], were applied to the fcc HP lattice model.

As a Metropolis importance sampling algorithm, the Monte Carlo simulations have also played a major role in solving the PSP. Typical examples are the multicanonical ensemble method [17], the broad histogram method [18], the flat histogram method [19], and the energy landscape paving method [20–22]. The Wang-Landau (WL) sampling method [23], which can reduce the computing time even for systems with complex energy landscapes, is a type of new Monte Carlo global optimization method that has been proved to be efficient for PSP on both two-dimensional (2D) and 3D HP lattice model proteins [24]. In this paper, an improved Wang-Landau (IWL) sampling method that incorporates the generation of initial conformation based on the greedy strategy and the neighborhood search strategy based on pull moves into the WL method is proposed for PSP on the fcc HP lattice model. Numerical results show that the IWL sampling method is an effective algorithm for solving PSP on the fcc HP lattice model.

The remainder of the paper is organized as follows. In Sec. II we briefly describe the fcc HP lattice model. In Sec. III we describe our method. The numerical results and a discussion of our implementation are shown in Sec. IV. We summarize in Sec. V.

II. THE FCC HP LATTICE MODEL

The fcc HP lattice model has been proved to be the densest sphere-packing model based on the full proof of Kepler conjecture [25]. It is a kind of HP lattice model. Before introducing the fcc HP lattice model, we first review a regular (square or simple cubic) HP lattice model.

A. The HP lattice model

The HP lattice model [1,2] is an abstraction of real proteins based on the belief that hydrophobic amino acids tend to come together and form a compact core to exclude water and interactions between hydrophobic amino acids greatly contribute to the free energy of the natural conformation of a protein. In a regular (square or simple cubic) HP lattice model, a protein is composed of only two types of amino acids, hydrophobic and hydrophilic. A protein sequence that can be regarded as a string with binary characters H and P is arranged as a self-avoiding walk (SAW) chain, where adjacent characters in the sequence occupy adjacent grid points in a regular (square or simple cubic) HP lattice model and no grid point in the lattice is occupied by more than one character. Two amino acids are topological adjacent if they are neighbors in the lattice, but are not adjacent in the sequence. A topological H-H bond is formed between two topological adjacent hydrophobic amino acids. If a conformation c has m H-H bonds, its free energy $E(c) = m(-1)$. Hence, the conformation with the lowest free energy corresponds to the conformation with the largest number of H-H bonds.

B. The fcc HP lattice model

The 2D square and 3D cubic models are the most frequently used HP lattice models. However, they have the drawback of allowing interactions only between amino acids of opposite parity in the sequence. That is to say, if two amino acids are at any even distance in the primary sequence, they cannot be neighbors in the lattices. In addition, their ability to approximate real proteins is poor. For these reasons, this paper considers the fcc HP lattice model [3], which is shown to yield very good approximations of real protein structures [5,10] and is parity problem free, which means an odd indexed amino acid in the sequence can be the neighbor of both odd and even indexed amino acids in the sequence and vice versa. This model has also been rigorously shown to be the closest packing geometry for identical spheres [4]. In fact, the 2D fcc lattice is the infinite graph $G = (V,L)$, where the vertex set $V = (\sqrt{3}\,\mathbb{Z} \times \mathbb{Z}) \cup [(\sqrt{3}\,\mathbb{Z} + \sqrt{3}/2) \times (\mathbb{Z} + 1/2)]$ and the edge set $L = \{(x,x') | x,x' \in V, \|x - x'\| = 1\}$. Here $\mathbb{Z}$ denotes the integer set and $\|x - x'\|$ denotes the Euclidean distance between x and x'. The 3D fcc grids can be described as a stack of 2D fcc grids, where every individual 2D grid is slightly offset with respect to the grids above and below it. The 2D and 3D fcc lattices are generated by the following basis vectors: (–1,0), (1,0), (1/2,$\sqrt{3}$/2), (–1/2, –$\sqrt{3}$/2), (–1/2,$\sqrt{3}$/2), (1/2,–$\sqrt{3}$/2) and (1,–1,0), (–1,1,0), (–1,–1,0), (1,1,0), (0,–1,1), (0,–1,–1), (1,0,1), (1,0,–1), (0,1,1), (–1,0,1), (0,1,–1), (–1,0,–1), respectively. Two 3D fcc points $P_i(x_i,y_i,z_i)$ and $P_j(x_j,y_j,z_j)$ are adjacent if and only if $(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 = 2$. Each grid point has 6 neighbors in the 2D fcc HP lattice model that form a hexagon (see Fig. 1) and 12 neighbors (see Fig. 2) in the 3D fcc HP lattice model [24]. With this, a protein conformation of the sequence can be placed as a SAW chain on a 2D or 3D fcc lattice. Then the energy of a given conformation is defined as the number of topological adjacent H-H bonds.

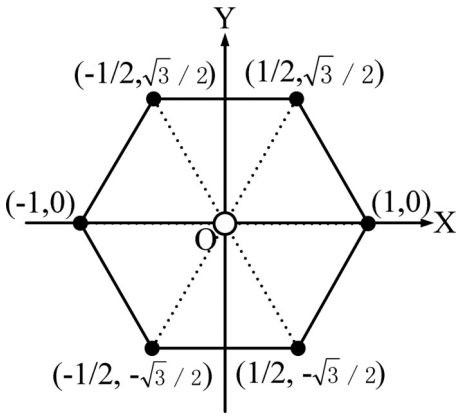

FIG. 1. Unit of the 2D fcc HP lattice model, where the six basis vectors are shown and grid point O has six neighbors.

III. METHODS

A. Wang-Landau sampling method

The WL sampling method was first introduced by Wang and Landau [23] in 2001. Unlike conventional Monte Carlo methods that directly generate a canonical distribution at a given temperature, this method estimates the density of states $g(E)$ for the range of possible energies accurately via a random walk. By using a carefully controlled modification factor, the estimate for g(E) is improved at each step of the random walk, which makes g(E) converge to the correct value very quickly. The method is based on the fact that a given energy level E is visited in energy space with a probability proportional to the reciprocal of the density of states $1/g(E)$.

At the very beginning of the Wang-Landau sampling method, $g(E)$ is unknown as is the range of possible energies. So the density of states is set self-adaptively, namely, every time the random walk finds a new energy, marked as visited, we set, respectively, its density of states and the corresponding histogram to 1. In the simulations on the fcc lattice model, we begin the random walk in energy space by randomly changing the conformations of the proteins, but the energy associated with each conformation is only accepted with a probability proportional to the reciprocal of the density of states. Therefore, the acceptance probability from energy E_1 to E_2 is as follows: $P(E_1 \to E_2) = \min(e^{\ln g(E_1) - \ln g(E_2)}, 1)$. If

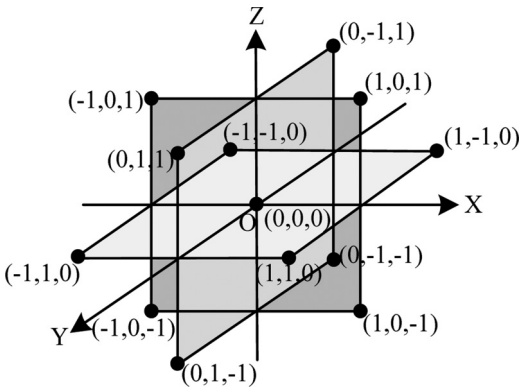

FIG. 2. Unit of the 3D fcc HP lattice model, where the 12 basis vectors are shown and grid point O has 12 neighbors.

E_2 is accepted, then $g(E_2)$ will be updated by multiplying it by a modification factor f_i, where f_0 is set equal to the value $e \approx 2.718\,28$, which Wang and Landau used in their original paper; at the same time, the histogram $H(E_2)$ will also be changed by adding 1, i.e., $\ln[g(E_2)] = \ln f_i + \ln[g(E_2)]$ $[g(E_2) = f_i g(E_2)]$ and $H(E_2) = H(E_2) + 1$. In contrast, if the state with energy E_2 is rejected, then $\ln[g(E_1)] = \ln f_i + \ln[g(E_1)]$ $[g(E_1) = f_i g(E_1)]$ and $H(E_1) = H(E_1) + 1$. Since the value of $g(E)$ will be too big to be shown as a double precision number, we choose the logarithmic formula. In the original work by Wang and Landau [23], the convergence of the WL algorithm was controlled by the flatness of the histogram. However, the flatness criterion is not strict [26–28]. Zhou and Bhatt [26] gave a proof of the convergence and analyzed the source of statistical error. Morozov and Lin [27,28] further identified estimations of the convergence and accuracy of the WL algorithm. In fact, in simulations on the fcc lattice model, the convergence of a random walk can be controlled by the criterion [28] $H(E) \geqslant \ln 2/\ln f_i$ for all visited energies E. After convergence we reduce the modification factor to a finer one using a function such as $f_{i+1} = \sqrt{f_i}$ (any function that monotonically decreases to 1 will do), reset $H(E)$ to 0 for all visited energies E, and begin the next random walk. The simulation continues until the modification factor f_i falls below a threshold (e.g., $f_{\text{final}} = 1.0001 \approx 1$) at which $g(E)$ has converged towards the true value of density of states with a statistical error $\sqrt{a \ln f_i}$, where the factor a is proportional to the local difference of the density of states [27]. Our goal is to find the conformation with the lowest energy, so in simulations we also keep the lowest energy E_{opt} and the corresponding conformation c_{opt} each time we find a new lower-energy conformation.

B. Generation of initial conformation

The Wang-Landau sampling method starts to search for low-energy conformations from a valid initial conformation. We use the greedy strategy to get the initial conformation for a given amino acid sequence with length n. The detailed steps are as follows. First, we put the first two amino acids at two adjacent fixed positions on the lattice. Subsequently, we pseudoplace the ith ($3 \leqslant i \leqslant n$) amino acid at every position that is adjacent to the (i–1)th amino acid and not occupied by other amino acids, where "pseudoplace" means that the ith amino acid is placed temporarily and will be removed after computing the energy of the partial conformation, which consists of the previous $i-1$ amino acids and the ith amino acid. If such positions exist, we formally put the ith amino acid at the position where the energy of the corresponding partial conformation is lowest; otherwise we remove the ($i-1$)th amino acid and continue to grow the partial conformation from the ($i-2$)th amino acid. Once a valid conformation with n amino acids is produced, the process is stopped.

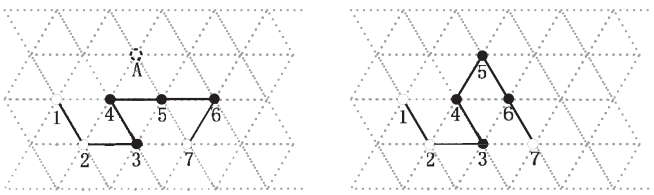

FIG. 3. Example of the single-point pull moves on the 2D fcc HP lattice model. Closed and open circles indicate the hydrophobic and hydrophilic amino acids, respectively. If position A is free, then amino acid 5 can be placed at A and a pull move in (a) can be executed, where amino acid 6 is moved to the position of amino acid 5 and then a valid conformation [indicated in (b)] is obtained.

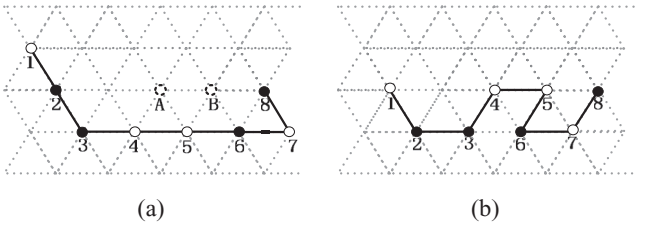

FIG. 4. Example of parallel pull moves on the 2D fcc HP lattice model. Closed and open circles indicate the hydrophobic and hydrophilic amino acids, respectively. If positions A and B are free [see (a)], then amino acids 4 and 5 can be placed at their adjacent parallel positions A and B. Then, to obtain a valid conformation, on the left side of amino acid 4, amino acid 3 is moved to the position of amino acid 4, 2 to 3, and 1 to 2 and on the right side of amino acid 5, amino acid 6 is moved to the position of amino acid 5 and 7 to 6 [indicated in (b)].

C. Pull moves

An efficient neighborhood search strategy is also impactful in the Wang-Landau sampling simulation. Here the pull move that was originally introduced by Lesh *et al.* [29] for square and cubic lattices is used to execute the neighborhood search. Different from the pivot moves [30], it allows for close-fitting motion of a polymer chain within a confining environment [29]; unlike end flips, corner flips, and crankshafts [30], it is complete and reversible [8,29], which makes it efficient to update the conformation and to guarantee the reachability of the global minimum.

According to the number of moved amino acids at the first step of the move, the pull moves can be divided into single-point pull moves and two-point pull moves. In single-point

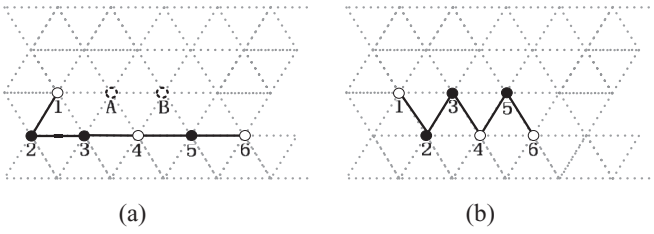

FIG. 5. Example of tilted pull moves on the 2D fcc HP lattice model. Closed and open circles indicate the hydrophobic and hydrophilic amino acids, respectively. If positions A and B are free [see (a)], then amino acid 3 can be placed at position A and 5 can be placed at position B. Then, to obtain a valid conformation, on the left side of amino acid 3, amino acid 2 is moved to the position of amino acid 3 and on the right side of amino acid 5, amino acid 6 is moved to the position of amino acid 5 [indicated in (b)].

TABLE I. Twelve instances for the fcc HP lattice model.

Instance	Length	Sequence
1	20	$HPHP_2H_2PHP_2HPH_2P_2HPH$
2	24	$H_2P_2(HP_2)_6H_2$
3	25	$P_2HP_2(H_2P_4)_3H_2$
4	36	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$
5	48	$P_2H(PH_3)_2P_5H_{10}P_6(H_2P_2)_2HP_2H_5$
6	50	$H_2(PH)_3PH_4P(HP_3)_3P(HP_3)_2HPH_4(PH)_4H$
7	54	$H_2(PH)_3PH_4P(HP_3)_4P(HP_3)_2HPH_4(PH)_4H$
8	60	$P(PH_3)_2H_5P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$
9	64	$H_{12}(PH)_2\ ((P_2H_2)_2P_2H)_3(PH)_2H_{11}$
10	85	$H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2(H_2P_2)_2HPH$
11	100_a	$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$
12	100_b	$P_6HPH_2P_5H_3PH_5PH_2P_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$

pull moves, at each move, only one amino acid can be moved and the entire sequence only moves to one direction. The main idea of the single-point pull move is as follows. We randomly choose an amino acid from the chain with length n. If there exists a "legal" position that is vacant in the grid and adjacent to both this amino acid and either its predecessor or successor in the chain, we put it at this position. If the chain has been broken, we put the successor (or predecessor) of this amino acid in its old position. Then we complete the remaining moves by the pull-move rules until a new legal conformation is reached. An example of single-point pull moves on the 2D fcc HP lattice model is shown in Fig. 3. References [29,31] give a description of the detailed steps of the single-point pull moves.

Correspondingly, in two-point pull moves, at each move, two amino acids can be moved at the same time and the entire sequence moves to two contrary directions. Two-point pull moves can be divided into parallel pull moves and tilted pull moves. In parallel pull moves, at the beginning of the move, two adjacent amino acids are moved and the directions in which they move are parallel. An example of parallel pull moves on the 2D fcc HP lattice model is shown in Fig. 4. Tilted pull moves can be performed on the condition that there exist three adjacent amino acids in the same line and there exist two free positions on the same side of the line that are adjacent to the middle amino acid. An example of tilted pull moves on the 2D fcc HP lattice model is shown in Fig. 5.

The pull move on the 3D fcc lattice is similar to that on the 2D fcc lattice. However, in a 2D fcc lattice, the chosen amino acid may at most be moved to six adjacent positions (see Fig. 1), but in a 3D fcc lattice it may be moved at most to 12 adjacent positions (see Fig. 2).

D. Description of the algorithm

By incorporating the generation of an initial conformation based on the greedy strategy and the neighborhood search strategy with pull moves into Wang-Landau sampling method, an IWL sampling method is proposed for the fcc HP lattice model. The calculating procedure is presented as follows.

(i) Generate randomly a valid initial conformation c_1 based on the greedy strategy. Compute the energy E_1 of c_1. Let $E_{\text{opt}} = E_1$ and $c_{\text{opt}} = c_1$. Let the set of visited energies $M = \{E_1\}$, the histogram function $H(E_1) = 1$, and density of states $g(E_1) = 1$. Let $i = 0$ and $f_0 = e \approx 2.718\,28$.

(ii) Let $N = \{1,2,\ldots,n\}$.

(iii) Choose randomly an integer j from N.

TABLE II. Comparison of computational results by different methods on the 2D fcc lattice model. Instances are taken from Table I. Numbers in bold indicate the lowest energies so far. Best denotes the lowest energy found in all runs for each instance. Avg. denotes the average lowest energy in all runs. SD denotes the standard error of the lowest energies in all runs.

							IELP			IWL		
Instance	SGA	HGA	HGATR	ERSGA	HHGA	TS	Best	Avg.	SD	Best	Avg.	SD
1	–11	**–15**	**–15**	**–15**	**–15**	**–15**	**–15**	–15.0	0.000	**–15**	–15.0	0.000
2	–10	–13	–13	–13	**–17**	**–17**	**–17**	–17.0	0.000	**–17**	–17.0	0.000
3	–10	–10	–10	**–12**	**–12**	**–12**	**–12**	–12.0	0.000	**–12**	–12.0	0.000
4	–16	–19	–19	–20	–23	**–24**	**–24**	–24.0	0.000	**–24**	–24.0	0.000
5			–32	–32	–41	–40	**–44**	–43.3	0.781	**–44**	–43.4	0.672
6				–30	–38		**–42**	–40.1	0.943	**–42**	–41.8	0.400
7	–21	–23	–23			–31	**–44**	–41.9	1.300	**–44**	–43.3	0.698
8	–40	–46	–46	–55	–66	–70	**–71**	–70.1	0.539	**–71**	–70.4	0.477
9	–33	–46	–46	–47	–63	–50	**–75**	–74.1	0.700	**–75**	–74.7	0.476
10							**–101**	–100.2	0.600	**–101**	–100.4	0.316
11							**–94**	–93.0	0.632	**–94**	–93.5	0.589
12							**–94**	–93.2	0.400	**–94**	–93.2	0.400

(iv) Execute single-point and two-point pull moves for all legal move positions of the jth amino acid of the current conformation c_1. If at least one pull move is executed successfully, choose randomly a legal conformation obtained by pull moves as an updated conformation of c_1, denoted by c_2, then compute the energy E_2 of c_2, let $g(E_2) = 1$ and $H(E_2) = 1$, and go to step (v); otherwise, let $N = N - \{j\}$ and go to step (iii).

(v) If $E_2 \notin M$, let $M = M + \{E_2\}$.

(vi) If random $(0,1) < \min(e^{\ln g(E_1) - \ln g(E_2)}, 1)$, let $\ln[g(E_2)] = \ln f_i + \ln[g(E_2)]$, $H(E_2) = H(E_2) + 1$, $E_1 = E_2$, and $c_1 = c_2$ and go to step (vii); otherwise, let $\ln[g(E_1)] = \ln f_i + \ln[g(E_1)]$ and $H(E_1) = H(E_1) + 1$ and go to step (viii).

(vii) If $E_2 < E_{\text{opt}}$, let $E_{\text{opt}} = E_2$ and $c_{\text{opt}} = c_2$.

(viii) If $H(E) \geqslant \ln 2/\ln f_i$ for all visited energies $E \in M$, go to step (ix); otherwise, go to step (ii).

(ix) Let $f_{i+1} = \sqrt{f_i}$, and $i = i + 1$.

(x) If $f_i \approx 1.0001$, output the lowest-energy conformation c_{opt} and stop; otherwise, reset $H(E) = 0$, keep $g(E)$ for all the energies $E(E \in M)$, and go to step (ii).

IV. NUMERICAL RESULTS

We run the IWL sampling method on both 2D and 3D fcc HP lattice models. The IWL method is coded in Java language and run on a desktop computer with an Intel Core 2 Duo 1.6-GHz processor and 2 GB of RAM. We test two sets of instances from the literature. For each instance, the IWL method is run 20 times independently.

A. Test instances

The first set of instances includes 12 general instances listed in Table I, some of which have been used in literature [9,11,15,16]. The second test set consists of five sets of larger-scale instances, denoted by the S, R, F90, F180, and CASP target instances. The S, R, F90, and F180 instances are taken from Ref. [13] and six CASP target instances are from the CASP website [32].

B. Numerical results and comparison

The first set of 12 instances listed in Table I is widely used to test the performance of the algorithms on square and simple cubic lattice models. Now we first test this set of general instances on the 2D fcc HP lattice model. The computational results are listed in Table II, in comparison with those from the SGA [6], HGA [6], HGATR [7–9], ERSGA [10], HHGA [10], TS [8], and improved energy landscape paving (IELP) methods [22]. Our algorithm can easily reach the lowest energies so far for four short instances 1–4, while for instances 5–9 both the IELP and IWL methods can find the same energies that are lower than the results from the other six methods. For instances 10–12, the IWL sampling method gets the same lowest energies that the IELP method does. However, our algorithm can get average energies for these three instances as good as or lower than those from the IELP method. To further compare the performances between the IWL and ELP methods, we also list the standard errors of the lowest energies in all runs. The standard error shows how much the variation or dispersion from the average value exists. As can be seen in Table II, for each sequence, the standard error by the IWL sampling method is as good as or lower than those by the IELP

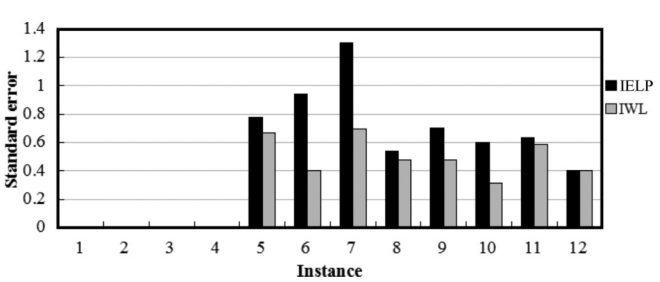

FIG. 6. Distribution of the standard errors of the lowest energies in all runs by the IELP and IWL methods for instances 1–12.

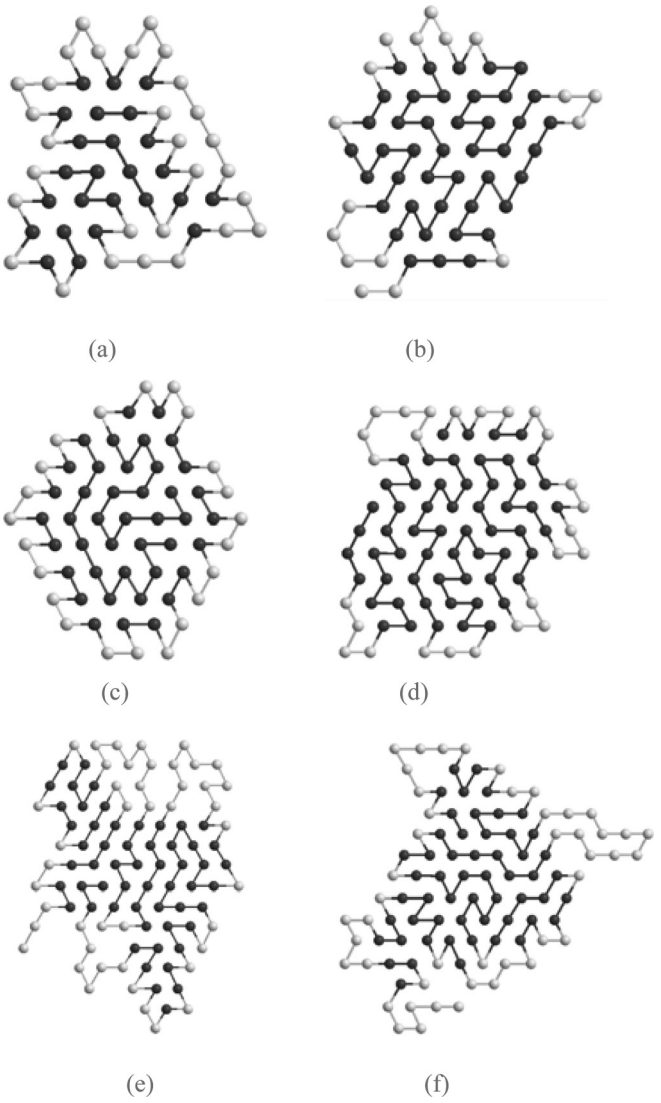

FIG. 7. Conformations with the lowest energies found by the IWL method on the 2D fcc HP lattice model. Black and white circles indicate the hydrophobic and hydrophilic amino acids, respectively. Typical conformations are shown with (a) $E = -44$ of instance 7, (b) $E = -71$ of instance 8, (c) $E = -75$ of instance 9, (d) $E = -101$ of instance 10, (e) $E = -94$ of instance 11, and (f) $E = -94$ of instance 12.

TABLE III. Comparison of computational results by different methods on the 3D fcc lattice model for the longest eight instances in Table I. Numbers in bold indicate the lowest energies by the six methods.

Instance	HGATR	MA	TS	EA	IELP			IWL		
					Best	Avg.	SD	Best	Avg.	SD
5	–69		**–74**	**–74**	**–74**	–74.0	0.000	**–74**	–74.0	0.000
6		–69		**–73**	**–73**	–72.6	0.663	**–73**	–72.6	0.497
7	–59		**–77**		**–77**	–76.6	0.663	**–77**	–76.6	0.490
8	–117	–122	**–130**	**–130**	**–130**	–130.0	0.000	**–130**	–130.0	0.000
9	–103	–114	**–132**	**–132**	**–132**	–132.0	0.000	**–132**	–132.0	0.000
10		–165			**–189**	–188.2	0.980	**–189**	–189.0	0.000
11		–156			**–186**	–185.0	0.775	**–186**	–185.2	0.548
12					**–181**	–180.2	0.600	**–181**	–180.5	0.592

method, which indicates that the lowest energy of each run by the IWL sampling method tends to be closer to the average value. Figure 6 shows the standard error for instances 1–12 by the IELP and IWL methods. Typical conformations by the IWL sampling method for instances 7–12 are shown in Fig. 7.

Further, to verify the effectiveness of the IWL sampling method, we apply it on the 3D fcc HP lattice model. We test 12 general instances listed in Table I. For instances 1–4, the IWL sampling method can easily obtain the optimal energies in the literature and for eight longer ones, the computational results by the IWL sampling method are listed in Table III, in comparison with those from the HGATR [7], MA [15], TS [8], EA [11], and IELP [22] methods. From Table III we can see that the lowest energies by the IWL method are as good as or lower than those by the HGATR, MA, TS, EA, and IELP methods. For instance 5, four out of the six methods (TS, EA, IELP, and IWL) find the lowest energy that is missed by the HGATR, while the MA does not report the result. For instance 6, the EA, IELP, and IWL methods can get lower energy than that by the MA. However, both the HGATR and TS do not report their results. The lowest energy by the TS, IELP, and IWL methods for instance 7 are lower than that by the HGATR, while the other two methods do not report their results. For instances 8 and 9, the IWL sampling method also gets the optimal energies that are obtained by the TS, EA, and IELP methods. For instances 10 and 11, it is obvious that the optimal energies by the IELP method and our method are much better than that by the MA. Only the IELP and IWL methods report the results for sequence 12 and they get the same lowest energy. However, the average lowest energies by the IWL sampling method are as good as or lower than those by the IELP method for all the instances. From Table III we can also see that, for each instance, the standard error by the IWL sampling method is as good as or better than that by the IELP method.

TABLE IV. Comparison of computational results by different methods for the S, R, F90, F180, and CASP target instances on the 3D fcc HP lattice model. Native E. is the optimal energy and is obtained by using HPSTRUCT [33]. Numbers in bold indicate the lowest energies by the six methods. The numbers in parentheses are the average lowest energies.

Instance	Length	Native E.	LST	LSM	SST	SSTHGA	IELP	IWL
S1	135	–357	–351(–341)		**–355**(–347)	–353(–349)	**–355**(–354.23)	**–355**(–354.40)
S2	151	–360	–355(–343)		–354(–347)	–355(–352)	–359(–356.84)	**–360**(–357.83)
S3	161	–367	–355(–340)		–359(–350)	–360(–355)	–364(–362.63)	**–366**(–364.80)
S4	164	–370	–354(–343)		–358(–350)	–363(–356)	–365(–362.63)	**–366**(–363.20)
R1	200	–384	–332(–318)	–353(–326)	–359(–345)	–364(–352)	**–369**(–362.44)	**–369**(–364.88)
R2	200	–383	–337(–324)	–351(–330)	–358(–346)	–364(–355)	–366(–362.60)	**–371**(–368.03)
R3	200	–385	–339(–323)	–352(–330)	–365(–345)	–366(–353)	–370(–362.82)	**–373**(–368.86)
F90_1	90	–168	–164(–160)		**–168**(–166)		**–168**(–166.23)	**–168**(–166.63)
F90_2	90	–168	–165(–158)		**–168**(–164)		**–168**(–167.13)	**–168**(–167.15)
F90_3	90	–167	–165(–159)		**–167**(–165)		**–167**(–166.00)	**–167**(–166.07)
F90_4	90	–168	–165(–159)		**–168**(–165)		**–168**(–167.24)	**–168**(–167.46)
F90_5	90	–167	–165(–159)		**–167**(–165)		**–167**(–166.18)	**–167**(–166.58)
F180_1	180	–378	–338(–327)	–360(–334)	–357(–340)	–359(–348)	–363(–357.68)	**–364**(–361.12)
F180_2	180	–381	–345(–334)	–362(–340)	–359(–345)	–365(–353)	–364(–362.83)	**–367**(–363.60)
F180_3	180	–378	–352(–339)	–357(–343)	–362(–353)	–371(–359)	–368(–363.45)	**–372**(–367.20)
3*mse*	179	–323	–266(–249)	–278(–254)	–289(–280)	–293(–286)	–291(–287.72)	**–294**(–290.31)
3*mr*7	189	–355	–301(–287)	–311(–292)	–328(–313)	–331(–320)	–351(–347.42)	**–353**(–349.75)
3*mqz*	215	–474	–401(–383)	–415(–386)	–420(–403)		–439(–435.88)	**–443**(–440.60)
3*no*6	229	–455	–390(–373)	–400(–375)	–411(–391)	–424(–406)	–415(–411.43)	**–425**(–418.12)
3*no*3	258	–494	–388(–359)	–397(–361)	–412(–393)	–426(–407)	–462(–457.32)	**–477**(–474.50)
3*on*7	279	?	–491(–461)	–499(–463)	–512(–485)	–526(–501)	–548(–546.85)	**–561**(–553.11)

Subsequently, we also perform simulations on five sets of larger-scale instances, denoted as the S, R, F90, F180, and CASP target instances, respectively. Table IV shows the computational results by the LST method [12,13], the LSM method [13], the SST algorithm [13], the SSTHGA [14], the IELP method [22], and the IWL sampling method on the 3D fcc HP lattice model. The lower bounds of the free energy values [in column 3 (Native E.) of Table IV] are obtained from [13]. However, the lower bound of *3on7* is unknown (presented as a question mark). From Table IV we can see that the IWL sampling method wins over the LST, LSM, SST, SSTHGA, and IELP methods for these instances on both the lowest energies and average lowest energies. For instances S2, S3, and S4 we get lower energies than those by the other four methods (LST, SST, SSTHGA, and IELP), except for the LSM, which does not report the results of these three instances, while for instance S1, SST, IELP, and our algorithm obtain the same energy. It is noted that, for instances R_2 and R_3, we can obtain lower energies missed by the other five methods, while for instance R1, the IELP and IWL methods obtain the lowest energy that is missed by the other four methods. From Table IV we can see that the SST, IELP, and IWL methods get the native energies for all the F90 instances that are missed by the LST method, but our method explores the conformation surfaces more efficiently than the SST and IELP methods; for instances F180_1, F180_2, and F180_3 we get lower energies than those by the other five methods. We also find lower energies than those by the LST, LSM, SST, SSTHGA, and IELP methods for all six CASP target instances except for instance *3mqz*, for which the SSTHGA does not report its result. For each instance, the native energy [in column 3 (Native E.) of Table IV] is obtained by using HPSTRUCT [33], which is a state-of-the-art software program and can give the native energy on the fcc HP lattice if one has access to the precomputed H cores [33] of the HP sequence by a different method. It is obvious that the software HPSTRUCT outperforms the IWL method for all instances, except for the five F90 instances; however, if an HP sequence has m H residues and there is no m-residue H core, then HPSTRUCT cannot run and if not all size m H cores are available, then HPSTRUCT may not converge. Even if H cores are available, HPSTRUCT may not converge within a prespecified time limit, in which case no answer is returned [16]. Figure 8 shows typical representatives of the lowest-energy conformations obtained by the IWL sampling method for six CASP target instances. It is obvious that each conformation possesses a compact hydrophobic core.

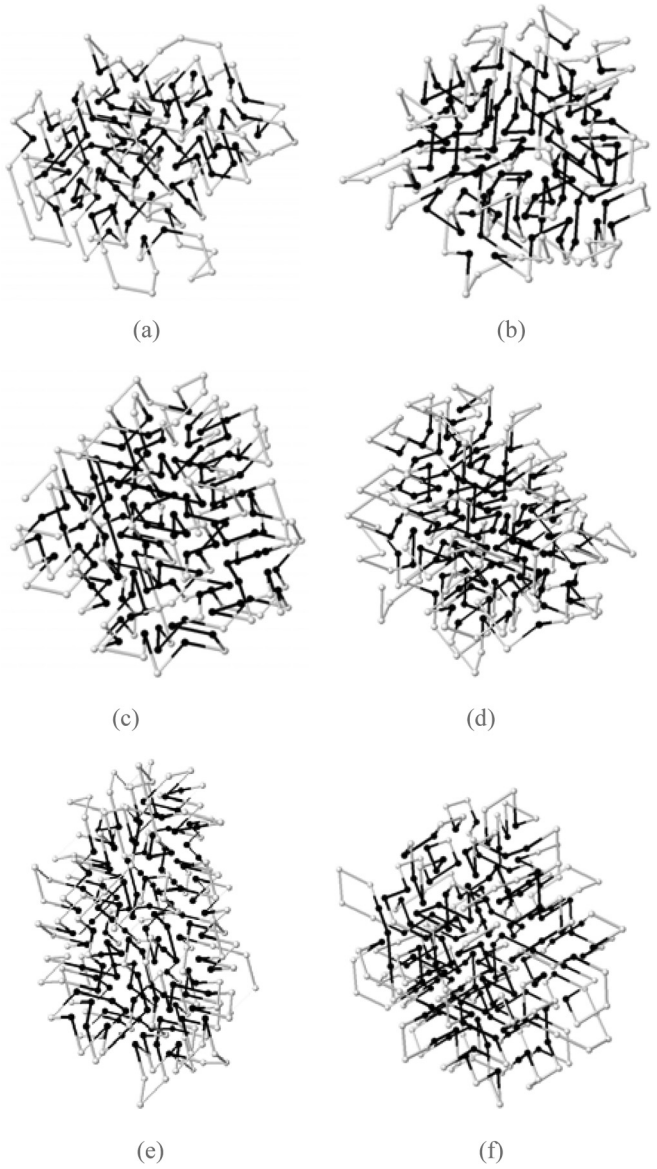

FIG. 8. Conformations with the lowest energies found by the IWL sampling method for six CASP target instances on the 3D fcc HP lattice model. The black and white balls indicate the hydrophobic and hydrophilic amino acids, respectively. Typical conformations are shown with (a) $E = -294$ of instance $3mse$, (b) $E = -353$ of instance $3mr7$, (c) $E = -443$ of instance $3mqz$, (d) $E = -425$ of instance $3no6$, (e) $E = -477$ of instance $3no3$, and (f) $E = -561$ of instance $3on7$.

V. CONCLUSION

It is easy for the search method to get trapped in local minima during the process of finding the ground-state conformations of a protein because of the huge search space of the protein free-energy landscape. To address this problem, we propose the IWL sampling method, which incorporates the generation of an initial conformation based on the greedy strategy and the neighborhood search strategy based on pull moves into the Wang-Landau sampling method for the protein structure prediction on the fcc HP lattice. By modifying the estimate of the density of states at each step of the random walk in energy space and carefully controlling the modification factor, we can determine the density of states very accurately. We compare our results with those by the SGA, HGA, HGATR, ERSGA, HHGA, TS, MA, EA, LST, LSM, SST, SSTHGA, and IELP methods, which achieved the state-of-the-art results for the same instances of the fcc HP lattices. The numerical results show that our method significantly outperforms or is as good as the other methods over the tested instances. Not unexpectedly, this is particularly pronounced for the five sets of larger-scale instances considered.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grants No. 61373016, No. 61103235, and No. 61403206), the "Six Talent Peaks" of Jiangsu Province (Grant No. DZXX-041), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20141005).

[1] K. A. Dill, Biochemistry **24**, 1501 (1985).
[2] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[3] W. E. Hart and S. Istrail, J. Comput. Biol. **4**, 241 (1997).
[4] R. Unger and J. Moult, Bull. Math. Biol. **55**, 1183 (1993).
[5] B. H. Park and M. Levitt, J. Mol. Biol. **249**, 493 (1995).
[6] M. T. Hoque, M. Chetty, and L. S. Dooley, in *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence, Hobart,* 2006, edited by A. Sattar and B. H. Kang (Springer, Berlin, 2006), Vol. 4304, p. 867.
[7] M. T. Hoque, M. Chetty, and A. Sattar, *Proceedings of the 2007 IEEE Congress on Evolutionary Computation, Singapore, 2007* (IEEE, Piscataway, 2007), p. 4138.
[8] H. J. Bökenhauer, A. D. Ullah, L. Kapsokalivas, and K. Steinhöel, in *Proceedings of the 8th International Workshop on Algorithms in Bioinformatics, Karlsruhe, 2008*, edited by K. A. Crandall and J. Lagergren (Springer, Berlin, 2008), Vol. 5251, p. 369.
[9] M. T. Hoque, M. Chetty, A. Lewis, and A. Sattar, IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 234 (2011).
[10] S. C. Su, C. J. Lin, and C. K. Ting, Proteome Sci. **9**, S19 (2011).
[11] S. C. Su and J. J. Tsay, *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, 2012* (IEEE, Piscataway, 2012), p. 1.
[12] M. Cebrián, I. Dotú, P. V. Hentenryck, and P. Clote, *Proceedings of the 23th AAAI Conference on Artificial Intelligence, Chicago, 2008* (AAAI, Menlo Park, 2008), p. 241.
[13] M. A. Rashid, M. A. H. Newton, M. T. Hoque, S. Shatabda, D. N. Pham, and A. Sattar, BMC Bioinf. **14**, S16 (2013).
[14] M. A. Rashid, M. A. H. Newton, M. T. Hoque, and A. Sattar, *Proceedings of the 2013 IEEE Congress on Evolutionary Computation, Cancun, 2013* (IEEE, Piscataway, 2013), p. 1091.
[15] J. J. Tsay and S. C. Su, *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshop, Atlanta, 2011* (IEEE, Piscataway, 2011), p. 315.
[16] I. Dotu, M. Cebrián, P. V. Hentenryck, and P. Clote, IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 1620 (2011).
[17] M. Bachmann, H Arkin, and W Janke, Phys. Rev. E **71**, 031906 (2005).
[18] P. M. C. de Oliveira, Eur. Phys. J. B **6**, 111 (1998).
[19] J. S. Wang, Phys. J. B **8**, 287 (1999).
[20] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[21] J. F. Liu, W. B. Huang, W. J. Liu, B. B. Song, Y. Y. Sun, and M. Chen, J. Korean Phys. Soc. **64**, 603 (2014).
[22] J. F. Liu, B. B. Song, Z. X. Liu, W. B. Huang, Y. Y. Sun, and W. J. Liu, Phys. Rev. E **88**, 052704 (2013).
[23] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[24] T. Wüst and D. P. Landau, Comput. Phys. Commun. **179**, 124 (2008).
[25] T. C. Hales, Ann. Math. **162**, 1065 (2005).
[26] C. G. Zhou and R. N. Bhatt, Phys. Rev. E **72**, 025701 (2005).
[27] A. N. Morozov and S. H. Lin, Phys. Rev. E **76**, 026701 (2007).
[28] A. N. Morozov and S. H. Lin, J. Chem. Phys. **130**, 074903 (2009).
[29] N. Lesh, M. Mitzenmacher, and S. Whitesides, in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin 2003*, edited by M. Vingron, S. Istrail, P. Pevzner, and M. Waterman (ACM, New York, 2003), p. 188.
[30] A. D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, 1995), Chap. 2, p. 47.
[31] J. F. Liu, G. Li, and J. Yu, Phys. Rev. E **84**, 031934 (2011).
[32] http://predictioncenter.org/casp9/targetlist.cgi
[33] R. Backofen and S. Will, in *Proceedings of the 19th International Conference on Logic Programming, Mumbai, 2003*, edited by C. Palamidessi (Springer, Berlin, 2003), Vol. 2916, p. 49.