Protein-folding simulations of the hydrophobic-hydrophilic model by combining pull moves with energy landscape paving

Jingfa Liu,[1,2,*] Gang Li,[3] and Jun Yu[4]

[1]*School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*Network Information Center, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[3]*School of Mathematics & Physics, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[4]*Department of Mathematics & Statistics, The University of Vermont, Burlington, Vermont 05405, USA*

(Received 4 June 2011; revised manuscript received 29 July 2011; published 30 September 2011)

The energy landscape paving (ELP) method is a class of heuristic global optimization algorithms based on Monte Carlo sampling. By incorporating the generation of an initial conformation based on a greedy strategy, the conformation update mechanism based on pull moves, and some heuristic off-trap strategies into an improved ELP method, we propose an alternative version of the ELP method, called the ELP–pull move method. We test the ELP-pull move method on both two-dimensional (2D) and 3D hydrophobic-hydrophilic protein-folding models. For ten 2D benchmark sequences of length ranging from 20 to 100, the proposed algorithm finds the lowest energies so far. Within the achieved results, the algorithm converges more rapidly and efficiently than previous methods. For all ten 3D sequences with a length of 64, the ELP–pull move method finds lower energies within comparable computational times. The numerical results demonstrate that our algorithm is a powerful method to study the lattice protein-folding model.

DOI: 10.1103/PhysRevE.84.031934 PACS number(s): 87.15.Cc, 87.15.ak, 05.10.Ln

I. INTRODUCTION

Predicting a protein's tertiary structure from its primary amino acid sequence is one of the most challenging problems in biology. Two major difficulties have been challenging researchers. One is the design of appropriate energy functions. The effective energy function can generally distinguish the native states from non-native states of a protein molecule. The other is the exploration of the vast space of all possible structures. The latter has been suggested as the current bottleneck [1] and attracts many experts who work on computational problems in local and global optimization. In this paper, we address to the latter, which can be investigated on a particular simplified model: the hydrophobic-hydrophilic (HP) lattice model [2] in both two and three dimensions, which is a widely studied abstract one and has been used by biochemists to evaluate hypotheses of protein structure formation.

In order to search for the minimum-energy conformations for the HP model, people have presented many heuristic algorithms, including the genetic algorithm (GA) [3] and its variations (the improved genetic algorithm (IGA) with a different selection scheme and multiple-point crossover [4] and guided genetic algorithm (GGA) [5]), particle swarm optimization (PSO) [6], the evolutionary Monte Carlo method [7], varieties of the Monte Carlo (MC) method [8], the hybrid of the GA and tabu search [9], the elastic net algorithm with local search [10,11], the gradient-directed Monte Carlo (GDMC) method [12], the GA based on optimal secondary structures (GAOSS) [13], and the hybrid Taguchi genetic algorithm [14], which combines a genetic algorithm, the Taguchi method, and PSO. Meanwhile, some statistic approaches such as fragment regrowth via energy-guided sequential sampling [15], equienergy sampling [16], extremal optimization with constrained structure [17], sequential importance sampling with pilot-exploration resampling (SISPER) [18], and the pruned-enriched Rosenbluth method [19] and its variations (nPERMis [20] and nPERMh [21]), have also been applied to the HP model.

In this paper, energy landscape paving (ELP) [22] with pull moves is proposed for the protein-folding problem in the HP lattice model. The ELP method is an improved Monte Carlo global optimization method that has been successfully applied to solving off-lattice protein models [23–25] and circular packing problems [26,27]. By incorporating the generation of an initial conformation based on a greedy strategy, the conformation update mechanism based on pull moves, and some heuristic off-trap strategies into the improved ELP method, an alternative version of the ELP method, called the ELP–pull move method, is put forth to solve the protein-folding problem in the HP lattice model.

II. THE HP MODEL

The HP model, which was proposed by Dill [28], is a free-energy model based on the belief that interactions between hydrophobic amino acids greatly contribute to the free energy of the natural conformation of a protein. In the HP model, a protein, which can be represented as a string with 20 different letters, is composed of only two types of amino acids: hydrophobic (*H* for nonpolar) and hydrophilic (*P* for polar). Based on this classification, the amino acid sequence of protein is abstracted to a binary string of *H* and *P* residues.

In the HP model, the sequence is folded on a regular (square or simple cubic) lattice and is restricted as a self-avoiding path. A folding of a protein in the HP model means that amino acids are embedded in the lattice such that adjacent residues in sequence occupy adjacent grid points in the lattice and no grid point in the lattice is occupied by more than one residue. Two amino acids are topological adjacent if they are neighbors in the lattice, but are not adjacent in sequence. Then

*Corresponding author: jfliu@nuist.edu.cn

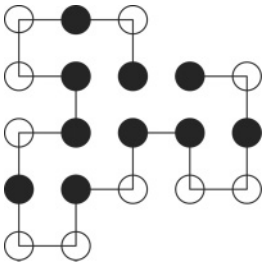

FIG. 1. Conformation of the sequence *HPHPPHHPHPPHPHH-PPHPH* with an energy of −9 in the 2D HP model, where • denotes a hydrophobic and ○ a hydrophilic amino acid.

a topological *H-H* bond is formed between two topological adjacent hydrophobic amino acids. The free energy of a conformation depends on the number of such *H-H* bonds. In other words, if a conformation denoted as $s = s_1 s_2 \ldots s_n$, where s_i is H if the ith amino acid in the sequence is hydrophobic and P if it is hydrophilic, has exactly k such *H-H* bonds, its energy $E(s) = k(-1)$. Figure 1 shows a conformation with an energy of −9 in the two-dimensional (2D) HP model.

The HP lattice protein-folding problem can be formally defined as follows: Given an amino acid sequence $s = s_1 s_2 \ldots s_n$, we try to find an energy-minimizing conformation of s, that is, $c^* \in T(s)$, so that $E(c^*) = \min\{E(c)|c \in T(s)\}$, where $T(s)$ is the set of all valid conformations of s. Even though the HP lattice model is highly simplified, it has been proven that the corresponding folding problem remains NP-complete [29].

III. METHODS

A. Energy landscape paving

The ELP method [22,23] is an improved MC global optimization method. As all good stochastic global optimizers, it is designed to explore low-energy conformations while avoiding entrapment in local minima. This is achieved by performing low-temperature MC simulations, but with a modified energy expression designed to steer the search away from regions that have already been explored. This means that if a conformation c is hit, the energy $E(c)$ is increased by a penalty and replaced by energy $\tilde{E}(c) = E(c) + f(H(q,t))$, where the penalty term $f(H(q,t))$ is a function of the histogram $H(q,t)$. In the present paper, we choose $\tilde{E}(c) = E(c) + kH(E,t)$ as the replacement for the energy E, where $H(E,t)$ is the histogram in energy at a MC sweep t and k is a constant. During the process of ELP,

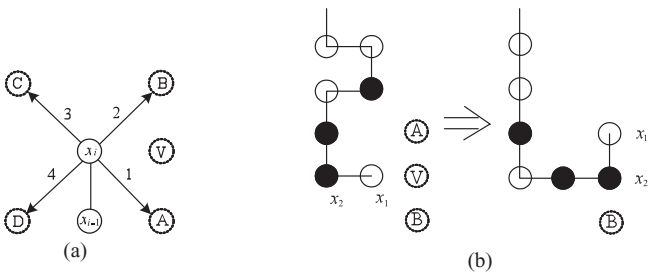

FIG. 2. Pull moves as the conformation update: (a) four move directions of amino acid i ($2 \leqslant i \leqslant n-1$) and (b) end moves of amino acid i ($i = 1$).

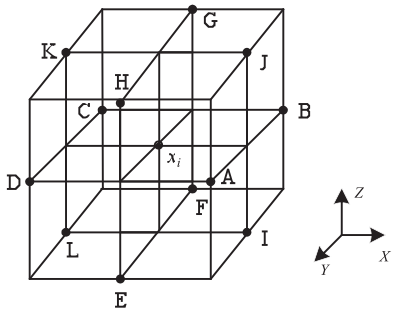

FIG. 3. Twelve potential diagonal move positions (i.e., indices A, B, C, D, E, F, G, H, I, J, K, and L) of the amino acid i ($1 \leqslant i \leqslant n$) in the 3D lattice.

we update the histogram by adding 1 to the frequency of the corresponding bin if the energy $E(c)$ of the newly generated conformation c falls into a certain bin, where a "bin" denotes an entry of the histogram, and all bins in the histogram are the same in size. We set the size of every bin $E_{\text{bin}} = 1$ in the HP model. The sampling weight for a conformation c is defined as $\omega(\tilde{E}(c)) = exp[-\tilde{E}(c)/k_B T]$, where $k_B T$ is the thermal energy at the (low) temperature T and k_B is the Boltzmann constant.

The ELP method has been applied successfully to rough energy landscapes to find low-energy conformations in the off-lattice protein-folding problems [23–25] and the circular packing problems [26,27]. In ELP minimization, the sampling weight of a local minimum conformation decreases with the time that system stays in that minimum and, consequently, the probability to escape the minimum increases. The accumulated histogram function $H(E,t)$ from all previously visited energy at the MC sweeps helps the simulation escape local entrapments and surpass the high-energy barrier more easily. There is, however, a technical flaw in the ELP [26,27]. From the current

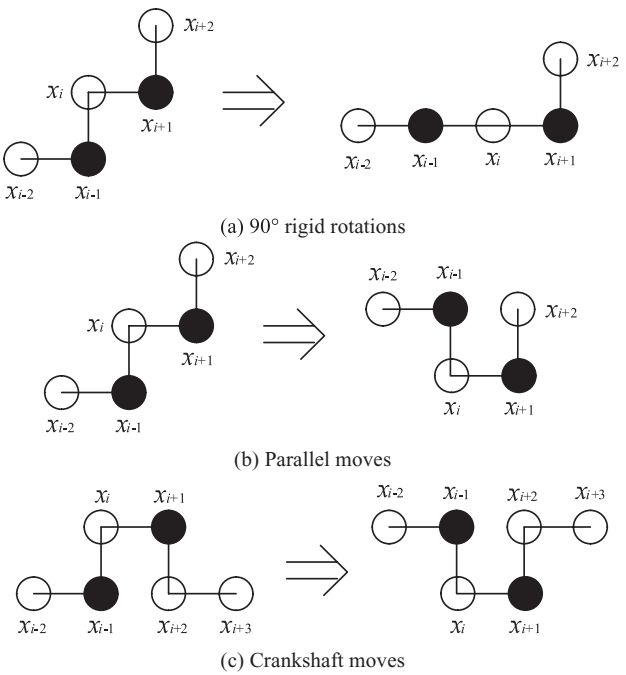

FIG. 4. Three off-trap strategies in the HP model.

TABLE I. Ten 2D HP sequences.

Seq. code	Length	Sequence
2D1	20	*HPHPPHHPHPPHPHHPPHPH*
2D2	24	*HHPPHPPHPPHPPHPPHPPHPPHH*
2D3	25	*PPHPPHHPPPPHHPPPPHHPPPPHH*
2D4	36	*PPPHHPPHHPPPPPHHHHHHHPPHHPPPPHHPPHPP*
2D5	48	*PPHPPHHPPHHPPPPPHHHHHHHHHHPPPPPPHHPPHHPPHPPHHHHH*
2D6	50	*HHPHPHPHPHHHHPHPPPHPPPHPPPPHPPPHPPPHPHHHHPHPHPHH*
2D7	60	*PPHHHPHHHHHHHHPPPHHHHHHHHHHPHPPPHHHHHHHHHHPPPPHHHHHHPHHPHP*
2D8	64	*HHHHHHHHHHHHPHPHPPHHPPHPPHPPHPPHPPHPPHPPHPPHPPHPHPHHHHHHHHHH*
2D9	85	*HHHHPPPPHHHHHHHHHHHHPPPPPPHHHHHHHHHHHPPPHHHHHHHHHHP* *PPHHHHHHHHHHPPPHPPHPPHPPHPH*
2D10	100	*PPPPPPHPHPPPPPHHHPHHHHPHPPPPHPPHPHPHHHHPHHHHHHHHHHPHPHPHHHHHHP* *PPPPPPPPPPHHHHHHPPHPHPPPPPHPHH*

conformation c_1, we consider an attempted move in ELP that yields a lower-energy minimum c_2 that has never been visited before whose energy happens to fall into the same bin that contains other energies previously visited in the earlier steps. The likelihood of accepting this energy minimum c_2 becomes undesirably small as the result of the penalty term $kH(E, t)$, i.e., the ELP search may miss this lower-energy conformation c_2 near c_1. To overcome this shortcoming, Refs. [26,27] present an improvement on the ELP method. In the improved ELP method, the acceptability of the conformation c_2 is determined by a comparison between $E(c_1)$ and $E(c_2)$, where two cases are possible: (i) $E(c_2) < E(c_1)$ and (ii) $E(c_2) \geqslant E(c_1)$. For case (i), the simulation unconditionally accepts the conformation c_2 and starts another round of iteration; for case (ii), if the conformation c_2 satisfies the condition $r(0,1) < \exp\{[\tilde{E}(c_1, t) - \tilde{E}(c_2, t)]/k_BT\}$, where $r(0,1)$ denotes a random number between 0 and 1, then the simulation accepts c_2 and starts another round of iteration; otherwise it does not accept c_2 and restores c_1 as the current conformation.

B. Generation of an initial conformation based on a greedy strategy

For the HP protein-folding problem in two and three dimensions, we build 2D and 3D Cartesian coordinate systems, respectively. The first amino acid and the second amino acid are put at (0, 0) and (0, 1), respectively, for two dimensions and at (0, 0, 0) and (0, 0, 1) for three dimensions. Subsequently the ith ($3 \leqslant i \leqslant n$) amino acid is first pseudoplaced 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 from the corresponding position after computing the energy of the partial conformation, which consists of the previous $i-1$ amino acids placed and the ith amino acid. Then we put the ith amino acid at the position where the energy of the corresponding partial conformation is lowest. This process repeats until a conformation with n amino acids is produced.

C. Conformation update based on pull moves

In ELP, each MC step must update the current conformation. We use a local move set (i.e., pull moves of Refs. [30,31]) as the conformation update move set. The set of pull moves is complete, reversible, and local [30], which makes it efficient for the conformation update. Any valid conformation is connected to all others by pull moves. Therefore, pull moves determine how other amino acids respond in the HP lattice model when one amino acid is moved in one direction.

First, we briefly describe the set of pull moves on 2D square lattice, i.e., a single XY plane. According to the pull move rules, four move directions, i.e., 1, 2, 3, and 4, are defined in Fig. 2. If

TABLE II. Ten 3D HP sequences.

Seq. code	Length	Sequence
3D1	64	*PPHHHHHPPPHHPPPPPHHPPPHPPPPPPHPHPPPHPPHPPHPPPPPHPPPPHHPHHPPHPPHP*
3D2	64	*PPHPHPPHPPHHHPHHHHPPHHHPPPPHPHPPPHPHPPPHPPPPPHPHPPHPHPPPPHPPHP*
3D3	64	*HPHHPPHHPHPPPPPHHHPHHHHPPHPPHPHHPPPPHPHPPHHHPHHPHPPPPPHHHHHHHHPPP*
3D4	64	*HPPHHPPHPPHPHPPHPPPPHPPPPPPPHPHPHHHPPHPHPPPHPHPPHHPPHPPHPPHPHHHPH*
3D5	64	*HPPPHHPPHPHPPPHPPPHPHHPPPHHPHPHHPHPPHPPPHPPHPHHHPPHPPHPPHHHPHHHH*
3D6	64	*HPPHHPHHHHPPPPPPHHPPHPPPPHHPPPHPPHPHHPHPPPPHHPPPPHPPPPPHPPPPHPHH*
3D7	64	*PPPPHPPPHPPPHHHHPHHPPPPPHPPHPHHPHPHPPPPPHPPPPPPPPPPHHHHPPPPHHPPH*
3D8	64	*PPPHHHPPHPHPPHPPHHPPPHPPHPPHHPHPPPHPPPPPPPHPHHHPHHHHHPPHHPPPHPPH*
3D9	64	*HPPHPPHHHPPPPHPHPPPHPHHPHHHHHPPPPHPHPHPPPPHPHPPPHHPHPPPPHPPHHPHP*
3D10	64	*PPHPPHPPHHHPPPHPHPPHPPHPPPPPPHPPHHHPPHPPHPHPHPPPPPPHHHPPPPPHPHP*

TABLE III. Comparison of performances of different methods on the ten 2D HP sequences listed in Table I. NA means data not available. The number in each cell is the minimum energy obtained by the corresponding method for the respective HP sequence. The numbers in parentheses are the numbers of valid conformations scanned before the lowest-energy values are found.

2D Seq.	MC[a]	GA[a]	SISPER[b]	GAOSS[c]	MC-pull move[d]	GDMC-pull move[d]	ELP-pull move[e]
2D1	−9(292 443)	−9(30 492)	−9	−9	−9(149)	−9(496)	−9(127)
2D2	−9(2 492 221)	−9(30 491)	NA	−9	−8(144)	−9(1248)	−9(441)
2D3	−8(2 694 572)	−8(20 400)	NA	−8	−8(674)	−8(1683)	−8(753)
2D4	−13(6 557 189)	−14(301 339)	−14	−14	−14(8895)	−14(4238)	−14(2103)
2D5	−20(9 201 755)	−22(126 547)	−23	−23	−23(29 269)	−23(43 434)	−23(1436)
2D6	−21(15 151 203)	−21(592 887)	−21	−21	−19(170 047)	−21(76 349)	−21(12 523)
2D7	−33(8 262 338)	−34(111 400)	−36	−36	−36(198 486)	−36(194 950)	−36(25 718)
2D8	−35(7 848 952)	−38(97 220)	−39	−42	−42(76 401)	−42(61 233)	−42(38 406)
2D9	NA	NA	−52	−52	−52(233 039)	−53(4 302 404)	−53(722 323)
2D10	NA	NA	−48	NA	−47(295 270)	−48(1 441 692)	−48(619 496)

[a]Values are from Ref. [3].
[b]Values are from Ref. [18].
[c]Values are from Ref. [13]
[d]Values are from Ref. [12].
[e]Values are from the present work.

the position in one direction (such as A, B, C, or D) is already occupied by the other amino acid, the move in that direction is not valid. If position A, B, C, or D is vacant, the moves to these positions may be permitted. Suppose the grid points x_{i-1} and x_i are occupied by the $(i-1)$th and ith amino acids. Consider the pull moves of the ith amino acid. If the grid point A is unoccupied, the fourth grid point in the minisquare including x_{i-1}, x_i, and A is labeled V [see Fig. 2(a)]. If $V = x_{i+1}$ [which is occupied by the $(i+1)$th amino acid], then the pull move simply operates by moving x_i to A and generates another conformation. If both V and A are unoccupied, the pull moves then operate by moving x_i to A, x_{i+1} to V, x_{i+2} to x_i, x_{i+3} to x_{i+1}, and so on, until a new legal conformation is reached by the least number of grid point moves. In the pull moves described above the amino acids are pulled one by one in descending order. By symmetry the amino acid positions can also be pulled in ascending order in a pull move. The first and last amino acids in the chain require special pull moves [30]. In Fig. 2(b), if position A or B is vacant, the moves to these positions may be permitted. Let A and V be two adjacent unoccupied grid points with V adjacent to the first amino acid position x_1. The end move displaces x_1 to A, x_2 to V, x_3 to x_1, x_j to x_{j-2}, and so on, until a new legal conformation is reached by the least number of grid point moves [Fig. 2(b)]. By symmetry, the end move on the last amino acid is similarly defined.

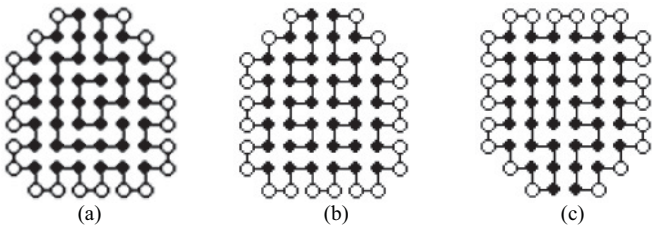

FIG. 5. Three conformations with an energy of −42 for the sequence 2D8 (the length-64 sequence) found by the ELP-pull move method.

In ELP, to update the current conformation, we consider in turn pull moves of every amino acid in the chain. Starting from the first amino acid, we execute pull moves for all legal move positions of each amino acid. The ith $(1 \leqslant i \leqslant n)$ amino acid may be moved at most to the four diagonal positions (see Fig. 2). First, we pseudomove it to every legal position. Then we complete the remaining moves by pull move rules until a new legal conformation is reached. After computing the energy of the corresponding conformation for each legal position, we move the ith amino acid to the position where the energy of the conformation obtained by pull moves is lowest. This process is repeated until a new conformation is accepted (the acceptance criterion is shown in Sec. III A) or pull moves on all n amino acids are executed.

For a 3D cubic lattice, the pull moves of a grid point can be similarly defined as above. It is noteworthy that in the 3D lattice an initial step of one pull move within one of the three planes may induce subsequent moves within the other planes. In addition, in a 2D lattice, the ith $(1 \leqslant i \leqslant n)$ amino acid (which occupies grid point x_i) may at most be moved to the four diagonal positions [see Fig. 2(a)], but in a 3D lattice it may be moved at most to the twelve diagonal positions (see Fig. 3).

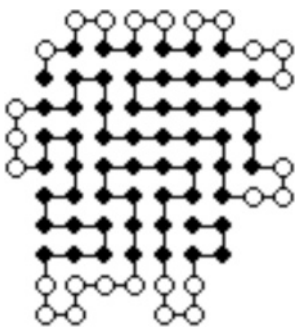

FIG. 6. Conformation with an energy of −53 for the sequence 2D9 (the length-85 sequence) found by the ELP-pull move method.

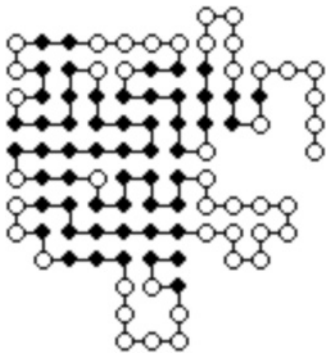

FIG. 7. Conformation with an energy of −48 for the sequence 2D10 (the length-100 sequence) found by the ELP-pull move method.

D. Off-trap strategy

When we apply pull moves to update the conformation in ELP, it is possible that each new conformation obtained by pull moves is not accepted. This means that the search is trapped in local minima that may be located at narrow and deep valleys of the energy landscape. To jump out of local minima, we execute the following three moves in turn: (i) 90° rigid rotations [Fig. 4(a)], (ii) parallel moves [Fig. 4(b)], and (iii) crankshaft moves [Fig. 4(c)]. In 90° rigid rotations, we suppose that grid point x_{i-1} (which is occupied by amino acid $i-1$) is a pivot and the grid points $x_i, x_{i+1}, x_{i+2}, \ldots, x_n$ compose a rigid body. The whole rigid body rotates 90° according to grid point x_{i-1} along the XY plane for a 2D lattice [see Fig. 4(a)] and along one of three planes XY, XZ, and YZ for a 3D lattice, gaining an updating conformation. In parallel moves, the grid points x_i, x_{i+1}, x_{i+2}, . . ., x_n compose a rigid body and the whole rigid body moves parallel along the x or y axis for a 2D lattice and along the x, y, or z axis for a 3D lattice until grid point x_i is still adjacent to x_{i-1} [see Fig. 4(b)]. In crankshaft moves, a U-shaped structure consisting of four connected neighbors in the sequence rotates 180° for a 2D lattice and 90°, 180°, or 270° for a 3D lattice according to the axis consisting of the grid points x_{i-1} and x_{i+2} [see Fig. 4(c)]. If these three moves cannot produce a valid conformation, we restore the previous conformation $\bar{c}$ of the current conformation c as the new current conformation and continue a new round of iteration of ELP.

E. Description of algorithm

By incorporating the generation of an initial conformation based on a greedy strategy, the conformation update mechanism based on pull moves, and some heuristic off-trap strategies into the improved ELP method, an alternative version of the ELP method, called the ELP–pull move method, is developed for the protein-folding problem in the HP lattice model. The calculating procedure is presented as follows.

(i) Generate initial conformation c based on a greedy strategy. Let $\bar{c} = c$ and $c_{\min} = c$. Set $t = 1$, $k = 0.5$, and $T = 5$. Compute $E(c)$ and initialize $H(E(c), t)$. Let $\tilde{E}(c) = E(c) + kH(E(c),t)$.

(ii) Let $i = 1$.

(iii) Execute pull moves for all legal move positions of the ith amino acid of the current conformation c. If at least a pull move is executed successfully, we compute the energies of the corresponding legal conformations obtained by pull moves and pick out the conformation with the lowest energy as a newly updated conformation of c, denoted as c', and go to (iv); otherwise we go to (vii).

(iv) Compute $E(c')$ and $H(E(c'), t)$ and let $\tilde{E}(c') = E(c') + kH(E(c'), t)$.

(v) If $E(c') < E(c)$, then let $c = c'$, $E(c) = E(c')$, $\bar{c} = c$, $c_{\min} = c$, and go to (ix); otherwise go to (vi).

(vi) If $r\ (0, 1) < \exp\{[\tilde{E}(c) - \tilde{E}(c')]/k_BT\}$, where $r(0,1)$ denotes a random number between 0 and 1, then let $c = c'$, $E(c) = E(c')$, $\bar{c} = c$, and go to (ix); otherwise go to (vii).

(vii) If $i > n$, then go to (viii); otherwise let $i = i + 1$ and go to (iii).

(viii) For the current conformation c, we produce the new conformation c' by off-trap strategies. If c' is a legal conformation, then update the current conformation c with c', i.e.,

TABLE IV. Comparison of performances of different methods on the ten 3D HP sequences listed in Table II. NA means data not available. The number in each cell is the minimum energy obtained by the corresponding method for the respective HP sequence. The numbers in parentheses are the numbers of valid conformations scanned before the lowest-energy values are found.

3D Seq.	MC[a]	GA[a]	IGA[b]	GGA[c]	PSO[d]	ELP-pull move[e]
3D1	−12(4 139 486)	−27(2 119 775)	−22	NA	−28(1 131 552)	−31(102 640)
3D2	−17(4 086 574)	−29(2 286 289)	−26	−30(14 186)	−31(456 877)	−36(209 922)
3D3	−24(3 958 530)	−35(1 831 102)	−35	−38(493)	−39(113 315)	−44(101 422)
3D4	−18(4 077 468)	−34(2 315 112)	NA	−36(27 450)	−36(1 730 129)	−39(1 007 703)
3D5	−20(4 027 596)	−32(2 040 915)	NA	NA	−38(1 602 646)	−41(737 062)
3D6	−16(4 114 480)	−29(2 160 690)	NA	−30(1899)	−31(410 586)	−34(405 219)
3D7	−15(4 128 584)	−20(2 317 862)	NA	NA	−27(1 296 319)	−28(51 015)
3D8	−19(4 067 513)	−29(2 391 876)	NA	NA	−35(1 113 330)	−37(87 188)
3D9	−19(4 053 207)	−32(2 121 287)	NA	−34(14 848)	−35(404 199)	−39(255 703)
3D10	−14(4 125 584)	−24(2 287 394)	NA	−25(362)	−27(175 053)	−31(111 953)

[a]Values are from Ref. [3].
[b]Values are from Ref. [4].
[c]Values are from Ref. [5].
[d]Values are from Ref. [6].
[e]Values are from the present work.

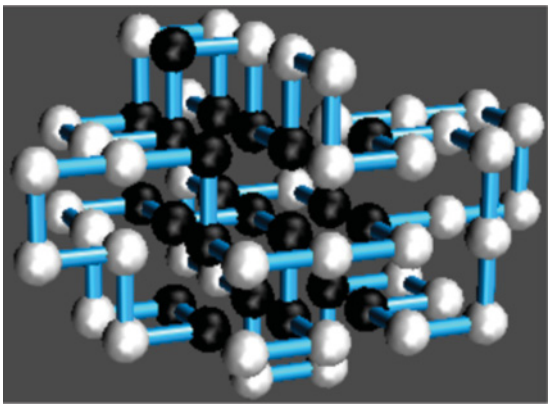

FIG. 8. (Color online) Typical conformation with $E = -31$ for the sequence 3D1.

let $c = c'$, $E(c) = E(c')$, and set $\bar{c} = c$; otherwise let $c = \bar{c}$ and $E(c) = E(\bar{c})$.

(ix) If $t > 5\times10^6$, then output the lowest-energy conformation $c_{\min}$ and stop; otherwise let $t = t + 1$ and go to (ii).

IV. COMPUTATIONAL RESULTS AND ANALYSIS

We test the ELP method combined with pull moves on the HP model in both two and three dimensions. The tested instances include ten 2D benchmark sequences with length ranging from 20 to 100 [12] and ten 3D benchmark sequences with a length of 64 [3]. These benchmark instances have in part been used extensively in the literature [3–21]. A complete listing of the 2D and 3D sequences can be found in Tables I and II, respectively. We implement the ELP-pull move method in Java language and run it on a Notebook PC with an Intel Core 2 Duo, 1.6-GHz processor and 1.0 GB of RAM. For each instance, the ELP–pull move algorithm is run five times independently.

For ten 2D sequences [12], Table III summarizes the ELP-pull move algorithm's performance as well as other methods' reported in the literature, including the MC method [3], the GA [3], the SISPER [18], the GAOSS [13], the MC method with pull moves [12], and the GDMC method with pull moves [12].

From Table III one can see that the lowest free energies for the three shortest protein sequences 2D1(20-mer), 2D2(24-mer), and 2D3(25-mer) obtained by the seven aforementioned methods are all the same, except for SISPER, which does not report the results for sequences 2D2 and 2D3. For the other four slightly shorter sequences 2D4(36-mer), 2D5(48-mer), 2D6(50-mer), and 2D7(60-mer), SISPER, the GAOSS, the GDMC–pull move method, and the ELP–pull move method find the ground-state conformations with the lowest free energies, but the three other methods miss the optimal results in some cases, for example, the MC method in sequences 2D4, 2D5, and 2D7; the GA in sequences 2D5 and 2D7; and the MC–pull move method in the sequence 2D6. For the sequence 2D8(64-mer), four out of the seven methods (the GAOSS, the MC–pull move method, the GDMC–pull move

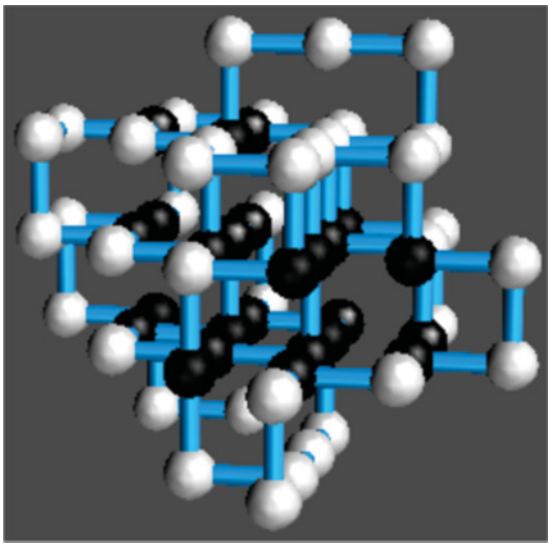

FIG. 9. (Color online) Typical conformation with $E = -36$ for the sequence 3D2.

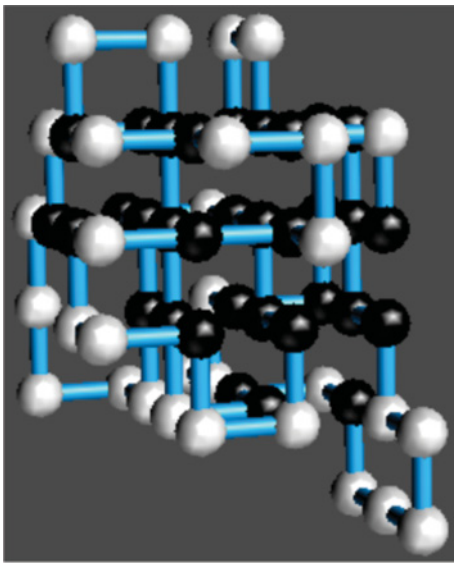

FIG. 10. (Color online) Typical conformation with $E = -44$ for the sequence 3D3.

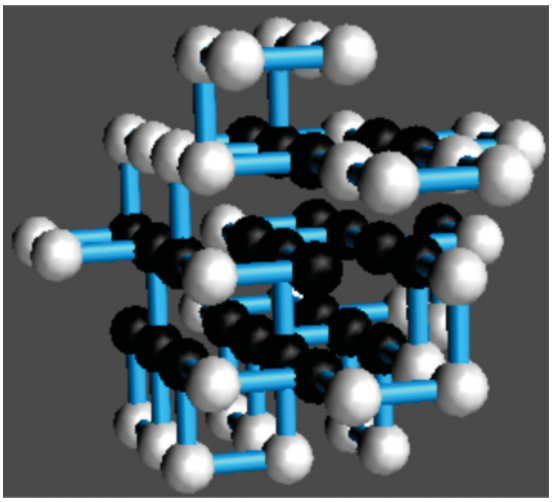

FIG. 11. (Color online) Typical conformation with $E = -39$ for the sequence 3D4.

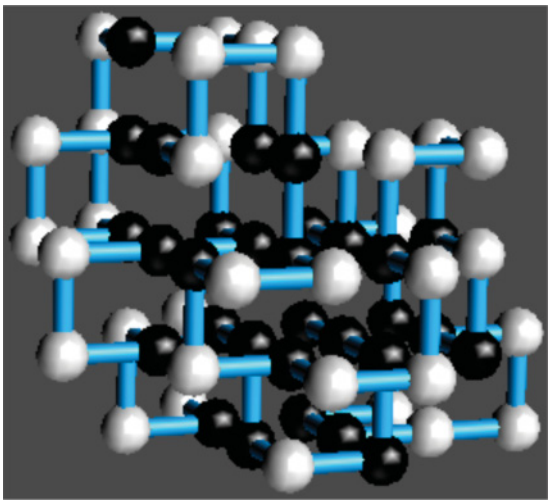

FIG. 12. (Color online) Typical conformation with $E = -41$ for the sequence 3D5.

method, and the ELP–pull move method) find conformations with an energy of -42. Three quite different conformations found by the ELP–pull move method are shown in Fig. 5. Shown in Figs. 5(a) and 5(c) are α-helix, β-sheet, β-turn, and their mixture secondary structures, while in Fig. 5(b) there are mainly α-helix secondary structures. For the sequence 2D9(85-mer), two out of the seven methods (the GDMC–pull move and ELP–pull move methods) find the ground-state conformations with the lowest free energy of -53, but SISPER, the GAOSS, and the MC–pull move method find an energy of -52; the other two methods (the MC method without pull moves and the GA) do not report results. One of the ground-state conformations by the ELP–pull move method for the sequence 2D9 is shown in Fig. 6. For the sequence 2D10(100-mer), three out of the seven methods (SISPER, the GDMC–pull move method, and the ELP–pull move method) find conformations with an energy of -48; the MC method without pull moves, the GA, and the GAOSS do not report results of this instance, whereas the MC–pull move method

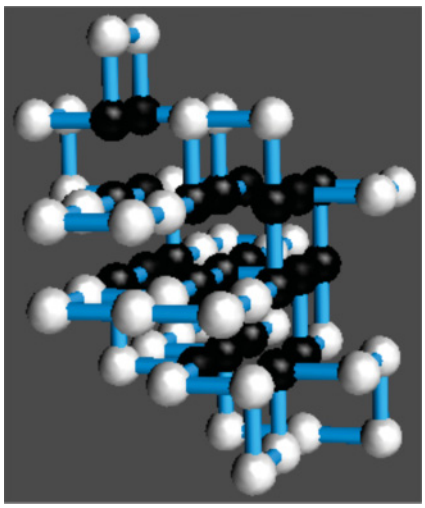

FIG. 13. (Color online) Typical conformation with $E = -34$ for the sequence 3D6.

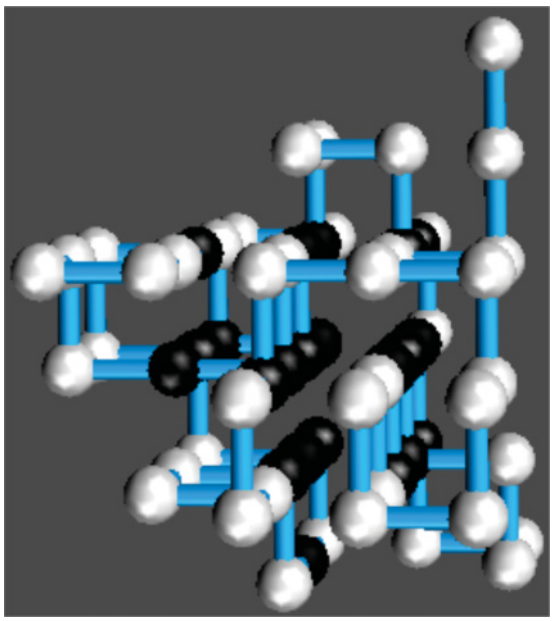

FIG. 14. (Color online) Typical conformation with $E = -28$ for the sequence 3D7.

finds an energy of -47. One of the ground-state conformations by the ELP–pull move method for the sequence 2D10 is shown in Fig. 7. From Figs. 5–7, one can see that all of the conformations possess a compact hydrophobic core.

In addition, the number of valid conformations scanned in the ELP–pull move method for each sequence is also listed in Table III in comparison with the results by the MC method [3], the GA [3], the MC–pull move method [12], and the GDMC–pull move method [12]. Table III shows that the

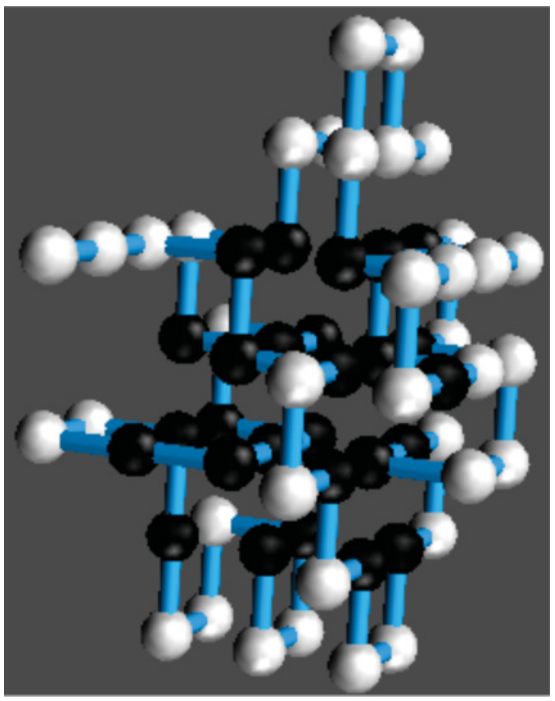

FIG. 15. (Color online) Typical conformation with $E = -37$ for the sequence 3D8.

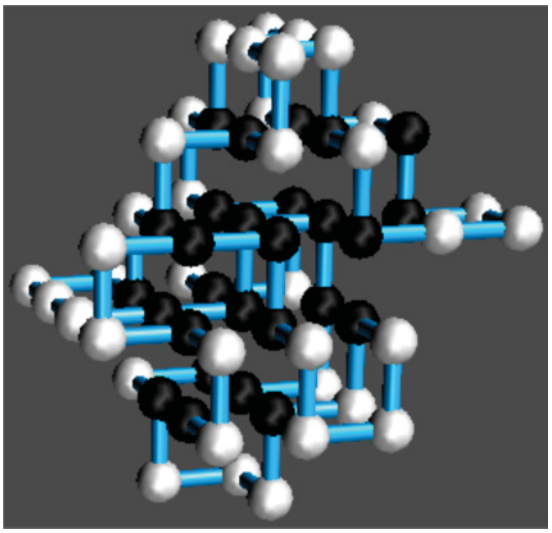

FIG. 16. (Color online) Typical conformation with $E = -39$ for the sequence 3D9.

ELP–pull move method takes less time to obtain the lowest energy than these four methods. There is no information about the time cost for SISPER [18] and the GAOSS [13], so we cannot make such a comparison. From Table III one can also see that the putative ground-state energies obtained by the GDMC–pull move and ELP–pull move methods for all ten 2D sequences are the same, but the ELP–pull move method requires much less time to obtain them than the GDMC–pull move method. Therefore, the ELP–pull move method explores the conformation surfaces more efficiently than the MC method, the GA, SISPER, the GAOSS, the MC–pull move method, and the GDMC–pull move method.

For ten 3D sequences with 64-mers [3], we list the results of the ELP-pull move algorithm in Table IV in comparison with those from the MC method [3], the GA [3], the IGA with a different selection scheme and multiple-point crossover [4], the GGA [5], and PSO [6]. From Table IV one can see that the ELP–pull move method finds different lower free energies than these five methods for all ten 3D sequences. Figures 8–17 show typical conformations of these putative lowest-energy states. It is obvious that each of these conformations possesses a compact hydrophobic core. Moreover, the ELP–pull move method scans fewer valid conformations than the MC method [3], the GA [3], and PSO [6] to obtain the lowest free energy for every sequence. In addition, one can notice that the GGA [5] spent less time obtaining results for sequences 3D2, 3D3, 3D4, 3D6, 3D9, and 3D10 than the ELP–pull move method, but the lowest energies from the GGA are far higher (at least a difference of -3) than those from our method. For the other four sequences, we cannot make a comparison between the ELP–pull move method and the GGA since the GGA does not report the corresponding results.

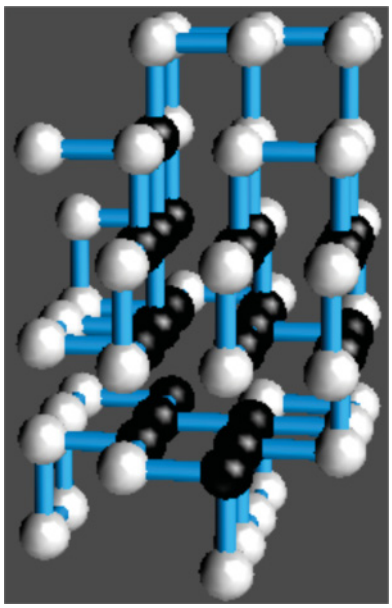

FIG. 17. (Color online) Typical conformation with $E = -31$ for the sequence 3D10.

V. CONCLUSION

The stochastic global optimization method of ELP has been applied successfully to the rough energy landscape of continuous space to find low-energy conformations, including the off-lattice protein-folding problems and the circular packing problems. However, few researchers have applied ELP to discrete combinatorial optimization problem. In this paper, to demonstrate the efficiency of ELP in discrete space, the protein-folding problem in both 2D and 3D HP lattice models is studied and the performance of the ELP procedure for each case is described. In ELP, we introduce pull moves to execute a local search and update conformation. Because the landscape for protein folding is rugged and the set of pull moves is local, ELP may trap the optimization in some local minima that are located at narrow and deep valleys of the energy landscape. Thus heuristic off-trap strategies are used to jump out of the local minima.

Our results indicate that for the 2D model the ELP method combined with pull moves explores the conformation surface more efficiently than the MC method, the GA, SISPER, the GAOSS, the MC–pull move method, and the GDMC–pull move method and finds the lowest known energies obtained by previous researchers in all cases. Moreover, for all ten 3D sequences, the ELP–pull move algorithm outperforms the MC method, the GA, the IGA, the GGA, and PSO and finds different lower energies within comparable computational times. Briefly, this paper presents an alternative, more efficient algorithm for finding lower-free-energy states of 2D and 3D HP lattice proteins. It is not hard to see that the proposed method is easy to extend to other discrete optimization problems. Furthermore, our study suggests that a proper combination of the normal stochastic global optimization method and local search can reduce the cost of the stochastic global search and enhance the efficiency of the search of the algorithm. It could be an efficient mechanism to construct a high-performance algorithm in a certain problem.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Jiangsu Province (Grant No. BK2010570), the China

Postdoctoral Science Foundation (Grant No. 20100471350), Special Foundation of China Postdoctoral Science Foundation (Grant No. 201104572), Jiangsu Planned Projects for Postdoctoral Research Funds (Grant No. 1001030B), Natural Science Foundation of Education Committee of Jiangsu Province (Grant No. 09KJB520008), the National Public Benefit Research Foundation of China (Grant No. GYHY200906006), and the Qing Lan Project.

[1] P. Bradley, K. M. Misura, and D. Baker, Science **309**, 1868 (2005).
[2] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[3] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993); *A Genetic Algorithm for 3D Protein Folding Simulations*, Proceedings of the Fifth International Conference on Genetic Algorithms, edited by S. Forrest (Morgan Kaufmann, San Francisco, 1993), p. 581.
[4] F. L. Custódio, H. J. C. Barbosa, and L. E. Dardenne, Genet. Mol. Biol. **27**, 611 (2004).
[5] M. T. Hoque, M. Chetty, and L. S. Dooley, in *Proceedings of the IEEE Congress on Evolutionary Computation*, Vancouver, 2006, edited by G. G. Yen, S. M. Lucas, G. Fogel, G. Kendall, R. Salomon, B.-T. Zhang, C. A. Coello Coello, and T. P. Runarsson (IEEE, New York, 2006), p. 2339.
[6] F. Kanj, N. Mansour, H. Khachfe, and F. Abu-Khzam, *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, Rabat, Morocco, 2009* (IEEE, Piscataway, NJ, 2009), p. 732.
[7] F. Liang and W. H. Wong, J. Chem. Phys. **115**, 3374 (2001).
[8] U. Bastolla, H. Frauenkron, E. Gerstener, P. Grassberger, and W. Nadler, Proteins Struct. Funct. Genet. **32**, 52 (1998).
[9] T. Z. Jiang, Q. H. Cui, G. H. Shi, and S. D. Ma, J. Chem. Phys. **119**, 4592 (2003).
[10] Y. Z. Guo, E. M. Feng, and Y. Wang, J. Chem. Phys. **125**, 154102 (2006).
[11] Y. Z. Guo and E. M. Feng, J. Chem. Phys. **125**, 234703 (2006).
[12] X. Q. Hu, D. N. Beratan, and W. T. Yang, J. Chem. Phys. **131**, 154117 (2009).
[13] C. H. Huang, X. B. Yang, and Z. H. He, Comput. Biol. Chem. **34**, 137 (2010).
[14] C. J. Lin and M. H. Hsieh, Expert Syst. Appl. **36**, 12446 (2009).
[15] J. F. Zhang, S. C. Kou, and J. S. Lin, J. Chem. Phys. **126**, 225101 (2007).
[16] S. C. Kou and J. Oh, J. Chem. Phys. **124,** 244903 (2006).
[17] H. Y. Lu and G. K. Yang, Comput. Math. Appl. **57**, 1855 (2009).
[18] J. L. Zhang and J. S. Liu, J. Chem. Phys. **117**, 3492 (2002).
[19] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[20] H. P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003); Phys. Rev. E **68**, 021113 (2003).
[21] W. Q. Huang and Z. P. Lü, Chinese Sci. Bull. **49**, 2092 (2004); W. Q. Huang, Z. P. Lü, and H. Shi, Phys. Rev. E **72**, 016704 (2005).
[22] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[23] J. F. Liu and W. Q. Huang, J. Theor. Comput. Chem. **5**, 587 (2006).
[24] A. Schug, W. Wenzel, and U. H. E. Hansmann, J. Chem. Phys. **122**, 194711 (2005).
[25] J. F. Liu, S. J. Xue, D. B. Chen, H. T. Geng, and Z. X. Liu, J. Biol. Phys. **35**, 245 (2009).
[26] J. F. Liu and G. Li, Sci. China Inf. Sci. **53**, 885 (2010).
[27] J. F. Liu, S. J. Xue, Z. X. Liu, and D. H. Xu, Comput. Indust. Eng. **57**, 1144 (2009).
[28] K. A. Dill, Biochemistry **24**, 1501 (1985).
[29] B. Berger and T. Leighton, J. Comput. Biol. **5**, 27 (1998).
[30] 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 Martin Vingron, Sorin Istrail, Pavel Pevzner, and Michael Waterman (ACM, New York, 2003), p. 188.
[31] A. A. Albrecht, A. Skaliotis, and K. Steinhöfel, Comput. Biol. Chem. **32**, 248 (2008).