Journal of the Korean Physical Society, Vol. 45, No. 6, December 2004, pp. 1450∼1454

Conformational Space Annealing and a Lattice Model Protein

Julian Lee*

Department of Bioinformatics and Life Science, Bioinformatics and Molecular Design Technology Innovation Center, and Computer Aided Molecular Design Research Center, Soongsil University, Seoul 156-743

(Received 14 September 2004)

The conformational space annealing (CSA) method is a powerful global optimization method for sampling low-lying local minimum energy conformations of a physical system. In this work, I apply CSA to the study of a two-dimensional HP lattice model of a protein, where a conformation is defined as a self-avoiding chain on a lattice. I study the 36-residue chain with a particular sequence $HPH_2P_2H_2P_3H_2PHP_3H_2P_2H_2P_4HPH_2PHP_2$ presented by Li *et al.*, for which by exhaustive enumeration of *compact* conformations, only one such conformation with the lowest energy was shown to exist. The CSA algorithm finds conformations with energies lower than those found by Li *et al.* for 100 independent runs, demonstrating that the global minimum energy conformation is not necessarily the most compact structure.

PACS numbers: 05.50.+q, 02.60.Pn, 87.14.Ee, 87.15.Cc
Keywords: Lattice model, HP protein, Conformational space annealing

I. INTRODUCTION

Finding the global minimum (GM) of a given function, called global optimization, is an important problem in various fields of science and engineering. One of the simplest algorithms for global optimization is the simulated annealing (SA) method [1], which has been applied to many systems. Although the SA is very versatile in that it can be applied to many problems, the drawback is that its efficiency is usually much lower than problem specific algorithms. This is especially problematic for NP-hard problems such as protein folding or molecular cluster optimizations. For this reason, it is important to find an algorithm which is as general as SA, and yet competitive with problem-specific ones.

Recently, a powerful global optimization method called conformational space annealing (CSA) was proposed and applied extensively and exclusively to various models of proteins [2–12] and a Lennard-Jone cluster [13]. The benchmark tests [2–4,10,11,13] demonstrated that it could not only find the known GM conformations with less computations than existing algorithms, but also provided new GMs in some cases [3,4].

It should be noted that although the CSA method has been applied, so far, only to off-lattice systems, it can be readily applied to lattice models [14,15]. In fact, if the CSA is to be applied to an optimization problem, usually only two things are necessary; a method for perturbing a seed configuration, and a distance measure between two configurations (See Sec. II). Additionally, since the local minimization is a crucial component of the CSA method, one has to make a suitable *approximate* definition of local minimization for the case of lattice models, in contrast to off-lattice systems where the local minimization is well defined.

In this work, I apply the CSA method for the global optimization of a lattice model, the HP model of proteins [16–27]. The HP model is a simplified model of proteins in which amino acids are modelled as a point with hydrophobic inter-residue energy only, and a protein conformation is defined as a self-avoiding chain on a lattice [28–32]. In particular, in Refs. 20 and 25, the authors studied all possible sequences of 27-residue protein on a three-dimensional lattice and 36-residue protein on a two-dimensional lattice. By exhaustively enumerating all *compact* conformations filling a cube of size $3 \times 3 \times 3$ and a square of 6×6, whose total numbers amount to 51704 and 28728 respectively, the sequences with unique ground states were shown to comprise only a small fraction of the set of possible sequences. In particular, for two-dimensions, the 36-residue sequence $HPH_2P_2H_2P_3H_2PHP_3H_2P_2H_2P_4HPH_2PHP_2$ was presented as an example of such a sequence, along with the corresponding global minimum energy conformation *among compact structures*. I sample conformations of this sequence with the CSA and find conformations with energies lower than the compact conformation presented in Ref. 20. The result shows that the ground-state structure is not necessarily the most compact structure and that it can be missed even by an exhaustive enumeration if that enumeration is restricted only to compact conformations.

*E-mail: jul@ssu.ac.kr;˜˜Fax: +82-2-812-5762

II. METHODS

In the original HP model, the conformations of a polymer chain with N monomers are modelled as two-dimensional self-avoiding chains of length N on a square lattice. The bond length is unity, so the position of monomer i is given by $\mathbf{r}_i = (k, l)$, where integers k and l are the Cartesian coordinates relative to an arbitrary origin. Chain connectivity requires $|\mathbf{r}_i - \mathbf{r}_{i+1}| = 1$. Because of the excluded volume, there can be no more than one monomer on each lattice site, $\mathbf{r}_i \neq \mathbf{r}_j$ for $i \neq j$. Beads on the chain are of two types, H and P standing for Hydrophobic and Polar. The potential energy is defined as

$$\begin{aligned} U &= \sum_{i<j} E_{\sigma_i\sigma_j} \Delta(\mathbf{r}_i - \mathbf{r}_j) \\ &= E_{HH} n_{HH} + E_{PP} n_{PP} + E_{HP} n_{HP}, \end{aligned} \tag{1}$$

where σ_i is the amino acid type of the i-th residue. Also, $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 1$ if $\mathbf{r}_i$ and $\mathbf{r}_j$ are adjoining sites but $i \neq j \pm 1$, and $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 0$ otherwise. In the second expression, $n_{\alpha\beta}$ is the number of contacts between residues of types α and β.

Since the numerical values $E_{HH} = -2.3$, $E_{HP} = -1$, and $E_{PP} = 0$ were used for the calculations in Refs. 20 and 25, I use the same relative scale for ease of comparison, but scale by overall factor of 10 to make them integer for the convenience of calculation. In this unit, the lowest energy *compact* conformation found by Li *et al.* [20] has energy

$$E^0_{compact} = -320, \tag{2}$$

with $n_{HH} = 10$, $n_{HP} = 9$, and $n_{PP} = 6$. If we allow only self-avoiding chains throughout the algorithm, it is extremely difficult to generate conformations because most of the conformations will be rejected due to overlapping residues. Therefore, in this work, I allow overlap of residues, but with an energy penalty. This additional term is of the form

$$U_{repulsion} = E_{repulsion} \sum_{i<j} \Delta'(\mathbf{r}_i - \mathbf{r}_j), \tag{3}$$

where $\Delta'(\mathbf{r}_i - \mathbf{r}_j) = 1$ only if $\mathbf{r}_i - \mathbf{r}_j = 0$ and $\Delta'(\mathbf{r}_i - \mathbf{r}_j) = 0$ otherwise. With a sufficiently large positive value of $E_{repulsion}$, conformations with overlapping residues get quickly eliminated during the energy minimization due to their relatively high energies, and the resulting conformations are the local minimum energy conformations of the original potential energy in Eq. (1). In this work, I use $E_{repulsion} = 1000$ and find that this value is good enough for my purpose.

To elaborate on the details of the CSA method, we first randomly generate a certain number of initial conformations (50 in this work) whose energies are subsequently minimized. We call the set of these conformations the *bank*. We make a copy of the bank and call it the *first bank*. The conformations in the bank are updated in later stages whereas those in the first bank are kept unchanged. Also, the number of conformations in the bank is kept unchanged when the bank is updated. The diversity of the bank conformations is controlled by a parameter $D_{\rm cut}$ (see below), and its initial value is set as $D_{\rm ave}/2$, where $D_{\rm ave}$ is the average distance between the conformations in the first bank. New conformations are generated by choosing a certain number (30 in this work) of seed conformations from the bank and by replacing parts of the seeds with the corresponding parts of conformations randomly chosen from either the first bank or the bank. A residue of a seed conformation is randomly selected, and a continuous segment starting from this residue is replaced, where the segment length is also randomly determined between 1 and $0.4N_{seq}$. In this work, 10 conformations are generated for each seed by using partial replacements. Then, the energies of these conformations are subsequently minimized (trial conformations).

A newly obtained local minimum conformation α is compared with those in the bank to decide how the bank should be updated, unless the energy of the conformation α is higher than those of the bank conformations. The definition of the distance measure between conformations is crucial for this procedure, and in this work, the distance measure $D(A, B)$ between two conformations A and B is defined as

$$D(A, B) = \frac{2}{\pi} \sum_{i=1}^{N_{seq}-1} |\theta_i(A) - \theta_i(B)|, \tag{4}$$

where $\theta_i(A)$ is the angle between the i-th and the $i+1$-th residues for the conformation A.

One first finds the conformation A in the bank that is closest to α with the distance $D(\alpha, A)$. If $D(\alpha, A) < D_{\rm cut}$, α is considered to be similar to A. In this case, the conformation with lower energy among α and A is kept in the bank, and the other one is discarded. However, if $D(\alpha, A) > D_{\rm cut}$, α is regarded as distinct from all conformations in the bank. In this case, the conformation with the highest energy among the bank conformations is discarded, and the rest are kept in the bank. We perform this operation for all trial conformations.

For efficient sampling of the conformational space, the diversity of sampling must be maintained in the early stages; then, the emphasis is gradually shifted toward obtaining low energy conformations by slowly reducing $D_{\rm cut}$. In practice, $D_{\rm cut}$ is reduced by a fixed ratio after the bank update has been attempted by all the newly generated trial conformations in such a way that $D_{\rm cut}$ reaches $D_{\rm ave}/5$ after 10000 local minimizations. Then, seeds which have not been used as seeds yet are selected again from the bank conformations, to repeat the aforementioned procedure. The value of $D_{\rm cut}$ is kept constant after it reaches the final value.

It should be noted that in the early stages of the CSA, the seed conformations are continuously being replaced

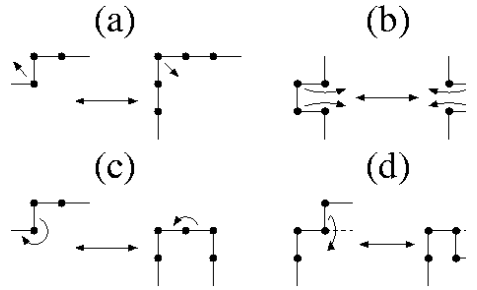

Fig. 1. Local moves used for the definition of local minimization consist of (a) three bead flips, (b) crankshaft movements, (c) rigid rotations, and (d) rigid reflections.

by low-energy local minima close to it. Therefore, when all of the bank conformations are used as seeds (one iteration completed), the procedure of updating the bank might have reached a deadlock. However, we give these conformations another chance by resetting them to be eligible for seeds, and repeat another iteration of search. After a preset number of iterations, we conclude that our procedure has reached a deadlock. When this happens, we enlarge the search space by adding more random conformations into the bank and repeat the whole procedure until the stopping criterion is met. In this work, after 3 iterations are completed, we increase the number of bank conformations by adding 50 randomly generated and minimized conformations into the bank (and also into the first bank), and reset $D_{\rm cut}$ to $D_{\rm ave}/2$. In the exhaustive searches I performed (See Sec.III), the algorithm was stopped when the number of bank conformations exceeded 1000.

It should be noted that although the local minimization is a crucial component of the CSA algorithm, the concept of the local minimum for the lattice model is not so straightforward as in the case of off-lattice systems. In contrast to the off-lattice model where the coordinates of the system are continuous and a local minimum conformation is defined as the one with vanishing gradient, care must be taken in order to suitably redefine the local minimum in the case of the lattice model. A natural definition of a local minimum conformation would be one whose energy does not get lowered by local moves only, so one must first define local moves. In this work, I use four types of local moves, three-bead flip, crank-shaft movement, partial rigid rotation, and partial reflection (Fig. 1). The first three movements were used by Chan and Dill [19] for the Monte-Carlo simulation of the HP model whereas the reflection is newly added in this work. The local minimization was performed by randomly selecting a residue and moving the residue with a method randomly selected from among the four methods with preset probabilities. The residues at both ends were specially treated, and they were given more chance for rigid rotation, which is also the method adopted by Chan and Dill [19]. A new conformation is accepted if its energy is lower than the original one, rejected if it has higher energy, and accepted with probability of 1/2 if its energy is the same as that of the original conformation. We repeat the local movements for $5N_{seq}$ times, regardless of success of failure to update, but repeat another $5N_{seq}$ local moves if the final energy value is positive, and so on, where N_{seq} is the chain length. The local minimization procedures stops unconditionally if the total number of attempted local moves exceeds $5N_{seq}^2$, although this case never happened in my computation. I find that a rigid rotation is most efficient for local minimization, so I used a probability of 0.7 for this move and 0.1 for all the others, which is an optimal value obtained from a trial run on shorter sequences.

Table 1. Summary of the 100 independent CSA searches. The lowest energy among each final 1000 bank conformations was either – 330, – 333, or – 336. NR denotes the number of corresponding runs, NC is the maximum number of distinct conformations one could obtain for the corresponding energy in the final bank, NF is the average number of local movements until the final minimum energy conformations were obtained, and the total degeneracy is the total number of distinct conformations with the given energy obtained from all 100 independent runs.

Minimum Energy	NR	NC	NF	total degeneracy
– 330	3	14	40467461.5	17
– 333	51	8	37802800.0	21
– 336	46	4	45273000.0	10

III. RESULTS

Using a single Intel Xeon CPU (2.4GHz), I performed several test runs for the sequence $HPH_2P_2H_2P_3H_2PHP_3H_2P_2H_2P_4HPH_2PHP_2$, and found that conformations with energies $E < E^0_{compact} = -320$ are located within two seconds of wall-clock time, within a few iterations with 50 bank conformations. I then performed exhaustive systematic runs for this sequence to look for the lowest energy conformation obtainable by the CSA. I performed 100 independent CSA runs with different values of the initial seed for the random number generator. Each run was terminated when the number of bank conformations reached 1000, and the search had reached a deadlock three times. It should be noted that typical CSA runs were terminated with bank conformations of 50 or 100 [13], so this is a very exhaustive search, indeed. However, all 100 runs took only 7 hours and 5 minutes, implying an average wall-clock time of only 4 minutes and 25 seconds for each run. The lowest energies found from the final bank conformations are $E = -3$ 30 for 3 runs, $E = -333$ for 51 runs, and $E = -336$ for the remaining 46 runs. The total number of distinct conformations for each of these energies, from all 100 runs, are 17, 21, and 10,

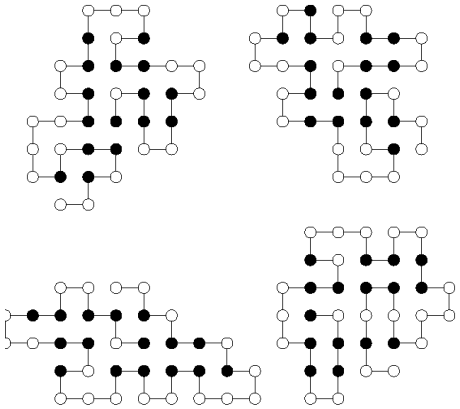

Fig. 2. Examples of conformations with $E = -336$ ($n_{HH} = 12$, $n_{HP} = 6$), for the sequence under study. The black and the white beads denote residues of type H and P, respectively.

respectively. These results are summarized in Table 1, along with the average number of local moves until all the lowest energy conformations in the final bank were obtained. The number of local moves is the analogue of the number of function calls in the off-lattice case, and I denoted it by NF in the table.

Although 10 distinct conformations with $E = -336$ are obtained, with $n_{HH} = 12$ and $n_{HP} = 6$, many of them are related to each other by local movements of several residues. Some of the distinct lowest energy conformations for $E = -336$ are displayed in Fig. 2. As we can see from the figure, the conformations have a more "protein-like" natural look compared to the conformation obtained by Li *et al.* [20], where the chain was artificially packed into a 6×6 square.

IV. CONCLUSION

In this work, I applied the CSA method for sampling the low-energy conformations of the 36-residue HP protein with sequence $HPH_2P_2H_2P_3H_2PHP_3H_2P_2H_2P_4$ HPH_2PHP_2. I could quickly obtain various conformations with energies lower than the one obtained by exact enumeration of *compact* conformations filling a 6×6 square, demonstrating that the CSA method is an efficient method for sampling low-energy conformations of a physical system. Although the CSA method can be easily adapted for parallel computation, the local minimization was performed very rapidly for the lattice system under study, and parallel implementation was not necessary for a chain length of 36.

The result of this work suggests that the conformation with the lowest energy is not necessarily the maximally compact conformation. Of course, one cannot be absolutely sure that the lowest energy conformations obtained in this work are truly the global minimum energy conformations of the sequence under study. In principle, the true global minimum energy and the corresponding conformations can be checked by exact enumeration of all possible conformations, which require tremendous amount of computer power and would be possible only by massively parallel computations. Such an endeavor is left for a future work.

ACKNOWLEDGMENTS

The author thanks Seung-Yeon Kim for useful discussions. This work was supported by the Soongsil University Research Fund.

REFERENCES

[1] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).
[2] J. Lee, H. A. Scheraga, and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).
[3] J. Lee, H. A. Scheraga, and S. Rackovsky, Biopolymers **46**, 103 (1998).
[4] J. Lee, H. A. Scheraga, and S. Rackovsky, Int. J. Quantum Chem. **75**, 255 (1999).
[5] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson, and H. A. Scheraga, Int. J. Quantum Chem. **77**, 90 (2000).
[6] J. Lee, A. Liwo, and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **96**, 2025 (1999).
[7] A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **96**, 5482 (1999).
[8] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, Proteins **Suppl 3**, 204 (1999).
[9] J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K. D. Gibson, R. J. Wawak, and H. A. Scheraga, Comput. Phys. Commun. **128**, 399 (2000).
[10] S-Y. Kim, S. J. Lee, and J. Lee, J. Chem. Phys. **119**, 10274 (2003).
[11] S-Y. Kim, S. J. Lee, and J. Lee, J. Korean Phys. Soc. **44**, 589 (2004).
[12] J. Lee, S-Y. Kim, K. Joo, I. Kim, and J. Lee, Proteins **56**, 704 (2004).
[13] J. Lee, I-H. Lee, and J. Lee, Phys. Rev. Lett. **91**, 080201 (2003).
[14] Y. Jung and Y. Kim, J. Korean Phys. Soc. **41**, 167 (2002).
[15] G-Y. Oh, J. Korean Phys. Soc. **42**, 714 (2003).
[16] K. A. Dill, Biochemistry **24**, 1501 (1985).
[17] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[18] C. J. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).
[19] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[20] H. Li, R. Helling, and C. Tang, Science **273**, 666 (1996).
[21] R. M'elin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).

[22] T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, J. Chem. Phys. **113**, 8329 (2000).
[23] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[24] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[25] C. Tang, Physica A **288**, 31 (2000).
[26] H-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, Phys. Rev. E **68**, 021113 (2003).
[27] H-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003).
[28] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[29] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1992).
[30] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[31] Y. Kim, J. Korean Phys. Soc. **28**, 539 (1995).
[32] J. Lee, J. Korean Phys. Soc. **44**, 617 (2004).