A Quasi-human Growth Algorithm for the Protein Folding Problem

Wen-Qi Huang, Zheng-Da Xiong
School of Computer Science and Technology
Huazhong University of Science and Technology,
Wuhan, China
xzda72@sina.com

Kun He, Ru-Chu Xu
School of Computer Science and Technology
Huazhong University of Science and Technology,
Wuhan, China

Abstract—**Inspired by the Chinese game "Weiqi", this paper proposes a Quasi-Human stochastic Growth Algorithm (QHGA) for the Simple Cubic Hydrophobic Polar (SC-HP) lattice model of the protein folding problem. Two concepts from the "Weiqi", the actual gain and the external potential, are used to evaluate different partial conformation. Experiments on several HP test sequences show the high performance of this new algorithm.**

Keywords-HP lattice model, heuristic, quasi-human, stochastic growth algorithm, external potential

I. INTRODUCTION

All manuscripts must be in English. These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts. Please follow them and if you have any questions, direct them to the production editor in charge of your proceedings at Conference Publishing Services (CPS): Phone +1 (714) 821-8380 or Fax +1 (714) 761-1784.

In molecular biology, a protein molecular is made up by an amino acid sequence. The biochemical character of a protein is directly decided by the folding form of the amino acid sequence in the space. Although people have been able to measure the amino acid sequence for many kinds of proteins, it is difficult to measure the folding conformation by experiments. The main reason is that the conformation achieved by the measurement is not the live conformation due to the restrictions on the environment and experimental tools. Therefore, people resort to the method of predicting protein conformation by establishing a corresponding mathematical model and then by solving the model on computers.

This paper addresses on the protein folding problem. Given the energy formula of the protein conformation, the protein folding problem is to find a conformation with the minimum energy, which actually corresponds to the realistic conformation according to the minimum energy principle.

There are many mathematical models for the protein folding problem, which can be divided into the following three categories: the lattice model, the off-lattice model and the all atoms model. Certainly, the off-lattice model is closer to the realistic conformation. However, the lattice model is more commonly used by the researchers because: (i) the mathematical model of the lattice model is simpler and it is then easier to be solved; (ii) the lattice model facilitates the observation on the geometry structure of the proteins, and the geometry overview of the lattice model is generally similar to that of the off-lattice model. Therefore, the lattice model not only reflects the shape of the protein folding but also be a powerful tool to predict the protein conformations.

II. THE SC-HP LATTICE MODEL

As a classic lattice model, the Simple Cubic Hydrophobic Polar (SC-HP) lattice model has been widely recognized in the academic community. The specific description of the SC-HP is as follows:

Given an amino acid chain with length N, which can be coded as a string compose by two types of characters: "H" and "P"; Here "H" denotes a hydrophobic amino acid, and "P" indicates a hydrophilic amino acid. For ease of description, we use black balls and white balls with unit diameter to represent "H" and "P" characters, respectively. So the whole chain of the amino acids can be considered as a chain lined up by black balls and white balls.

The problem is how to place the balls on lattice points in a two- or three-dimensional space so that they satisfy the following constraints:

1) The adjacent balls on the chain are still adjacent;

2) Different balls should be placed on different lattice points.

A feasible placement is called a conformation of the protein chain. And we define an energy value E for each conformation: the energy e_{ij} is -1 if two black balls i, j are adjacent, or else the e_{ij} is 0, and the energy of the chain is the total energy of the ball pairs, as shown in formula (1).

$$E = \sum_{1 \le i < i+1 < j \le N} e_{ij} \tag{1}$$

The objective is to find a conformation with the lowest energy.

If we put one ball at each time,， then the problem can be regarded as a "self-avoiding walk" problem. In many disciplines, it has important theoretical and practical significance. For a protein chain with length N, the total number of conformations is about $R(N) = A\mu^N N^\gamma$. Here A is a constant, $\mu \approx 2.63$ and $\gamma \approx 0.333$ For the SC-HP lattice model, the solution space is very large, but for many instances, the number of the lowest energy conformation is

much smaller. And the problem of solving the SC-HP lattice model has been proved to be NP-hard [1].

Algorithms for solving the SC-HP lattice model can be divided into complete algorithm and heuristic optimization algorithm. Yue and Dill [2] proposed a complete algorithm for the SC-HP lattice model, which could solve instances with *N* up to 88. For instances with *N* larger than 88, the running time is too long to be accepted. So researchers begin to design non-complete algorithms to find satisfactory solutions within reasonable times. The major heuristic algorithms are: genetic algorithm proposed by Ron Unger and John Moult[2], the H kernel constructing algorithm by Beutler and Dill[3], the importance sampling method by Zhang and Liu[4], PERM by Grassberger et al.[5], simulated annealing method by Chou *et al*[6], and the Evolutionary Monte Carlo by Liang[7]. In these algorithms, the PERM algorithm and its improved version have better efficiency and quality than the others.

III. ALGORITHM DESCRIPTION

The stochastic growth algorithm places the balls one by one on different lattice points. Suppose *n*-1 balls have been put down, the algorithm evaluates each candidate position for the *n*th ball and selects a location based on some probabilistic rules. And the growing procedure keeps working until all the balls have been placed, and the conformation of the entire chain is completed. The key points for a growth algorithm are how to design the evaluation strategy and how to define the probability rule.

Analogy or comparison with Chinese game "Weiqi", we propose a quasi-human approach for the SC-HP lattice model. In Chinese game "Weiqi", the player's goal is to get as much space. In the course of chess, not every step can get more space; however, the current pattern will seriously affect the opportunity to get space in the future. The space acquired currently is called the actual gain; and the space expected to acquire in the future is called the external potential. The actual gain could be calculated by the "Weiqi" rules; the external potential is an expectation of the influence of the current pattern to the future course. There is no unified formula for the external potential, and it can only be estimated by experienced player.

We assume that the balls of a chain are numbered from *0* to *N*-1, and the first ball is placed in the origin. Corresponding to the SC-HP lattice model, we define the actual gain and the external potential as follows:

Definition 1. Partial *n* conformation. In the process of placing the ball, the current situation is defined as a partial *n* conformation if ball 0 to *n*-1 have been placed and ball *n* is going to be placed on a lattice point .

Definition 2. Actual gain E_n. The actual gain of a partial *n* conformation is the number of neighbor black pairs.

Definition 3. Weight W_n .The weight is as Equation (2) shows, where *T* is a temperature constant.

$$W_n = e^{-\frac{E_n}{T}} \tag{2}$$

Definition 4. Potential point. For a partial *n* conformation, each free lattice point around a black ball is defined as a potential point.

Definition 5. External potential. For a partial *n* conformation, consider each black ball among the *n*+1th to *N*th balls of the chain, and the sum of the total number of the black balls that can be placed on each potential point, is defined as the external potential of this partial *n* conformation.

Then, how can we deduce whether a black ball can be placed on a potential point? Taking into account the parity of the lattice point, an odd-numbered ball must be placed on an odd lattice, and an even-numbered ball must be placed on an even lattice. For a partial *n* conformation and a black ball that haven't be placed, if the number is *i*, the distance between the black ball and ball *n*-1 is less than or equal to *i-n+1*. If we want the black ball contact with the partial *n* conformation, it can only occupies the potential point which distance less than or equal to *i-n+1* with the last ball of the partial *n* conformation. For these reasons, we will take the formula of outer potential as:

$$S_n = \sum_{i \geq n,\ \text{ith ball is black}} \text{number of potential point} \tag{3}$$

The main idea of the Quasi-Human stochastic Growth Algorithm (QHGA) is as follows.

1) During the operation of the algorithm, for *n* that *1<n<N-1*, compute C_n , the number of partial *n* conformation that have been generated by the algorithm, the average weight $\overline{W}_n$, and the average outer potential $\overline{S}_n$.
2) Randomly generate an initial complete conformation, initialize $C_n, \overline{W}_n$ and $\overline{S}_n$.
3) For a partial *n* conformation, consider the free lattice point near the last ball, placed the *n*th ball on these positions, compute the incremental energy $\Delta E_{n+1}^{(i)}$, weight $W_{n+1}^{(i)}$, and external potential $S_{n+1}^{(i)}$.
4) Given the upper and lower threshold coefficient $c_>, c_<$, calculate the upper and lower threshold: $W^> = c_> \overline{S}_{n+1} \overline{W}_{n+1}, W^< = c_< \overline{S}_{n+1} \overline{W}_{n+1}$
   Compare $S_{n+1}^{(i)}$, $W_{n+1}^{(i)}$ with $W^>, W^<$, and there are three kinds of possibilities and the corresponding strategies:
   a) T1: $S_{n+1}^{(i)} W_{n+1}^{(i)} > W^>$: For each conformation in T1, go into new conformation;
   b) T2: $S_{n+1}^{(i)} W_{n+1}^{(i)} \in [W^<, W^>]$: proportional to the corresponding data, select a conformation according to probability, go into new conformation;
   c) T3: $S_{n+1}^{(i)} W_{n+1}^{(i)} < W^>$: with probability 0.5 to decide whether growth, then proportional to the corresponding data, select a conformation

according to probability, go into new conformation.

The detailed pseudo-code of the QHGA is as follows:

Algorithm: QHGA(n)

1. If n=N-1
2. Place the last ball.
3. Return.
4. end if
5. $W^{>} = c_{>}\overline{S}_{n+1}\overline{W}_{n+1}$
6. $W^{<} = c_{<}\overline{S}_{n+1}\overline{W}_{n+1}$
7. for each free lattice point near the last ball，
8. $W_0 = W$
9. if have least one free dot near it
10. Compute S
11. $W = W_0 e^{-\frac{\Delta E}{T}}$
12. renew C_n, $\overline{W}_n$ and $\overline{S}_n$
13. if $SW > W^{>}$
14. Insert SW into Set T_1
15. else if $SW < W^{<}$
16. Insert SW into Set T_3
17. else
18. Insert SW into Set T_2
19. end if
20. end if
21. end for
22. for each in T_1
23. QHGA(n+1)
24. end for

Proportional to the data in T_2, select a conformation

25. according to probability，QHGA(n+1)
26. generate a random number $r \in [0,1)$
27. if $r>0.5$
28. Proportional to the data in T_3, select a conformation according to probability，QHGA(n+1)
29. end if

V. CONCLUSION

Based on the definitions of the actual gain and the external potential, we proposed a straightforward and efficient heuristic for the protein folding problem with HP lattice model. The proposed heuristic works well on several classical instances. QHGA achieved optimal conformations on each of the six instances, which to our knowledge is the same as the current best result.

Our future work is how to improve the local search strategy such that the heuristic could get good results on three-dimensional instances.

IV. RESULT AND ANALYSIS

We tested QHGA on the sequences for HP lattice model that were present by Unger and Moult[2]. Table 1 shows the HP sequences and the energies that QHGA obtained. QHGA can find the lowest energy conformation for each sequence within short times. And the conformations obtained by QHGA are as Figure 1 shows.

As a non-complete algorithm, we cannot guarantee that the solution obtained is the optimal solution. However, for HP lattice model, we can analyze whether they have the lowest energy. In the conformation (c)-(e), the “H”s formed a square layout. No doubt, they must have the lowest energy. To get this kind of conformation, the HP sequences should meet the following conditions:

1) No “P” in the sequence has two adjacent “H”;

2) Up to four “H” in the sequence has two adjacent “P”.

If the HP sequence does not meet above conditions, the lowest energy can be estimated by analyzing the energy loss.

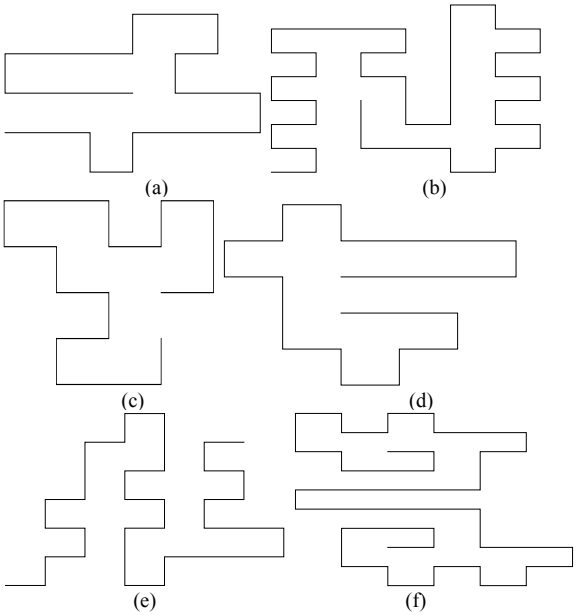

Figure 1. The conformations of the sequences

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (Grant No. 61070235).

REFERENCES

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The

sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Electronic Publication: Digital Object Identifiers (DOIs):

Article in a journal:

[9] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

Article in a conference proceedings:

[10] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

TABLE I. TABLE I. HP SEQUENCES FOR MODEL PROTEINS

No	Length	Sequence	E
1	20	HPHPPHHPHPPHPHHPPHPH	-9
2	24	HHPPHPPHPPHPPHPPHPPHPPHH	-9
3	25	PPHPPHHPPPPHHPPPPHHPPPPHH	-8
4	36	PPPHHPPHHPPPPPHHHHHHHHPPHHPPPPHHPPHPP	-14
5	48	PPHPPHHPPHHPPPPPHHHHHHHHHHPPPPPPHHPPHHPPHPPHHHHHH	-23
6	50	HHPHPHPHPHHHHPHPPPHPPPHPPPPHPPPHPPPHPHHHHPHPHPHPHH	-21