http://www.pjbs.org

PJBS

ISSN 1028-8880

Pakistan Journal of Biological Sciences

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Pakistan Journal of Biological Sciences 11 (14): 1815-1819, 2008
ISSN 1028-8880

A New Algorithm for 2D Hydrophobic-Polar Model: An Algorithm Based on Hydrophobic Core in Square Lattice

Wang Wei and Tang Yanlin
College of Science, Guizhou University, Guiyang, Guizhou Province, China

Abstract: This study was engage in a new algorithm which was used to solve the problem of protein folding. The conformation of hydrophobic core of protein was key factor of structure of protein. So, in our algorithm, we set a hydrophobic core which was restricted by new aggregate. Then, the hydrophilic residues between two hydrophobic residues were ranged, the optimal conformation was gained if all residues were not overlap and continuous. The algorithm in this study can be prevented effecttively falls into partially smallest energy.

Key words: 2D HP model, protein structure, new aggregate, hydrophobic core

INTRODUCTION

Determining the functionality of a protein molecule from amino acid sequence remains a central problem in computational biology, molecular biology, biochemistry and physics. Even the experimental determination of these conformations is often difficult and time consuming. It is common practice to use models that simplify the search space of possible conformation. These models try to generally reflect different global characteristics of protein structures. In the Hydrophobic-Polar (HP) model (Dill, 1985) the primary amino acid sequence of a protein (which can be represented as a string over twenty-letter alphabet) is abstracted to a sequence of hydrophobic (H) and polar (P) residues that is represented as a string over the letter H and P. In the model, the amino acid sequence is abstracted to a binary sequence of monomers that are either hydrophobic or polar. The structure is a chain whose monomers are on the vertices of a three dimensional cubic lattice. The free energy of a conformation is defined as the negative number of non-consecutive hydrophobic-hydrophobic contacts. A contact is defined as two nonconsecutive monomers in the chain.

Occupying adjacent sites in the lattice. In spite of its apparent simplicity, finding optimal structures of the HP model on a cubic lattice is NP-complete problem (Berger, 1998).

In this research, we induce a new algorithm in which hydrophobic core was formed based an aggregate. The aggregate was a series of possible coordinate of next hydrophobic residues in sequence, the amount of conformation of hydrophobic core are largely reduced when the new aggregate was applied in process of searching optimal conformation.

The 2D HP (2-dimension hydrophobic-polar) model was introduced (Dill, 1985) and it is the most widely studied discrete model for protein folding in the recent literature. It models the concept that the major contribution to the free energy of the native conformation of a protein is due to interactions among hydrophobic residues. They tend to form a core in the protein structure while surrounded by hydrophilic residues that interface to the environment.

In the HP model, the 20 standard amino acids are divided into two types, according to its affinity to water: hydrophobic (H for non-polar) or hydrophilic (P for polar). As it is a lattice model, the amino acid chain is embedded in a 2- or 3-dimensional square lattice and the movements of the chain are restricted to angles of 90°. In a legal conformation, the adjacent residues in the sequence must be adjacent in the lattice and each lattice point can be occupied by only one residues.

The free energy of a conformation is inversely proportional to the number of hydrophobic residues occupy adjacent grid points in the lattice but are not consecutive in the sequence. Each such interaction contributes with -1 to the energy value.

Since, the processes involved in the folding of proteins are very complex and only partially understood simplified models like Dill's Hydrophobic-Polar (HP) model have become one of the major tools for studying proteins. The HP model is based on the observation that hydrophobicity of amino acids is the main force for protein folding and conformation of small globular protein. In the HP model, the primary amino-acid sequence of a protein is abstracted to a sequence of hydrophobic (H) and polar (P) residues, HP model considered here, a 2-dimensional square lattice is used.

Corresponding Author: Wang Wei, College of Science, Guizhou University, Guiyang, Guizhou Province, 550025, China

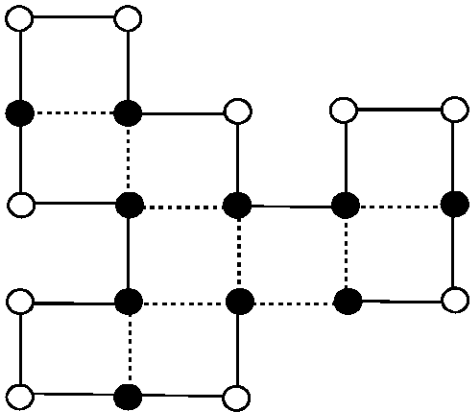

Fig. 1: A sample protein conformation in 2D HP model

An example for a protein conformation under the 2D HP model is shown in Fig. 1.

One of the most common approaches to protein structure prediction is based on the thermodynamic hypothesis which states that the native state of the protein is the one with the lowest Gibbs free energy. In the HP model, based on the biological motivation given above, the energy of a conformation is defined as a number of topological contacts between hydrophobic amino-acid that are not neighbors in the given sequence.

A number of wellknown heuristic optimization methods have been applied to the 2D HP protein folding problem, including evolutionary algorithms (Eas) (Shmygelska, 2007; Hoque *et al*., 2006; Konig and Dandekar, 1999; Krasnoger *et al*., 1999; Unger and Moult, 1993; Verma *et al*., 2007) and Monte Carlo (MC) algorithms (Bastolla *et al*., 1998; Chikenji *et al*., 1999; Chris *et al*., 2007; Hsu *et al*., 2002; Liang and Wong, 2001) and Simulated Annealing (SA) algorithm (Li, 2007; Liu and Tao, 2006). The latter have been found to be particular robust and effective for finding high-quality solutions to the 2D HP protein folding problem.

An early application of EAs to protein structure prediction was presented by Unger and Moult. They presented a nonstandard EA incorporating characteristic of Monte Carlo methods, which was able to find high-quality conformations for a set of protein sequences of length up to 64 amino-acids. Unfortunately, it is not clear how long their algorithm ran to achieve these results.

The Core-directed Chain Growth (CG) (Zhang and Liu, 2002) method of Beutler *et al.* (1996) approximates the hydrophobic core of the protein with a square.

Finally, Ant Colony Optimization (ACO) (Shmygelska *et al*., 2002; Shmygelska and Hoos, 2005; Dorigo *et al*., 1999; Chu *et al*., 2005) is a population-based approach to solving combinatorial optimization problems that is inspired by the foraging behavior of ant colonies. The fundamental approach underlying ACO is an iterative process in which a population of simple agents (ant) repeatedly constructs candidate solutions. This construction process is probabilistically guided by heuristic information on the given problem instance as well as by a shared memory containing experience gathered by the ants in previous iterations (pheromone trails). Following the seminal study by Dorigo *et al.* (1991), ACO algorithms have been successfully applied to a broad range of hard combinatorial problem.

These algorithms to 2D HP model, such as ACO (Ant Colony Optimization) and EA (Evolution Algorithm) are expensive in terms of computation and time. A new algorithm will decrease the cost in this study.

Currently, none of these algorithms appears to completely dominate the others in terms of solution quality and run time.

MATERIALS AND METHODS

The shape of the hydrophobic core is the key factor which influence the forming of protein conformation because the protein conformation is determined after formed hydrophobic core, so an algorithm has been constructed based on the principle that the H residues have the tendency to forming the impact hydrophobic core and lowest energy in this study.

Firstly, all hydrophobic residues in the sequence are arranged to an optimal core according to some principle. Secondly, the hydrophilic residues are distributed between hydrophobic residues.

Definition of shaping of hydrophobic core in the algorithm

Aggregate of coordinate of hydrophobic residues: If s_i is hydrophobic residues and its coordinate is (x_i,y_i). s_j is the next adjacent hydrophobic residues in the sequence. The coordinate of s_j should satisfy follow condition because of the restriction of hydrophilic residues between s_i and s_j.

$$\left|x_i - x_j\right| + \left|y_i - y_j\right| = \left|i - j\right| - 2 \times n > 0 \qquad (n = 0, 1, 2...)$$

The aggregate of x and y which satisfy above restriction are the potential coordinates of s_j (Fig. 2). As shown in the Fig. 2: s_i locates in center and possible s_j are presented by dots. By this rule, the number of potential conformation of hydrophobic core rapidly reduces.

Path in adjacent hydrophobic residues: The path from can be found after the coordinate of them were fixed. A new method which is used to find all path from s_i and s_j (s_i and

s_j are hydrophobic residues and j-i = 1) is introduced in the follow. S_k(i<k<=j) is a polar residue between s_i and s_j, (x_k, y_k) is the coordinate of S_k, The coordinate of S_k satisfy the follow relation.

$$|x_k - x_j| + |y_k - y_j| = |j - k|$$
$$|x_k - x_{k-1}| = 1 \qquad |y_k - y_{k-1}| = 1$$

All S_k between s_i and s_j compose the possible path from s_i to s_j. The coordinate of s_i and s_j is (0, 0) and (1, 2), all paths between s_i and s_j can be found based on above relation (Fig. 3).

Deducing the optimal conformation of core: In this research, we will introduce how to deduce the optimal hydrophobic core.

Start from the hydrophobic residue h_1, it consists of n_1 (n_1 = 1) conformation of core, based on the above rule, h_2 which is the next adjacent hydrophobic residue of h_1 in the sequence has n_2 possible positions in the lattice. So, there are n_2 newest conformation of core, in the same way, h_3 is the next adjacent hydrophobic residue of h_2, it have n_3 possible positions. There are $n_1 \times n_2 \times n_3$ newest conformations of core. If there are j hydrophobic residues in the sequence then there are $n_1 \times n_2 \times \ldots \times n_i \times \ldots \times n_{j-1} \times n_j$ conformation of core.

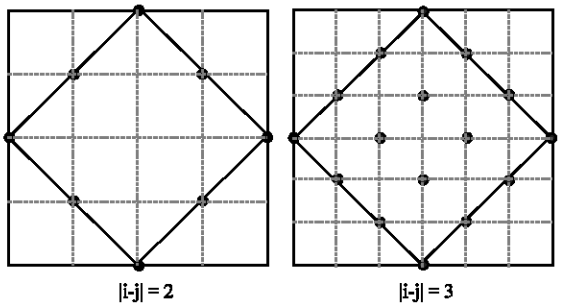

Fig. 2: The aggregate of coordinate of s_i

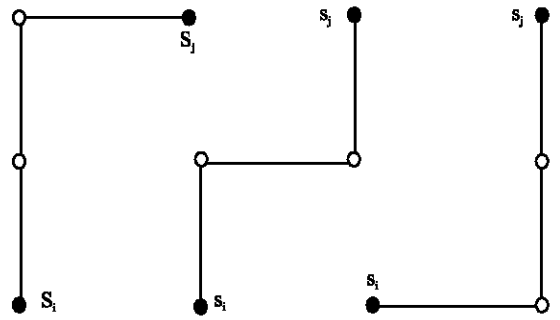

Fig. 3: All paths between s_i and s_j , each of them different from each other

The amount, $n_1 \times n_2 \times \ldots \times n_i \times \ldots \times n_{j-1} \times n_j$, of the conformation of core is very large despite the new aggregate is adopt, so follow methods are used to reduce the amount of newest conformation of core in step i (i = 1, 2 ,...j-1, j).

Eliminating the similar and invalid conformation: Similar conformation is a kind of conformation which is based on the x-axis or y-axis or origin symmetric, this kind of conformations is redundancy; invalid conformation is the kind of conformation in which coordinate of residues is overlap.

Reduce the amount of conformation of core via energy: The energy of new conformation of core is calculated after every step. In case of that the lowest energy of all new conformation is E_{min}, the conformations which energy locates between E_{min} and αE_{min} are reserved. Conformation searching can be prevented effectively falls into partially smallest if α is proper.

Restriction of polar residues: The distribution of polar residues between s_i and s_j is a key factor to judge whether the new conformation of core is valid. When all paths between s_1 and s_i were found after s_i is fixed. A conformation of core is valid if there is one path which is not overlapping or crosses at least.

RESULTS AND DISCUSSION

We used the benchmark sequence HPHPPHHPHPP-HPHHPPHPH to test the algorithms. The results are shown as Fig. 4, dots are hydrophobic residues and circles are polar residues.

There are two optimal conformations after calculation because the benchmark sequence is palindrome, but these two conformations are similar.

Some benchmark sequences, as shown in Table 1, are used to test algorithms in this study and performance on finding the lowest energy conformations compare with other four algorithms are shown in Table 2.

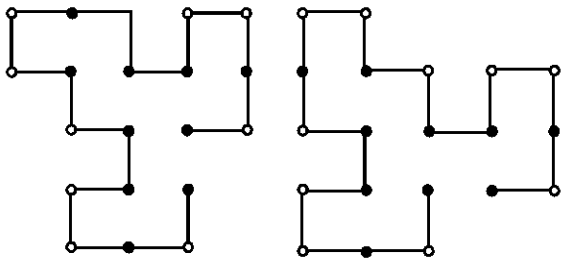

Fig. 4: The optimal conformation of benchmark

Table 1: The 10 benchmark sequences for algorithm evaluation

Length	Sequence
20	HPHPPHHPHPPHPHHPPHPH
24	HHPPHPPHPPHPPHPPHPPHPPHH
25	PPHPPHHPPPPHHPPPPHHPPPPHH
36	PPPHHPPHHPPPPPPHHHHHHHHPPHHPPPPHHPPHPP
48	PPHPPHHPPHHPPPPPHHHHHHHHHHPPPPPPHHPP HHPPHPPHHHHH

Table 2: Performance comparison of the four algorithms*

Length	Optimal	MC	GA	ACO
20	-9	-9	-9	-9
24	-9	-9	-9	-9
25	-8	-7	-8	-8
36	-14	-12	-14	-14
48	-23	-20	-22	-23

*Performance comparison on finding the lowest energy conformations of the four algorithms, including Monte Carlo (MC), Genetic Algorithm (GA), Ant Colony Optimization (ACO) (Guo *et al*., 2006)

The strong point of our method which used to find the optimal conformation of protein in 2D HP square lattice are list follows:

- During the process of inducing optimal conformation of core, the amount of experimental conformation of core is largely reduced because we adopt the new aggregate to abandon the invalid core conformation.
- The process of searching the optimal conformation can be prevented effectively falls into partially smallest energy when the coefficient α is proper.

But, the above algorithms could not be applied in the sequence with more than one hydrophobic core because we can not judge how many hydrophobic cores in the sequence. That mean, in the future work, it is important to confirm every hydrophobic core in the sequence. Furthermore, the application of this method in the 3D lattice should be researched.

REFERENCES

Bastolla, U., H. Frauenkron and E. Gestner, 1998. Testing a new Monte Carlo algorithm for the protein folding problem. J. Proteins Struct. Funct. Genet., 32: 52-66.

Berger, B.T., 1998. Leighton, protein folding in the hydrophobic-hydrophilic (HP) model is np-complete. J. Comput. Biol., 5: 27-40.

Beutler, T.C. and K.A. Dill, 1996. A fast conformational search strategy for finding low energy structures of model proteins. Protein Sci., 5: 2037-2043.

Chikenji, G., M. Kikuchi and Y. Iba, 1999. Multi-self-overlap ensemble for protein folding: Ground state search and thermodynamics. J. Phys. Rev. Let., 83: 1886-1889.

Chris, T., A. Shmygelska and H.H. Hoos, 2007. A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinform., 8: 342-362.

Chu, D., M. Till and A. Zomaya, 2005. Parallel ant colony optimization for 3D protein structure prediction using the HP lattice model. IEEE Comput. Soc., 07: 193-200.

Dill, K.A., 1985. Theory for the folding and stability of globular proteins. J. Biochem., 24: 1501-1509.

Dorigo, M., G. Di Caro and L.M. Gambardella *et al*., 1999. Ant algorithms for discrete optimization. J. Artificial Life, 5: 137-172.

Guo, Y.Z., E.M. Feng and Y. Wang, 2006. Exploration of two-dimensional hydrophobic-polar lattice model by combining local search with elastic net algorithm. J. Chem. Phys., 125: 102-108.

Hoque, T., M. Chetty and L.S. Dooley, 2006. A guided genetic algorithm for protein folding prediction using 3D hydrophobic-hydrophilic model. J. Evolut. Comput., 10.1109/CEC.2006.1688597 http://ieeexplore.ieee.org/Xplore/login.jsp? url=/iel5/11108/35623/01688597.pdf?tp=&arnumber=1688597&isnumber=35623.

Hsu, H.P., V. Mehra and W. Nadler, 2002. Growth algorithm for lattice heteropolymers at low temperatures. J. Chem. Phys., 118: 444-451.

Konig, R. and T. Dandekar, 1999. Improving genetic algorithms for protein folding simulations by systematic crossover. J. BioSyst., 50: 17-25.

Krasnoger, N., W.E. Hart and J. Smith, 1999. Protein structure prediction with evolutionary algorithms. Proceedings of the Genetic and Evolutionary Computation Conference, pp: 1596-1601

Li, X., 2007. Protein folding based on simulated annealing algorithm. Natural Comput., 4: 256-259.

Liang, F. and W.H. Wong, 2001. Evolutionary Monte Carlo for protein folding simulation. J. Chem. Phys., 115: 3374-3380.

Liu, Y.L. and L. Tao, 2006. An improved parallel simulated annealing algorithm used for protein structure prediction. Mach. Learn. Cybernet., 10.1109/ICMLC.2006.258721 http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/4028021/4028022/04028455.pdf?temp=x.

Shmygelska, A., R. Aguirrc-Hernandez and H.H. Hoos, 2002. An ant colony optimization algorithm for the 2D HP protein folding problem. Lecture Notes Comput. Sci., 2463: 40-53.

Shmygelska, A. and H.H. Hoos, 2003. An improved ant colony optimization algorithm for the 2D HP protein folding problem. Adv. Artificial Intell., 2671: 400-417.

Shmygelska, A. and H.H. Hoos, 2005. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics, 6: 30-52.

Shmygelska, A., 2007. An external optimization search method for the protein folding problem: The go-model example. Genetic and Evolutionary Computation Conference. July 7, ACM , New York, NY, USA pp: 2572-2579.

Unger, R. and J. Moult, 1993. Genetic algorithms for protein folding simulations. J. Mol. Boil., 231: 75-81.

Verma, A., S.M. Gopal, J.S. Oh, K.H. Lee and W. Wenzel, 2007. All-atom de novo protein folding with a scalable evolutionary algorithm. J. Comput. Chem., 28: 2552-2560.

Zhang, J.L. and J.S. Liu, 2002. A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model. J. Chem. Phys., 117: 3492-3498.