A Global Optimization Algorithm for Protein Folds Prediction in 3D Space

Xiaoguang Liu, Gang Wang, Jing Liu

Department of Computer Science, Nankai University, Tianjin, 300071, China
{liuxg74, wgzwp}@hotmail.com,jingliu@nankai.edu.cn

Abstract. Protein folds prediction is one of the most important problems in computational biology. In previous works, local optimization algorithms were used mostly. A new global optimization algorithm is presented in this paper. Compared with previous works, our algorithm obtains much lower energy states in all examples with a lower complexity.

1 Introduction

Predicting the structure of proteins, given their sequence of amino acid, is one of the core problems in computational biology. With the rapid advances in DNA analysis, the number of known amino acid sequences has increased enormously. However, the progress in understanding their 3D structure and their functions has lagged behind owing to the difficulty of solving the folding problem.

Since the problem is too difficult to be approached with fully realistic potentials, many researchers have studied it in various degrees of simplifications. By the simplifications, protein fold prediction is converted to a combinatorial optimization problem. Its main target is to design algorithms which can find the lowest energy states of the amino acid sequences in three-dimensional space. The most popular model used in related works is HP model [1,2] which only consider two types of monomers, H (hydrophobic) and P (polar) ones. Hydrophobic monomers tend to avoid water which can only attract mutually by themselves. All the monomers are connected like a chain. There are repulsive or attractive interactions among neighboring monometers. The energies are defined as $\varepsilon_{HH}= -1$, and $\varepsilon_{HP}=\varepsilon_{PP}= 0$.

Many computational strategies have been used to analyze these problems, such as Monte Carlo simulations[3], chain growth algorithms[4], genetic algorithms[5], PERM and improved PERM[6], etc. Most models mentioned above are discrete. It's possible that some potential solutions are missed by the discrete models in 3D space. In reference 7, Huang devised a continuous model for 3D protein structure prediction. But the results from reference 7 had some errors owing to the defects in algorithm . Following the idea of Huang's model, we present a continuous optimization algorithm in the paper.

2 The Algorithm

In HP model, all amino acid monomers are connected and form a n-monomer chain. It's easy to understand that every monomer can be considered as a rigid ball. In order to present more succinctly, hydrophobic monomers are denoted as H balls and polar monomers are denoted as P balls in the following sections.

If the number of the balls in the chain is n and the radius of every ball is one, then protein folds prediction can be transformed into discovering the fit positions of these balls in 3D Euclidean space. It requires all the neighboring balls connected each other are tangent and all H balls are close as much as possible.

More precisely, the algorithm wants to obtain a n-dimensional position vector P$(P_1,P_2\ldots P_n)$ in 3D Euclidean space satisfying the following three conditions:

$$d_{i,j} \geq 2 \tag{1}$$

Where $d_{i,j}$ is the distance between position P_i and P_j ,($i \neq j, i, j = 1,2\ldots n$)

$$d_{i,i+1} = 2 \qquad i = 1,2,\ldots,n-1 \tag{2}$$

$$\text{Minimized } E \text{ where } E = \sum_{i=1}^{n}\sum_{j>i}^{n} \varepsilon / d_{i,j} \text{ , } \varepsilon_{HH} = -1 \text{ , } \varepsilon_{HP} = \varepsilon_{PP} = 0 \tag{3}$$

E is the gravitational energy of all balls.

We can consider all the balls in the chain are connected by a spring. Thus there are three types of forces in the n-ball chain, the pull forces of spring between the adjacent balls, the repulsion forces between two embedded balls and the gravitational forces between two H balls (since $\varepsilon_{HP}=\varepsilon_{PP}= 0$).

At any time, the external force that each ball received is the sum of forces that all the other balls in the same chain imposing on it. From the initial state, all the balls in the chain will be moved continuously driven by the external force. The n-ball system keeps moving until all the forces reach the equilibrium. During the process, the pull and repulsion forces drive the system to meet the requirements of equation (1) and (2), the gravitational forces among all H balls pull them together as close as possible. In the equilibrium state, the position vector of all the balls *P(P1,P2…Pn)* represent a best fit approximation to 3D protein structure prediction. The value of P can be determined according to equations (1) ,(2) and (3).

Considering the pull forces that ball i put on ball j ,

$$\overrightarrow{F_{pij}} = \begin{cases} \dfrac{k_p \times (d_{ij} - 2) \times (\vec{r_i} - \vec{r_j})}{d_{ij}} & if\ d_{ij} > 0 \\ 0, \qquad if\ d_{ij} = 0 & \end{cases} \tag{4}$$

Where $\vec{r_i}$ ($\vec{r_j}$) is the vector pointing to the position of ball i (j) from grid origin, d_{ij} is the distance between ball i and j , and k_p is the elastic coefficient of the spring

in the chain. It's easy to understand that there is only one pull force to the first and the last ball in the n-ball chain. To the others, the pull forces will be produced by the previous and the following balls. Obviously, the pull forces will be changed into push forces if $0 < d_{ij} < 2$ according to equation (4).

To the repulsion forces between ball *i* and *j*,

$$\overrightarrow{F_{rij}} = \begin{cases} \dfrac{k_r \times (2 - d_{ij}) \times (\vec{r_i} - \vec{r_j})}{d_{ij}}, d_{ij} < 2 \\ 0, \qquad d_{ij} \geq 2 \end{cases} \tag{5}$$

Where k_r is the repulsive coefficient of the balls in the case that two balls are embedded each other. To the gravitational forces between two H balls,

$$\overrightarrow{F_{gij}} = \begin{cases} k_g \times (\vec{r_i} - \vec{r_j}) \big/ d_{ij}^3, & if\ d_{ij} \geq 2 \\ 0, & if\ d_{ij} < 2 \end{cases} \tag{6}$$

According to equations (4), (5) and (6), the force $\overrightarrow{F_i}$, which exerted to ball *i* at any time, is the composition of the forces giving by the other balls in the chain.

$$\overrightarrow{F_i} = \sum_{j=i-1,i+1} \overrightarrow{F_{pji}} + \sum_{j=1, j\neq i}^{n} \overrightarrow{F_{rji}} + \sum_{\substack{j=1, j\neq i, \\ i, j \in H-balls}}^{n} \overrightarrow{F_{gji}} \tag{7}$$

Our algorithm can be described as following,

Initially, all the balls in the chain are distributed orderly on the surface of a virtual sphere in the 3D Euclidean space as even as possible. Therefore every ball will be coequal in the initial state. In the next period, each ball is moved in a small distance by the composition of external forces. This process repeats continuously until the n-ball system reaches the equilibrium. The positions of all the balls in the equilibrium state should be the solution to 3D protein structure prediction.

The pseudocode of the algorithm

```
Initialization.
for (t=0;t<t_MAX;t++)
  for (i=0;i<n;i++)
  {
  F_i =Compute_Force(i);//Computing the Force to ball i;
   r_i^(t+1) = r_i^t + λ × F_i ;
   };
```

Where $t_{\max}$ the upper bound of the periods, and λ is is the movement coefficient in the iterative equation.

3 Experimental Results

Four 3D HP sequences with length N equals to 58,103,124 and 136 respectively were described in reference 6 as models of actual proteins. Using our algorithm, we re-calculated the four sequences and found much lower energy states for all these sequences.

In our experiments, we set the maximal number of periods $t_{\max} = 3.0 \times 10^9$, set the movement coefficient $\lambda = 2 \times 10^{-7}$.

To the other coefficients used in the algorithm,

$$K_P = \begin{cases} 500 + 0.5 \times t, n = 58 \\ 800 + 0.5 \times t, n = 103{,}124 \\ 1000 + 0.5 \times t, n = 136 \end{cases} \tag{8}$$

$$K_r = \begin{cases} 500 + 0.6 \times t, n = 58 \\ 800 + 0.6 \times t, n = 103{,}124 \\ 1000 + 0.6 \times t, n = 136 \end{cases} \tag{9}$$

$$K_g = \begin{cases} 30, & n = 58 \\ 40, & n = 103{,}124 \\ 50, & n = 136 \end{cases} \tag{10}$$

Where *t* is the current period.

Furthermore, the precision is set to 1×10^{-7} in order to determine whether the two H balls are tangent or not. The details of the results are shown in table 1.

Among all previous work, the growth algorithm in reference 6 provided the best results. Compared with the results from reference 6, our algorithm can find much lower energy states for all the four HP sequences. The big gap is mainly caused by the differences in algorithm.

The growth algorithm used in reference 6 is a depth-first *implementation* of the "go-with-the-winners" strategy. Actually, it can be regarded as a special type of greedy algorithm. It always takes the local optimal solution while resolving the problem. As we known, a winner of a battle may not be a winner of the whole war. The solutions found by the growth algorithm may not be the global optimal to the problems.

On the contrary, the algorithm in present paper is a global optimization algorithm. All the monomers in the HP sequences have the same weights initially. The n-

monomer chain can move in many directions. After long-time iterations, many possible positions can be reached following our algorithm.

For the shorter chains, our algorithm is more time consuming than growth algorithm used in reference 6, but the situation inverses to the longest chain. Indeed, growth algorithm requires exponential time consuming with the increase of the chain's length. Comparing the results, it is apparently that the times increase much slower in our experiments which demonstrate that our algorithm has better performance when applied to longer HP sequences.

Table 1. Experimental results

N	(ε_{HH}, ε_{HP}, ε_{PP})	*Sequence*	E_{min}[a]	E_{min}[b]	*CPU time*[c]	*CPU time*[d]
58	(-1,0,0)	$PHPH_3PH_3P_2H_2PHPH_2PH_3PHPHPH_2P_2H_3P_2HPHP_4HP_2HP_2H_2P_2HP_2H$	-63	-44	3.81	0.19
103	(-1,0,0)	$P_2H_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HPHP_2HP_5H_3P_4H_2PH_2P_5H_2P_4H_4PHP_8H_5P_2HP_2$	-88	-54	10.39	3.12
124	(-1,0,0)	$P_3H_3PHP_4HP_5H_2P_4H_2P_2H_2P_4HP_4HP_2HP_2H_2P_3H_2PHPH_3P_4H_3P_6H_2P_2HP_2HPHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2PHPHPHPH$	-109	-71	24.25	12.3
136	(-1,0,0)	$HP_5HP_4HPH_2PH_2P_4HPH_3P_4HPHPH_4P_{11}HP_2HP_3HPH_2P_3H_2P_2HP_2HPHPHP_8HP_3H_6P_3H_2P_2H_3P_3H_2PH_5P_9HP_4HPHP_4$	-117	-80	36.01	110

[a] *Lowest energies found in present work*

[b] *Lowest energies found in reference 6*

[c] *CPU times (hours) cost on 3.0 GHz Intel P4(results in our experiments)*

[d] *CPU times (hours) cost on 667 MHz DEC ALPHA 21264 (results from reference 6)*

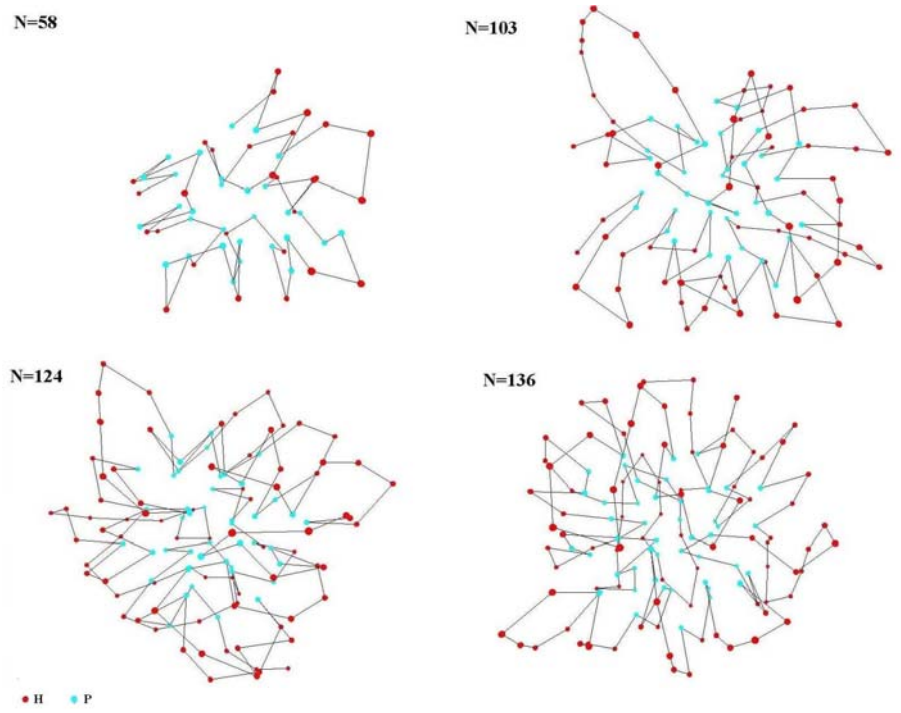

Fig. 1. The experimental results shown in 3D space

4 Discussion

In this paper we present a new continuous optimization algorithm for 3D protein structure prediction. The main idea of the algorithm is that all monomers share the same initial weight and will move in continuous three-dimensional space following certain physical theories. As a global optimization algorithm, our algorithm can search many potential solutions to find the optimum solution to the problems.

Comparing our results to the best results in previous works, we obtain lower energy states in all 3D cases. Moreover, our algorithm has lower time complexity than previous work. It will show more advantages to the proteins which have longer amino acid sequence.

Following the way of previous work, we used HP model, an abstract model of protein folds prediction, to study the problem. Actually, our algorithm can be used for a much wider range of applications. We anticipate it can be applied to more realistic protein models.

In the future work, we will try to add more information about the proteins into our algorithm, such as molectronics, experiential data, and examine the improvement on performance of the enhanced algorithm.

Acknowledgements

This paper is sponsored by NSF of China (No. 60273031), Education Ministry Doctoral Research Foundation of China (No. 20020055021) and Nankai university ISC. And the proofreading by Dr. Gu Dayong

References

1. K.F.Lau, K.A.Dill, A lattice statistical mechanics model of the cobformation and sequence space of proteins, Macromolecules, 22, 2002
2. Shortle, D., H.S. Chan, and K.A. Dill, Modeling the Effects of Mutations on the Denatured States of Proteins, Protein Science, 1（1992）: 201-215.
3. J. M. Deutsch, Long range moves for high density polymer simulations, J. Chem. Phys. 1997,106,8849-8856
4. E.M. O'Toole,A.Z. Panagiotopoulos, Effect of sequence and intermolecular interactions on the number and nature of low-energy states for simple model proteins,J. Chem. Phys.1993,98(4),3185-3190.
5. R. König and T. Dandekar Solvent entropy driven searching for protein modeling examined and tested in simplified models. Protein Eng.,2001, **14**, 329-335.
6. Hsiao-Ping Hsu, Vishal Mehra, Walter Nadler and Peter Grassberger, Growth Algorithms for Lattice Heteropolymers at Low Temperatures, J. Chem. Phys.2003(118): 444-448 .
7. Huang Wen-Qi, Huang Qin-bo, Shi He, An Quasiphysical Algorithm for 3D Protein Structure Prediction,J. of Wuhan University,2004,50(5)：586-590