http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

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

Information Technology Journal 12 (23): 7734-7738, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7734.7738
© 2013 Asian Network for Scientific Information

Improved PSO in the Application of the Protein Structure Prediction

Shigang Wang, Fangfang Zhou and Fengjuan Wang
School of Mechatronics Engineering, Qiqihar University, 161006, Heilongjiang Qiqihar, China

Abstract: Protein structure prediction occupies an important position on bioinformatics science. In this study, basic theory of Particle Swarm Optimization (PSO) and some theory models of protein folding study are introduced. Using modified particle swarm optimization, the protein structure prediction is predicted and good performance of algorithm is verified by testing results of Fibonacci sequence. From the experimental results it can be seen that PSO-SA algorithm accurately obtain the optimal solution of the lowest energy value. The algorithm can jump out of local optima and has good convergence performance. The experimental result is very satisfactory.

Key words: AB the lattice model, PSO, the fibonacci sequence, protein structure prediction

INTRODUCTION

Protein structure prediction problem, namely protein folding problem is one of core problems on bio informatics science. And research on protein spatial structure has important significance on biology field because the biological performance of protein depends on a great extent on protein spatial folding structure. Protein is a complex biology condensed matter system which contains many alignment polymer chain tissues formed by glycolipid polymerization of a amino units. Polymer system has many theoretical problems remaining to deeply study because of self-avoidance effects and interactive complexions (Kinjo and Nakamura, 2012). Meanwhile, biology system is a complex coupled open system which makes protein research to be a challenge subject (Yang *et al*., 2012).

How to predict the functional conformation using amino acid sequence of protein is an important problem in bio informatics science which explains protein structure prediction (Liu and Shen, 2013). The mastery of protein structure prediction can be conducive to understand the relationship between spatial structure and its function. Based on this understanding, the structure of natural protein can be recreated in order to improve its performance with purpose. Or unnatural protein with special function is designed according to true requirement (Li, 2012).

AB NON-LATTICE MODEL OF PROTEIN STRUCTURE PREDICTION

The main folding problem of computer simulation is the distinction between the large conformational space and the limited computing power of protein. Therefore, a lot of different simplified models are put forward in recent years. Also, the simplified model of proteins has become a new hotspot (Zou *et al*., 2011). This study will focus on the AB-lattice model which is based on the amino acid classification.

Stillinger *et al*. (1993), who proposed AB-lattice model (AB Off-Lattice Model), this model takes into account both hydrophobic and hydrophilic considered (Zhang and Liu, 2002). Angle of the model in the three amino acid residues connected between keys can be rotated to the 20 kinds of amino acids of this model based on affinity, hydrophobicity divided into two categories, respectively and expressed. Protein sequence of any one of n amino acids which corresponds to an angle of n-2 angles namely $\theta_2,...,\theta_{n1}$, as shown in Fig. 1 which is a predetermined range:

$$\pi \le \theta_i \le \pi \quad (i = 2,..., n-1)$$

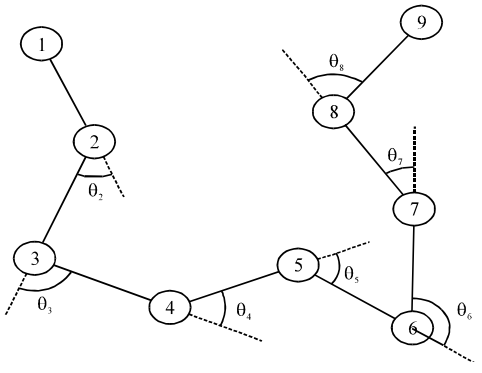

Fig. 1: Folding show of protein sequences in the two-dimensional space

Corresponding Author: Shigang Wang, School of Mechatronics Engineering, Qiqihar University, 161006, Heilongjiang Qiqihar, China

When $\theta_i = 0$, the three residues adjacent to the same line, the angle between the corresponding two keys of π which means that when θ_i is a positive reverse rotation angle.

Main chain energy V_1 is closely related with the folding angle regardless polar residues, not only from the energy and V_2 residues between the two, but also on interest-bearing and polar residues on both at the same time, every key on the main chain Effect of change is a residue of non-direct contact to the latter. For simplicity, we will use a set of residues encoded binary variables ξ_i, if the i-th residue is A, $\xi_i = -1$, if the i-th residue isB, $\xi_i = 1$. Protein sequence length is defined as the energy shown:

$$\varphi = \sum_{i=2}^{n-1} V_1(\theta_i) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} V_2(r_{ij}, \xi_i, \xi_j) \quad (1)$$

Which recorded as a function of distance r_{ij} bond angles:

$$r_{ij} = \left\{ \left[\sum_{k=i+1}^{j-1} \cos\left[\sum_{l=i+1}^{k} \theta_l \right] \right]^2 + \left[\sum_{k=i+1}^{j-1} \sin\left[\sum_{l=i+1}^{k} \theta_l \right] \right]^2 \right\}^{\frac{1}{2}} \quad (2)$$

where, V_1 is aboutsimple trigonometric, if no special force, will tend to be continuous linear bond ($\theta_i = 0$):

$$V_1(\theta_i) = \frac{1}{4}(1 - \cos\theta_i) \quad (3)$$

Expressions V_2 function is:

$$V_2(r_{ij}, \xi_i, \xi_j) = 4(r_{ij}^{-12} - C(\xi_i, \xi_j) r_{ij}^{-6}) \quad (4)$$

where, the coefficients v as follows:

$$C(\xi_i, \xi_j) = \frac{1}{8}(1 + \xi_i + \xi_j + 5\xi_i\xi_j) \quad (5)$$

The results show that the two hydrophobic residues have a stronger gravitational for each other and vice versa. There is little repulsion between the hydrophilic and hydrophobic residues. All the phenomena can reflect the original nature of proteins. In the two-dimensional non-lattice AB model, the protein sequence of length n where we want to get the most appropriate set of θ_i (i = 2,…, n-1) values to calculate the energy Eq. 1 minimum value. In this way, we can find a two-dimensional structure of protein sequences.

BASIC PSO AND PARALLEL PARTICLE SWARM OPTIMIZATION

PSO was first introduced by Kennedy and Eberhart (1995). In basic PSO model, each object is regarded as a particle with no mass or volume in n-dimension search space. And they have a certain flight speed and can adjust the flight speed of their own according to own experience and their fellow's experience.

Let:

- $X_i = (x_{il}, x_{i2},\ldots, x_{in})c_2$ is the current location of particle i
- $V_i = (v_{il}, v_{i2},\ldots, v_{in})$ is the current flight speed of particle i
- $P_i = (p_{il}, p_{i2},\ldots, p_{in})$ is the best passed location of particle i, namely the passed location with best value of particle i which is called the best location (Ceng *et al*., 2004). For the minimization problem, the smaller is the objective function value, the better the corresponding fitness

Let f(x) be the minimum object function and the current location of particle i is determined by the equation:

$$P_{i(t+1)} = \begin{cases} P_i(t) & \text{if } f(X_i(t+1)) \geq f(P_i(t)) \\ X_i(t+1) & \text{if } f(X_i(t+1)) < f(P_i(t)) \end{cases} \quad (6)$$

the number of this particle group is s and the best location of all the particles in the group is $P_g(t)$, namely best global location. Then:

$$P_g(t) \in \{P_0(t), P_1(t),..., P_s(t)\} \mid f(P_g(t)) = \min \{f(P_0(t)), f(P_1(t)), ..., f(P_s(t))\} \quad (7)$$

Whereas the upper definition, the modified equation of particle swarm optimization is described as:

$$V_{ij}(t+1) = V_{ij}(t) + C_1 r_{1j}(t)(P_{ij}(t) - X_{ij}(t)) + C_2 r_{2j}(t)(P_{gj}(t) - X_{ij}(t)) \quad (8)$$

$$X_{ij}(t+1) = X_{ij}(t) + V_{ij}(t+1) \quad (9)$$

Which, j means the exon j dimension of the particle, i mean the exon i dimension of the particle, t means the exon t generation. c_1, c_2 denotes speed constant ranging from 0 to 2 usually. $r_1 = U(0, 1)$, $r_2 = U(0, 1)$ are the independent random function.

The main defect of particle swarm optimization is that the speed of most particle approaches or even equals

zero, the location of the particle can't be updated and algorithm is at a standstill which can lead the whole to fall into local optimal state, because of small change of values of individual and population as well as higher and higher similarity of the particles. Then, the modified formula is listed as followed:

$$V_{ij}(t+1) = V_{ij}(t) + C_1 r_{1j}(t)(P_{ij}(t) - X_{ij}(t)) + C_2 r_{2j}(t)(P_{gj}(t) - X_{ij}(t)) + r * c_i r_{ij} \quad (10)$$

$$X_{ij}(t+1) = X_{ij}(t) + V_{ij}(t+1) \quad (11)$$

where, r is a small constant in Eq. 10 and other parameters are the same as those in basic particle swarm optimization. Modified algorithm is called as Parallel Particle Swarm Optimization (PPSO).

The basic step of PSO is listed as followed:

Step 1: Based on the initial process, the location and speed of the particles is initialized according to Eq. 6-7

Step 2: Calculate the proper value of each particle

Step 3: For each particle, compare the proper value with the best passed location. If better, change the location corresponding to the proper value for the best passed location

Step 4: Compare the proper value of each particle with that of best passed global particles. If better, change it for the best passed global location

Step 5: According to Eq. 10-11, correct the location and speed of the particle

Step 6: Judge whether to meet the complete requirement. Usually in order to get a value good enough or presupposition maximum iteration number, else return to step 2

Due to the high velocity of particle and the strong global search capability for algorithm, in the early of algorithm running, the fourth for $r \times c_i r_{ij}$ will much smaller than three items in before and has a little effect on the search capabilities of the algorithm. But in the medium-late period of algorithm running, the higher and higher similarity and more and more low-speed of the particle, not only the fourth item ensure that the particle search speed could not reach for 0, but also the particle search can continue to carry out local search.

FIBONACCI TEST SEQUENCES AND TEST RESULTS

In order to evaluating and comparing the performance of algorithm, this study carry out PPSO algorithm of folded structure prediction experiments on real-protein sequences. All experiments are conduct in using C language in Linux environment. In the algorithm, particle number is set to 20 and weights from 0.9 linear down to 0.4, c_1 and c_2 is set to 2.0 both, the maximum speed for particle of π in half.

Fibonacci is widely used as experimental standard artificial protein sequence (Tian *et al*., 2012) and defined as follows:

$$S_0 = A,\ S_1 = B,\ S_{i+1} = S_{i-1} * S_i$$

In which "*" indicates that the connection operation. Such as: S_2=AB, S_3=BAB, S_4=ABBAB and so on. A is a hydrophobic residue, it will always appear separately in the sequence and B is a hydrophilic residue, appear either alone or in pairs, then the Fibonacci sequence there is a hierarchy of expression:

$$S_i = S_{i-2} * S_{i-3} * S_{i-2} = (S_{i-4} * S_{i-5} * S_{i-4}) * (S_{i-5} * S_{i-6} * S_{i-5}) * (S_{i-4} * S_{i-5} * S_{i-4}) = \ldots$$

The Fibonacci series are shown in Table 1. Experimental results were compared with other methods experimental results obtained by the improved algorithm PPSO, the results shown in Table 2.

In which E_{HTML} experimental results obtained through the HTML algorithm (Li, 2006), E_{PERM} is the minimum energy by using the PERM algorithm (Liang, 2004) and the E_{PPSO} is the corresponding by using improved algorithm PPSO. Four different lengths Fibonacci sequence applied PSO-SA algorithm predicted conformation diagram shown in Fig. 2.

For the Fibonacci sequence length Len = 13, 21, 34, 55 of four Fibonacci, this study presents the lowest energy configuration which differs from the PSO and SA algorithms. From these configuration diagrams, it can be seen that: Len = 13 is the protein sequence conformation; Class A residue is surrounded by class B residues. And this set of phenomenon accords with the characteristics of protein. In Len = 21, 34, 55 conformation of the protein

Table 1: Fibonacci series (when 13≤n≤55)

A.S.	Len.	Sequence
S_{13}	13	ABBABBABABBAB
S_{21}	21	BABABBABABBABBABABBAB
S_{34}	34	ABBABBABABBABBABABBABABBABBABABBAB
S_{55}	55	BABABBABABBABBABABBABABBABBABABBABBABABBABABBAB

Table 2: Lowest energy values obtained by different algorithms (2D)

A.S.	E_{HTML}	E_{PERM}	E_{PPSO}
S_{13}	-3.2235	-3.2167	-3.2441
S_{21}	-5.2881	-5.7501	-5.9980
S_{34}	-8.9749	-9.2195	-9.4341
S_{55}	-14.4089	-14.9050	-15.4574

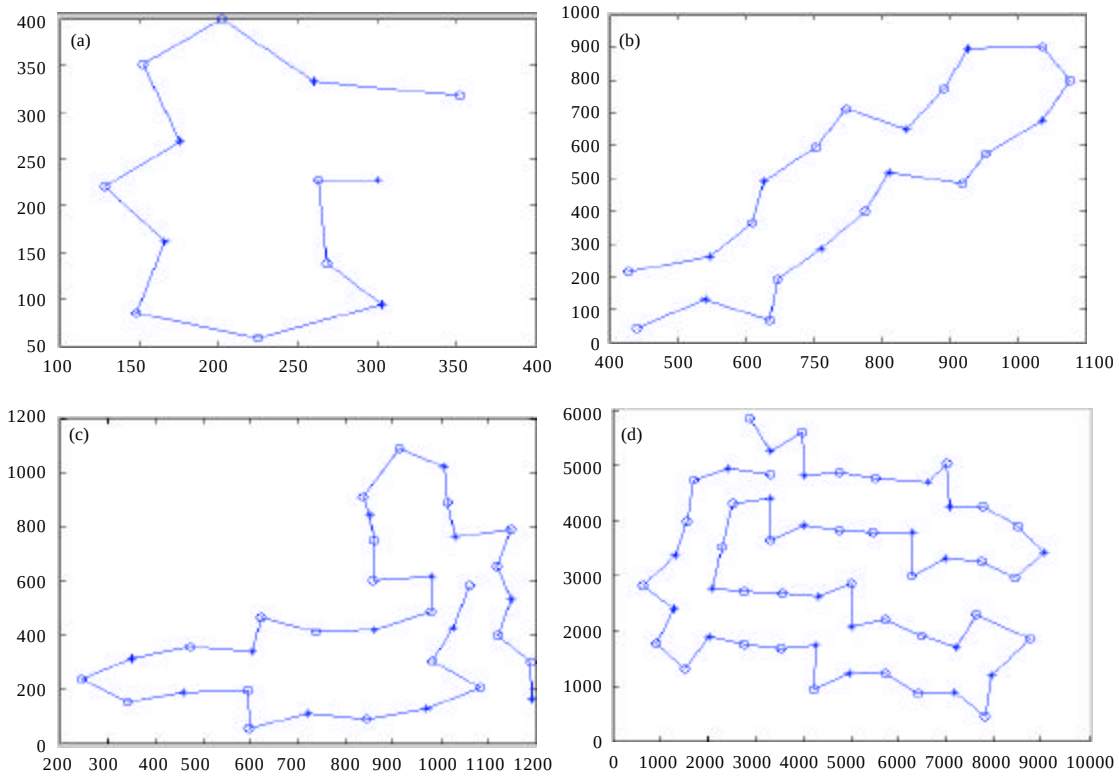

Fig. 2(a-d): Conformational map of length of 13, 21, 34, 55 (a) Len = 13, (b) Len = 21, (c) Len = 34 and (d) Len = 55

sequence, class A residues formed multiple beam and it is basically surrounded by Class B residues. Therefore, it is by and large in line with the true characteristics of the protein.

CONCLUSION

Protein structure prediction is a long-term strategic task. So far, although amazing progress has been made, there is no a complete protein structure prediction methods with no deviation. The author improves the speed of the particle swarm optimization and use it to experiment, Got the lengths Len = 13, 21, 34, 55 of four conformations Fibonacci sequence diagram. As comparing the results with other algorithms, we confirm that the PPSO algorithm has a better computational performance. In recent years, we can predict that protein structure prediction will soon have the breakthrough with the development of computer and bio informatics research (Zheng, 2012). And we also believe that protein structure prediction will play a more important role in protein engineering, molecular design and screening of drugs, gene therapy and molecular biology research (Lian and Xiong, 2013).

ACKNOWLEDGMENT

This study was supported by the Natural Science Foundation of Heilongjiang Province, China under Grant E201106 and the Project of Department of Education of Heilongjiang Province, China under Grant 12521597.

REFERENCES

Ceng, J.C., J. Jie and Z.H. Cui, 2004. The Particle Swarm Optimization. Science Press, China, pp: 13-16.

Kennedy, J. and R. Eberhart, 1995. Particle swarm optimization. Proc. IEEE Int. Conf. Neural Networks, 4: 1942-1948.

Kinjo, A.R. and H. Nakamura, 2012. Composite structural motifs of binding sites for delineating biological functions of proteins. PloS One, Vol. 7. 10.1371/journal.pone.0031437

Li, D.P.A., 2012. Novel structural position-specific scoring matrix for the prediction of protein secondary structures. Bioinformatics, 28: 32-39.

Li, Y.L., 2006. Wavelet nuclear protein secondary structure prediction based on support vector machines. Shenzhen Univ., 23: 117-121.

Lian, Y.J. and H.L. Xiong, 2013. Multi-kernel learning method for predicting protein secondary structures. J. Comput. Appl., 33: 43-45.

Liang, F., 2004. Annealing contour Monte Carlo algorithm for structure optimization in an off-lattice protein model. J. Chem. Phys., 120: 6756-6763.

Liu, Y. and S.P. Shen, 2013. An overview of protein function prediction methods. Chin. J. Bioinformat., 11: 33-38.

Stillinger, F.H., T. Head-Gordon and C.L. Hirshfeld, 1993. Toy model for protein folding. Phys. Rev. E, 48: 1469-1477.

Tian, P.F., J. Wu and H.B. Zhu, 2012. Clonal selection algorithm based on quantum-lattice AB protein folding prediction. J. Wuhan Univ., 58: 144-148.

Yang, A., R. Li, W. Zhu and G. Yue, 2012. A novel method for protein function prediction based on sequence numerical features. Match-Commun. Math. Comput. Chem., 67: 833-843.

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.

Zheng, L.J., 2012. Protein folding research and application of molecular chaperones. J. Agric. Process. Complex, 12: 65-68.

Zou, D., Z. He, J. He and Y. Xia, 2011. Supersecondary structure prediction using Chou's pseudo amino acid 2composition. J. Comput. Chem., 32: 271-278.