Proceedings of the 10th
World Congress on Intelligent Control and Automation
July 6-8, 2012, Beijing, China

Hybrid Artificial Bee Colony and Particle Swarm Optimization Approach to Protein Secondary Structure Prediction*

Mengwei Li
Science and Technology on Aircraft Control Laboratory
Beihang University
Beijing 100191, P R China,
lmwwjl@163.com

Haibin Duan, *Senior Member, IEEE*, Dalong Shi
State Key Laboratory of Virtual Reality Technology and Systems
Beihang University
Beijing 100191, P R China

***Abstract* - Proteins are crucial in the biological processes, and their structure determines whether they can function well or not. Since the theory presented by Anfinsen that proteins' space structure is entirely determined by the primary structure came out, it is possible for us to predict the structure of proteins through their primary structure without any experiment. In order to reach this target, the prediction problem can be formulated as an optimization problem that is set to find the lowest free energy conformation. In this paper, a hybrid Artificial Bee Colony (ABC) with Particle Swarm Optimization (PSO) Algorithm is used to solve this problem. Considering that the two algorithms have complementary characteristics, we combine them together and find out a better optimization results through this new approach. Experimental results have demonstrated the feasibility and effectiveness of our proposed approaches.**

Index Terms* - *Protein structure prediction, Artificial Bee Colony, Particle Swarm Optimization.

I. Introduction

Protein is a most important substance in organisms' body, and most of the biological processes are accomplished by it. There are two main factors that can influence the function of protein. One is the sequence of amino acids, the other is the folding structure. Nowadays, benefited by the research of DNA conducted by Watson and Crick, the scientists now can obtain the sequence of amino acids only with the help of genetic material. However, the protein folding structure prediction is regarded as a grand challenge and is one of the great puzzling problems in computational biology [1].

NMR and X-ray Crystallography are the two main experimental ways to solve this problem. But the processes are so time consuming that the progress is far more behind than the discovery of protein sequences. Therefore, the prediction through computation is a necessity. In 1950s, Anfinsen [2] proposed a theory that proteins' folding structure was entirely determined by the primary structure, which gives us a possibility to solve this problem as long as we know the sequence of the amino acids [3]. According to the laws of thermodynamics and kinetic law, any molecule's structure tends to have the lowest free energy so that it can be most stable. Therefore, the prediction problem can be converted into an optimization problem whose goal is to find out the lowest free energy of the protein. The corresponding protein structure is exactly the prediction result we respected. Thus, the computation methods can be used to solve this problem with computer, which will save a lot of time and money.

Since the protein structure prediction can be regarded as an optimization problem, there must be an energy function of the protein which can properly represent the true energy situation between component units of protein according to the relative position. By using specific method, the energy function can be calculated to find out the least value. Thus, the prediction task can be accomplished.

Considering that there is a torsion angle between every two amino acids, one specific protein must have many degrees of freedom. Therefore, the energy function is multivariable and multimodal. Finding its least value will be a difficult issue. Most researchers at present tend to solve this problem by genetic algorithm and have gained decent results. While swarm intelligence is a new active research area [4], we presented a hybrid Artificial Bee Colony (ABC) with Particle Swarm Optimization (PSO) for solving this complex optimization problem, and the comparative results show the effectiveness of our proposed approach.

II. Protein Structure Model For Prediction

Every protein molecule tends to have the lowest free energy so that the structure is the most stable in specific condition. Thus, we need an energy function that can truly reflect the performance of protein and find out the least value of the function computationally. The corresponding solution reflects the most stable protein structure.

In this paper, we choose AB Off-Lattice Model as our protein structure model. AB Off-Lattice Model was first represented by Frank H. Stillinger [5] in 1993 given the HP Lattice Model. In this model, all the amino acids are divided into two categories: one is hydrophobic, expressed as A; the

* This work is partially supported by Natural Science Foundation of China(NSFC) grant #60975072, Program for New Century Excellent Talents in University of China grant #NCET-10-0021, Aeronautical Foundation of China under grant #20115151019, Open Fund of the State Key Laboratory of Virtual Reality Technology and Systems under grant #VR-2011-ZZ-01, the Fundamental Research Funds for the Central Universities of China, and Student Research Training Program (SRTP) of Beihang University.

other one is hydrophilic, expressed as B. In addition, there are some rules in this model. First is that the bond length between every two amino acids is undistinguishable. Secondly, each amino acid is simplified as a sphere in specific dimension. Thirdly, the torsion angel of two contiguous bond ranges from $-\pi$ to π in two dimensions. The presentation of the structure in this model is shown in Fig.1 below. On the basis of the above rules, knowing the sequence of the amino acids, the protein structure of n amino acids is determined by $n-2$ variables $\theta_2,...,\theta_{n-1}$.

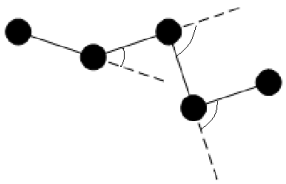

Fig. 1 The presentation of the protein structure in AB Off-Lattice Model in two dimensions

The energy function of AB Off-Lattice Model can be defined as follows [5]:

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

Where V_1 represents the energy of the backbone and is only determined by the sequence of amino acids. Where V_2 represents the energy of the two amino acids that are not contiguous and is determined by not only the sequence but also the distance between each other. The notation ξ_i represents the category of the amino acid. When the amino acid is hydrophobic, $\xi_i = 1$; on the contrary, $\xi_i = -1$. The notation r_{ij} is equal to the distance between amino acid i and amino acid j, and can be defined as following equation:

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

Since V_1 is a function only relative to θ_i, it can be defined as

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

The variable 's expression is as follows

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

Where the notation $C(\xi_i, \xi_j)$ is represented as:

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

Using the above free energy function, when the sequence of the amino acids is given, all of the corresponding torsion angels are regarded as the variables of the function. The task is to find out the optimal solution of this energy function, and the solution can reflect the prediction results of the protein structure. Thus, the goal of the prediction can be achieved.

III. Artificial Bee Colony With Particle Swarm Optimization Algorithm

A. Artificial Bee Colony Algorithm

Artificial Bee Colony Algorithm was first put forward by Karaboga [6] from Turkey in 2005. It simulates the artificial bees to find out the best nectar source with swarm intelligence. In the ABC, every single solution reflects an independent artificial bee, and every numerical value of the target function is equivalent to a nectar source. The task is to find out the best source for all the artificial bees in the algorithm, which means to find out the least value of the target function.

Like other swarm intelligence, the intelligence in artificial bees depends on the cooperation with each other. The exchange of information in artificial bee colony plays an important role in the cooperation. This kind of exchange is accomplished by means of waggle dance, odor and so on [7]. Just like the artificial bee colony in reality, at this algorithm, all the artificial bees are mainly divided into two categories. One is called employed foragers. Their job is to gather honey from their corresponding nectar source, and to exchange information of their source with other bees. The specific employed foragers whose source is the best at the present will become the ones who lead others to their source. The other one is the unemployed foragers. They are the one who don't find out the suitable source by themselves. They can keep looking for the source or follow the lead foragers to gather honey. The source is suitable or not is decided by the fitness value of the specific solution. The larger the fitness value, the better the source's quality is. That is, if we want to get the least value of the target function, we can let fitness value be the opposite number of the function value.

Although this algorithm is newly out forward, it has shown its advantages compared to genetic algorithm and particle swarm optimization algorithm. Because it does local search and global search in every generation, which can raise the probability of finding out the optimal solution and avoid falling into the local optimal solution in great degree. [8]

B. Particle Swarm Optimization Algorithm

In 1995, an American psychologist Kennedy and an electrical engineer Eberhart proposed a swarm intelligence algorithm simulating the process of the birds' foraging called Particle Swarm Optimization algorithm [9, 10]. Similarly, in this algorithm, the every single solution stands for an individual. However, the individual in this kind of algorithm is a particle without mass and volume in dimensions. The only characteristics that determine each individual is its location and velocity, and the location is expressed by the solution of the function. Each solution also has a corresponding fitness value that is related to the function value. That is, the goal to find out the solution which has the

best fitness value is equal to the goal to find out the best location of those that all the particles have ever gone through.

The evolutionary equation and the comparison among the locations that all the particles have gone through are the two most important parts of the algorithm. For every single particle, if the current location is better than all the locations that it itself has ever experienced before, then the particle's location is replaced by the current one. And for every single particle, if its current location is better than any other locations that all the particles have ever experienced before, then the current global best location is replaced by it. After the comparison, the location and the velocity of each single particle is changed by the following equation:

$$v(t+1) = v(t) + c_1 r_1(t)\left(p(t) - x(t)\right) + c_2 r_2(t)\left(p_g(t) - x(t)\right) \quad (6)$$

$$x(t+1) = x(t) + v(t+1) \quad (7)$$

Where the notation t represents the iteration time, and the notation v and x represent the velocity and the location of the particle. The notation p and p_g the notation represent the best location of each individual and the global best location. The notations c_1 and c_2 are two constants varies from zero to two usually. And the functions $r_1 \sim U(0,1)$, $r_2 \sim U(0,1)$ are two independent random functions.

Because of the advantages that it has simple model and no need for gradient information, it is easy to achieve, it has less parameters in the algorithm and so on, PSO is widely applied to many kinds of optimal problems [8].

C. Artificial Bee Colony with Particle Swarm Optimization Algorithm

Artificial Bee Colony Algorithm is a newly proposed optimization algorithm and is becoming a hot topic nowadays. Because its high probability of avoiding the local optimal, it can make up the disadvantage of Particle Swarm Optimization Algorithm. Moreover, Particle Swarm Optimization Algorithm is in easy model, which can help us to find out the optimal solution more easily. In such circumstances, we bring the two algorithms together so that the computation process can possesses both of the advantages. We call it ABC-PSO.

In our combination model, we divided the colony into two parts: one possesses the swarm intelligence of Artificial Bee Colony; the other one has the particle swarm intelligence. We assume that there is cooperation between the two parts. In each iteration time, one part which finds out the better solution will share its achievement with the other part. In other words, the inferior solution will be replaced by the better solution and will be substituted into the next iteration time.

The process of ABC-PSO is as follows:

Step 1. Initialization of Parameters: set the number of individuals of the swarm; set the maximum circle-index of the algorithm; set the search range of the solution; set the other constants needed in both ABC and PSO.

Step 2. Initialization of the colony: firstly, generate a colony with specific number of individuals. Then on one hand, as a bee colony, it is divided into two categories, employed foragers and unemployed foragers according to each individual's fitness value; on the other hand, as a particle swarm, calculate the fitness value of each particle and take the best location as the global best location. We assume that the cyclic number is represented by $iter$, and $iter = 1$.

Step 3. In bee colony, for each employed forager, it keeps searching other sources around it while gathering honey from the current corresponding source. Once the new source is better than the current, it will turn to the new one. And for all the unemployed foragers, the number of them to follow each employed forager is depending on the quality of the corresponding source. That is, the better the quality of the source, the more unemployed foragers will turn to it. The same as the employed ones, they are keeping searching the new sources around them and determine which one is turned to according to the fitness value of each source. After all the choices above have been made, the best solution is generated in this iteration which we called it $GlobalMin$.

Step 4. In particle swarm, after the comparison among the solutions that each particle has experienced and the comparison among the solutions of all the particles have ever experienced, the best location in this iteration will be found out which is called Ta_best. Then all the particles will evolve according to the evolution equation (6) and (7).

Step 5. The minimum of the value $GlobalMin$ and the value Ta_best which we call it $GlobalMins$ is defined by the following equation:

$$GlobalMins = \begin{cases} GlobalMin & if\ GlobalMin \le Ta\_best \\ Ta\_best & if\ Ta\_best \le GlobalMin \end{cases}$$

And the value $GlobalMin$ and the value Ta_best will both be equal to the value $GlobalMins$, and will be substituted into next iteration $iter = iter + 1$.

Step 6. Check out that if the number of the circles is greater than the maximum of the circle-index. If not, return to Step 2; if it is, end the computing process and record the value $GlobalMins$.

The flow chart of ABC-PSO process is also described in Fig.2.

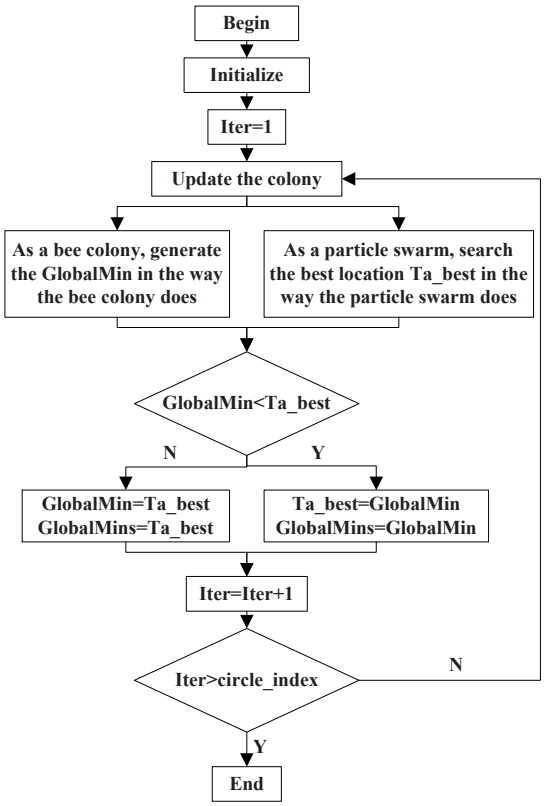

Fig. 2 The flow Chart of ABC-PSO

IV. Experimental Results

In order to verify the feasibility and effectiveness of our proposed ABC-PSO, series of experiments are conducted. The results are showed in Fig.3-Fig.5. It is obvious that our proposed ABC-PSO can achieve better solutions, and has a faster convergence speed.

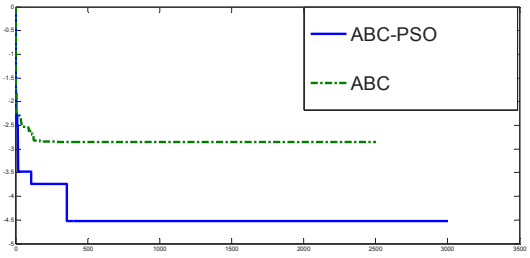

Fig. 3 The lowest free energy both ABC and ABC-PSO achieved of the sequence of amino acids AAAAA

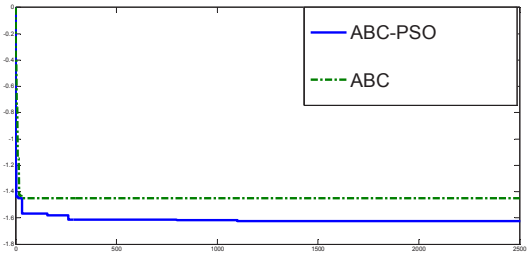

Fig. 4 The lowest free energy both ABC and ABC-PSO achieved of the sequence of amino acids AABA

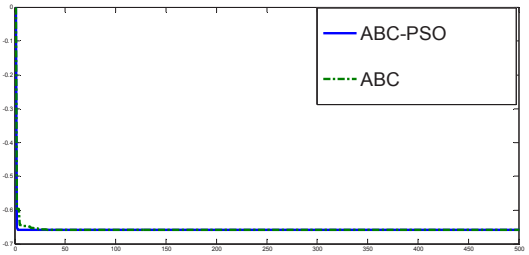

Fig. 5 The lowest free energy both ABC and ABC-PSO achieved of the sequence of amino acids ABA

With ABC-PSO, we have achieved a breakthrough in predicting the short chain proteins. The result is shown in the following table 1.

SEQUENCE	NEWENERGY	ORIGINAL [Reference 11]
AAA	-0.658205	-0.65821
AAB	0.0322266	0.03223
ABA	-0.658205	-0.65821
ABB	0.0322266	0.3223
BAB	-0.0302734	-0.03027
BBB	-0.0302734	-0.03027
AAAA	-2.30218	-1.67633
AAAB	-0.603223	-0.58527
AABA	-1.61781	-1.45098
AABB	0.0672041	0.06720
ABAB	-0.665732	0.64938
ABBA	-0.933634	-0.03617
ABBB	0.00470414	0.00470
BAAB	-0.183991	0.06172
BABB	-0.247224	-0.00078
BBBB	-0.309654	0.13974
AAAA	-4.19007	-2.84828
AAAAB	-2.07544	-1.58944
AAABA	-3.14916	-2.44493
AAABB	-0.55749	-0.54688
AABAA	-3.22415	-2.53170
AABAB	-1.54562	-1.34774
AABBA	-1.72809	-0.92662
AABBB	0.0401702	0.04017
ABAAB	-1.54099	-1.37647
ABABA	-2.22020	-2.22020
ABABB	-0.790606	-0.61080
ABBAB	-1.11429	-0.00565
ABBBA	-0.882661	-0.39804
ABBBB	-0.253233	-0.06596
BAAAB	-0.715991	-0.52108
BAABB	-0.317672	0.09621
BABAB	-0.844718	-0.64803
BABBB	-0.439788	-0.18266
BBABB	-0.591503	-0.24020
BBBBB	-0.737172	-0.45266

Table 1Comparison between ABC-PSO's results and original results

V. Conclusion And Further Work

This paper has proposed a new way to the protein secondary structure prediction, which is based on our proposed ABC-PSO model. Because of the complementary advantages of the two swarm intelligence algorithms, ABC-PSO has shown the superiority. The corresponding results of the experiments are also shown in this paper, which explain

that ABC-PSO has a breakthrough in the short chain protein structure prediction than ever before.

Our further work will focus on applying evolving ABC-PSO to the long chain protein structure prediction, which is also a challenging issue.

REFERENCES

[1] Chiu, T.-L. and R. Goldstein, Optimizing energy potentials for success in protein tertiary structure prediction. Folding and Design, 1998, Vol.3, No.3, pp. 223-228.

[2] C. B. Anfinsen, Principles that govern the folding of protein chains [J]. Science, 1973, 181(4096): 223-227.

[3] K. F. Lau, K. A. Dill, A lattice statistical mechanics model of the conformational and sequence space of proteins. Macromolecules, 1989, 22: 3986-3997.

[4] Yang, X. S., Engineering Optimizations via Nature-Inspired Virtual Bee Algorithms, in Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach. 2005, pp. 317-323.

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

[6] Karaboga D, An idea based on honey bee swarm for numerical optimization. Kayseri: Erciyes University, 2005.

[7] Frisch K V. The Dance Language and Orientation of Bees. Cambridge: Harvard University Press, 1967.

[8] Haibin Duan, Xiangyin Zhang, Chunfang Xu, Bio-inspired Computation. Beijing: Science Press, 2011.

[9] Kennedy J, Eberhart R. Particle Swarm optimization. Proceedings of IEEE International Conference on Neural Networks, 1995: 1942-1948.

[10] Eberhart R, Kennedy J. A new optimizer using particle swarm theory. Proceeding of the 6th International Symposium on Micro-Machine and Human Science, 1995: 39-43.

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

[12] Chunfang Xu, Haibin Duan, Fang Liu, Chaotic Artificial bee colony approach to Uninhabited Combat Air Vehicle (UCAV) path planning, Aerospace Science and Technology, 2010, Vol. 14, No. 8, pp. 535-541.

[13]Haibin Duan, Chunfang Xu, Senqi Liu, Shan Shao. Template matching using chaotic imperialist competitive algorithm. Pattern Recognition Letters, 2010, Vol.31, No.13, pp.1868-1875.