Application of improved intelligent ant colony algorithm in protein folding prediction

Fengjuan Wang, Cheng Xu, Shufeng Jiang and Fengxia Xu

Abstract
While the single ant colony algorithm and the fish swarm algorithm have many advantages, they also have various shortcomings. After analyzing the advantages and disadvantages of the ant colony algorithm and the fish swarm algorithm, this paper uses the complementary principle of the two algorithms to effectively fuse the two population intelligent algorithms. The improved swarm intelligence algorithm is applied to the well-considered protein folding prediction problem, and the simplified protein structure Toy model is verified, and the ideal results are obtained. The improved algorithm enhances the search ability, and the computational efficiency is greatly improved, ensuring the accuracy of the operation.

Keywords
Ant colony algorithm, fish swarm algorithm, fusion of intelligent algorithms, two-dimensional Toy model, protein folding prediction

Received 7 December 2018; revised 4 February 2020; accepted 12 May 2020

Introduction
Protein folding prediction in bioengineering has always been a focus of attention. According to the statistics of relevant departments, about one-third of the proteins in normal human cells will produce folding errors, but the intracellular mechanism will promptly discover the existence of these error proteins and quickly remove them. When the amount of protein folding is too large to be treated by these mechanisms, it can cause various diseases, similar to Alzheimer’s disease (commonly known as Alzheimer’s disease), cavernous hemangioma, family genetic cholesterol, rabies, etc. Therefore, the study of protein folding is not only significant but also urgent.

Ant colony algorithm and artificial fish swarm algorithm (AFSA) are called two emerging cluster intelligence algorithms. Although it has not been long since they were born, the speed of their progress is very fast, and they have been tried by many scholars in the structural model of protein, and has achieved certain results. This paper analyzes the characteristics of the two algorithms and uses the principle of complementing the advantages and disadvantages to effectively fuse the two algorithms. This proves that this swarm intelligence performs better than a single algorithm and can achieve the optimal effect of the target problem.

Description of protein folding prediction
Protein folding structure prediction refers to predicting the natural structure of a protein from a given protein sequence. Proteins with a complete primary structure must have normal biological functions and must be folded into specific spatial structures. In other words, the biological properties of a protein are largely determined by the multidimensional spatial structure of its peptide chain. In the 1930s, Chinese scientist Wu Xian first pointed out that changes in the external environment can lead to the destruction of protein spatial structure and the loss of biological activity, but it...
does not destroy its primary structure. The denatured protein usually becomes a stretched peptide chain, but under certain conditions, it can be refolded into the original spatial structure and restored to its original activity. This shows that the internal structural properties of the protein are contained in the protein chain, and the spatial configuration of the protein can be inferred through the protein chain. As early as 1973, Anfinsen put forward the famous theory that “the natural conformation of protein is the lowest conformation of free energy” in the journal “Science”, which laid the theoretical foundation for the prediction of protein folding structure, that is, by calculating the lowest energy of protein.

Theoretical basis for predicting protein thermodynamics structure. Therefore, from a mathematical point of view, we can attribute the problem of protein folding structure prediction to a global optimization problem.

The general flow of protein structure prediction is to first analyze the sequence of a protein after a given protein sequence, and find a homologous protein in a library of known protein sequences. If found, and the spatial structure of the protein is determined, it is possible to use homologous information to establish a tertiary structure model of the target protein. However, in actual research, we generally first predict the secondary structure of the protein, and then predict the tertiary structure. The accuracy of the tertiary structure prediction depends on the results of the secondary structure prediction, see in Figure 1. Searching the entire conformational space of the protein finds the lowest energy conformation as the conformation of the target protein, which is the protein folding prediction.

At present, the main problems in protein structure prediction are: the accuracy of prediction is not high enough, the calculation speed is not fast enough, and there is a certain distance between the model of structure prediction and real protein. In recent years, the swarm intelligence has brought new hopes for solving the above problems, especially in the aspect of solving complex problems, which shows the characteristics of dynamic, adaptability, and robustness, which helps to compensate for the existing protein structure prediction problems, weaknesses and defects. In recent years, with the rise of various intelligent random algorithms and widely used in various fields, it has been found that the use of these algorithms to solve the problem of protein folding has unique advantages. In this paper, a new algorithm combining ant colony and AFSA is applied to the protein folding prediction problem, and good results have been achieved.

**Principles of ant colony algorithm**

In many years of research, biologists have discovered that ants in nature are invisible, and that the intelligence and mobility of a single ant are negligible. Only relying on communication with peers, as well as the special material left behind when walking—pheromone, to change their direction of action. Based on this, Italy’s Dorgio and its research team proposed an algorithm in the early 1990s, named Ant Colony Optimization (ACO). The algorithm is also called ant colony optimization and is a probability algorithm. This is a general heuristic method. The concept of ant colony algorithm is convenient to understand, it has a few of parameters and does not involve complex situations such as mutation and crossover. The application field of the algorithm is very wide. From appearance to present, research on ant colony algorithms continues to rise in various industries, and the algorithm can be seen in almost every optimization field.

**Description of ant colony algorithm**

When the ant passed from $i$ to $j$, the path node in the beginning can be set as $A, A = \{0, 1, \ldots, n-1\}$. Suppose the quantity of ants is $n$, and the information-heuristic factor is $\alpha$. This indicates the relative importance of the trajectory, reflecting the role of pheromones accumulated by ants during exercise during ant movement. The larger $\alpha$ is, it indicates that the high probability for subsequent ants to choose this path. The expected heuristic factor is $\beta$, which indicates the relative importance of visibility. And it reflects the importance of the heuristic factor $\alpha$ in the entire path during the movement of the ant. Pheromone volatilization coefficient is $\rho$, which is a constant that represents weight. The times of iteration is $N_c$. The tattoo table is tabu. The initial setting is $\phi$. Select the starting point $S$ and the target point $S$, and allow the ant colony to move from the starting point $S$. The probability for the ant to move from $i$ to $j$ is calculated with formula (1)

$$p_{ij}^t(t) = \frac{\tau_{ij}^t(t)\eta_{ij}^t(t)}{\sum_{j \in c} \tau_{ij}^t(t)\eta_{ij}^t(t)}$$ (1)
In which, $\sum t_{ij}^g(t)\eta_{ij}^g(t)$ is the denominator, so it is not 0. $t_{ij}^g(t)$ refers to the pheromones when the ant $k$ passed from node $i$ to $j$, and $\eta_{ij}(t)$ represents a pheromone heuristic factor from node $i$ to $j$, and $\eta_{ij}(t)$ is expressed in formula (2)

$$\eta_{ij}(t) = \frac{1}{d_{ij}} \quad (2)$$

In formula (2), $d_{ij}$ represents the distance for the ant to move from $i$ to $j$; $\eta_{ij}$ means the heuristic degree for the ant to move from $i$ to $j$; $\tau_{ij}$ refers to pheromones. For each movement of the ant $k$, the node $j$ will be added to the tattoo table till all the circling ants have reached the destination. Calculate the length of the path taken by each ant at the end point and save it. The updated pheromone is represented by equation (3)

$$\tau_{ij}(t + 1) = \rho\tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^k(t, t + 1) \quad (3)$$

In the formula, when the ant $k$ passes $j$ from the node $i$, the formula (3) is $\Delta\tau_{ij}^k(t) = \frac{Q}{t}$, or it is 0.

**Thoughts on improving ACO–AFSA algorithm**

The ant colony algorithm has good parallelism and high precision of the algorithm, which is a prominent advantage of the ant colony algorithm. Artificial ants can search the entire search space, and information exchange between artificial ants can be achieved, and finally the purpose of optimization is achieved. In the ant colony algorithm, in fact, each artificial ant interacts with the environment, and each artificial ant is connected to each other by means of pheromone “bridge”. Specifically: ants must first have some kind of inertia, keep moving forward, rather than walking around. Second, after the ants have fixed the trend of movement, they must also maintain a certain degree of randomness. Taboo, like the particles in the particle swarm algorithm, has been moving in a crazy linear motion.11 This allows the ants to do their best to follow the original trend and try to generate new search ranges. But ants also have mistakes, so do not move to places with high pheromones, and walk blindly, which produces a local optimal situation, which is also a shortcoming in the ant colony algorithm.

The AFSA is a group intelligent bionic algorithm that simulates the behavior of fish in nature. By optimizing the four underwater behaviors of a single artificial fish, local optimization is achieved, and finally the global optimum is achieved. In the fish’s eating behavior, the number of trials is set less, the probability of artificial fish random walk is increased, the ability of local optimal solution can be avoided, and the global optimality is obtained.12,13 The use of the crowding factor can limit the size of the fish population, which allows the artificial fish to search more freely. The occurrence of trailing behavior can make the artificial fish move toward a better state, and can also cause the artificial fish caught in the local optimum to escape from the local optimal region and move toward the global optimal direction. A typical disadvantage of the AFSA is its low accuracy.

Through the above description, it can be concluded that the ant colony algorithm and the AFSA have complementary relationships, and the intelligent algorithms of the two populations are improved, and the effective fusion can improve the performance and efficiency of the algorithm. The improved group intelligence algorithm can avoid the disadvantages of the single algorithm, strengthen the advantages of the algorithm, and form a more stable and comprehensive group intelligent algorithm.14

**Steps to improve the ACO–AFSA algorithm**

This paper is based on a single ant colony algorithm and AFSA to improve and merge into a new algorithm—ant colony fish swarm algorithm, referred to as ACO–AFSA. The steps of the ACO–AFSA algorithm are as follows:

- **Step 1:** Initial setting of the ant colony algorithm;
- **Step 2:** Initial setting of the AFSA;
- **Step 3:** Generate a new initial solution;
- **Step 4:** Calculate fitness;
- **Step 5:** Update pheromone concentration and bulletin board information;
- **Step 6:** Determine if the condition is met.

Now, we will elaborate on the above steps.

1. Initialize the number of artificial ants $m$, heuristic factor, hope heuristic factor, information persistence factor, and information intensity $Q$.
2. Initially set the number of artificial fish $M$, step, artificial fish field of view visual, congestion factor.
3. The ant colony algorithm and the AFSA alternately iterate, and use comparison to determine the superiority of the solution. In the same problem, if the ant colony algorithm first searches for a good solution, it is randomly placed in an artificial fish position in the artificial fish group. If the AFSA first searches for a good solution, then it is randomly
placed in an artificial ant position in the artificial ant colony. Meanwhile, judge whether it is crowded according to the formula $q_{ij} = \frac{2^{n(t)}}{\sum_{i=1}^{N} \tau_{ij}(t)}$. If it is within the allowable range, then the population continues to optimize according to the original plan. Otherwise, it will go further in the allowed range.\cite{15}

4. Get the adaptation value according to the formula $Y = f(X)$.

5. Follow the instructions to update the pheromone concentration and update the bulletin board information.

6. Determine whether the conditions for ending the algorithm process are met. If it is satisfied, the output solution is the optimal solution. If it is not satisfied, return to step 3 to continue the operation.

**Prediction of protein folding in the Toy model**

Protein folding prediction faces two major problems. One is how to convert the cumbersome protein structure into a simple mathematical model to simplify the problem; the other is how to find efficient search methods to efficiently search for protein conformations.\cite{16} At present, the simplified protein structure models mainly include Toy model, H-hydrophobic amino acid, P-hydrophilic amino acid (HP) grid model, three-dimensional cubic model, and hexagonal lattice model.\cite{17} In this paper, the improved ant colony fish swarm algorithm is applied to the protein folding prediction problem of the two-dimensional Toy model, and the effect is ideal.

**Two-dimensional Toy model**

The Toy model was proposed by Stillinger and his companions in the 1990s and is now also known as the AB non-grid model.\cite{18} In this model, the common 20 amino acids are represented by B (hydrophilic or non-polar) and A (hydrophilic or polar) to represent hydrophilic amino acids and hydrophobic amino acids.\cite{19} The adjacent amino acids are connected by a steel bond, and the angle between the two amino acid atoms can be formed, that is, the bond angle, as shown in Figure 2.

The amino acid sequence in Figure 2 is nine and has seven angles, respectively $\theta_{2}, \theta_{3}, \theta_{4}, \theta_{5}, \theta_{6}, \theta_{7}, \theta_{8}$. For N amino acids of other sequence lengths, there are also N-2 $\theta$ perspectives, from $\theta_{2}$ to $\theta_{n-1}$. Each bond angle in the sequence is $\theta \in [-\pi, \pi]$. When $\theta_{i} = 0$, the three amino acid atoms are in a line.

No matter how long the amino acid sequence is, their energy functions are the same, which is the sum of Van der Waals gravitational potential energy and bending potential energy.\cite{20} The length of a protein is $N$, and the Van der Waals potential between the separated residues is

$$E_1 = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} 4(r_{ij}^{-12} - C_i(\xi_i, \xi_j)r_{ij}^{-6}) \quad (4)$$

$\xi_i$ refers to amino acid residue species. It is regulated that $\xi_i$ when the residue is represented by A, the value is 1; and when the residues are B, the value of $\xi_i$ is -1. $C_i(\xi_i, \xi_j) = \frac{1}{8}(1 + \xi_i + \xi_j + 5\xi_i \xi_j)$, its value is only 1/2, -1/2, or 1.

These three values represent strong attraction between AA, mutual exclusion between AB, and weak attraction between BB, showing the nature of the protein from the side.\cite{21} It can be seen from equation (4) that Van der Waals gravitational potential energy is determined by the hydrophobicity, polarity, and distance of non-adjacent residues.

The main chain bending potential energy is

$$E_2 = \sum_{i=2}^{N-1} \frac{1}{4}(1 - \cos \theta_i) \quad (5)$$

It can be seen from equation (5) that the magnitude of the potential energy and the protein sequence itself are not determined by the chain $\theta$. Based on this, the spatial conformational energy function will be smooth, that is, $E = E_2 + E_1$

$$E = \sum_{i=2}^{N-1} \frac{1}{4}(1 - \cos \theta_i) + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} 4(r_{ij}^{-12} - C(\xi_i, \xi_j)r_{ij}^{-6}) \quad (6)$$

In the two-dimensional Toy model, this paper transforms the protein folding prediction problem into a
function optimal value problem, that is, how to obtain N-2 bond angles to minimize the energy value. In this way, we abstract the abstraction into concrete, simplify the complex problem, and change the structure of the real protein without changing its characteristics.

Prediction of two-dimensional Toy models protein folding

When using the Toy model for protein folding prediction, the most common method is to use the Fibonacci sequence as the amino acid sequence of the protein. The Fibonacci sequence is also called the golden section, and the form of the series is: 0, 1, 1, 2, 3, 5, 8, 13, 21… The third item of the series begins with each item being the sum of the previous two items, which is expressed as equation (7)

\[
F_0 = 0, \quad F_1 = 1, \quad F_n = F_{n-1} + F_{n-2} \quad (n \geq 2, \ n \in N^*)
\]

When performing protein folding prediction, convert it to \( S_0 = A, S_1 = B, S_{i+1} = S_i \times S_{i-1} \). It is to be noted that "\(*\)" in the conversion formula does not represent multiplication, but represents a connection between two atoms. Therefore, \( S_3 = BAB \), and \( S_4 \) can be expressed as \( S_4 = ABBAB \). It can be found from the expression of \( S_3 \) and \( S_4 \) that: A does not appear in pairs, and B may appear alone or in pairs. In the real protein sequence, hydrophobic residues are also present alone, and polar residues can exist alone and in pairs. This also proves that the use of Fibonacci for protein prediction is reasonable and correct.

Table 1 shows the lengths of the protein sequences calculated using the Fibonacci sequence, which are 5, 8, and 13, respectively. This paper experiments on Intel (R) Core i3 2.40GHz CPU, 2.00GB RAM, and 32-bit operating system, using MATLAB7.0 software to test these three sequences. The important parameters in the algorithm are set to: \( \alpha = 2.0, \beta = 3.0, \rho = 0.65, Q = 100, \delta = 0.60 \). The maximum number of iterations and the number of attempts are 80. The results of the test are shown in Table 2.

The protein folding prediction of the two-dimensional Toy model was performed using the improved algorithm of this paper. Under the premise of guaranteeing the quality of the solution, the three sequences can find the lowest energy value. In order to increase the persuasiveness of the improved ant colony fish swarm algorithm in this paper, it is indicated that the improved algorithm can be applied to the protein folding prediction problem. The results of other literature calculations are listed below, as shown in Table 3.

It can be seen from Table 3 that with the improved algorithm of this paper, the energy value can also be obtained, and the effect is higher than that of the existing research. In order to explain more intuitively the results of this paper, the experimental results will be compared with the line graph, as shown in Figure 3. It can be seen from the line graph of Figure 3 that the new algorithm has a lower trend than the original algorithm. It shows that the convergence speed of the algorithm is faster, and new algorithm can find better

Table 1. Fibonacci series.

| Serial number | Sequence length | Sequence   |
|---------------|----------------|------------|
| 1             | 5              | ABBAB      |
| 2             | 8              | BABABBAB   |
| 3             | 13             | ABBABBABABBAB |

Table 2. Test result.

| Serial number | Sequence length | The energy value |
|---------------|-----------------|------------------|
| 1             | 5               | -1.726           |
| 2             | 8               | -3.109           |
| 3             | 13              | -6.175           |

Table 3. Result comparison.

| Serial number | Sequence length | In this paper, the energy value | Literature energy value |
|---------------|-----------------|--------------------------------|------------------------|
| 1             | 5               | -1.726                         | -1.000                 |
| 2             | 8               | -3.109                         | -2.000                 |
| 3             | 13              | -6.175                         | -5.000                 |

Source: reproduced with permission from Hou, 2014.24
results in a shorter time. The structure of real proteins is extremely simple, using computer simulations of protein structures for protein folding prediction. Although the true structure of the protein cannot be completely reflected, the energy value can provide a reliable reference value for the stability of the natural protein structure. The data shown in the table not only show the feasibility of the algorithm, but also show its effectiveness. The energy constellations of these three sequences are shown in Figures 4 to 6. Through these three energy constellation diagrams, we can also find that the longer the energy constellation of the sequence, the better the effect, and the closer it is to the appearance of the real protein structure.

Applying the improved ant colony fish swarm algorithm to the two-dimensional Toy model in the protein folding prediction problem, the Fibonacci sequence is tested and the results are valid. Compare the test results with the results in other literatures, and draw a constellation diagram, the effect is outstanding. The results of this paper are slightly better, and the conformational map can basically reflect the physical properties of proteins.

Conclusion
Protein folding prediction is a very hot topic in the field of bioengineering, involving many aspects. The use of swarm intelligence algorithms for protein folding prediction is a novel topic and a trend for future research. In this paper, the improved ant colony fish swarm algorithm is applied to the protein folding prediction problem. The artificial protein Fibonacci sequence was tested using a two-dimensional Toy model. The test results are compared with the results in other literature and an energy conformation map is given. There are still many places in this paper that can be improved through research. The improved algorithm in this paper can also be applied to higher domain layers through deformation. These ideas will be verified in later research.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The paper is supported by Heilongjiang University education reform project (SJGY20180565), 2018 Annual Education Department Basic Research Business Youth Innovation Talent Project (135309376), Intelligent
Manufacturing Equipment Innovation Team–Heilongjiang Province Intelligent Manufacturing Equipment Industrialization Collaborative Innovation Center (135409102), and 2019 Qiqihar City-level Science and Technology Plan General Project Contract Number (GYGG-201919).

ORCID iD
Fengjuan Wang https://orcid.org/0000-0002-9829-0565

References
1. Tao HB, Ma S, Cao JJ, et al. Advances in research on protein structure prediction methods. Technol Forum 2016; 43.
2. Chu DL, Zhao WC and Lin CC. Research status and development trend of group intelligence algorithm. Inf Commun 2015; 11: 38–39.
3. Dorigo M, Maniezzo V and Colomn A. Ant system: optimization by a colony of cooperating agents. Systems, man, and cybernetics, part B: cybernetics. IEEE Trans Syst, Man, Cybern B 1996; 26: 29–41.
4. Fei T and Zhang LY. Research on modern intelligent optimization algorithm. Inf Technol 2015; 10: 26–29.
5. Liu X. Study of particle swarm optimization for protein structure prediction. Fujian Agriculture & Forestry University, 2011. pp.3–5.
6. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973; 181: 223–230.
7. Xu B. An efficient ant colony algorithm based on wake-vortex modeling method for aircraft scheduling problem. J Comput Appl Math 2017; 317: 157–170.
8. Asghari S and Azadi K. A reliable path between target users and clients in social networks using an inverted ant colony optimization algorithm. Karbala Int J Mod Sci 2017; 3: 143–152.
9. Lin Y, Gong YJ and Zhang J. An adaptive ant colony optimization algorithm for constructing cognitive diagnosis tests. Applied Soft Computing 2017; 52: 1–13.
10. Yilmaz H and Yilmaz M. A mathematical model and ant colony algorithm for multi-manned assembly line balancing problem. Int J Adv Manuf Technol 2017; 89: 1935–1939.
11. Li WJ, Zhou BY, Zhu XF, et al. Hybrid swarm intelligent parallel algorithm research based on multi-core clusters. Microprocess Microsyst Part A 2016; 47: 151–160.
12. Kumar KP, Saravanans B and Swarup KS. Optimization of renewable energy sources in a microgrid using artificial fish swarm algorithm. Energy Procedia 2016; 90: 107–113.
13. Luan XY, Li ZP and Liu TZ. A novel attribute reduction algorithm based on rough set and improved artificial fish swarm algorithm. Neurocomputing Part A 2016; 174: 522–529.
14. Zhang QK, Liu WG, Meng XX, et al. Vector coevolving particle swarm optimization algorithm. Inform Sci 2017; 394–395: 273–298.
15. Zhao M, Yin H, Zheng LJ, et al. Flexible job shop scheduling based on improved artificial fish swarm algorithm. China Mech Eng 2016; 27: 1059–1065.
16. Yan RX, Wang XF, Xu WM, et al. A review of protein folding recognition methods. Bioinformatics 2015; 13: 231–238.
17. Lin ZX, Guo JW and Lin J. Node importance algorithm under Hadoop framework for protein function prediction. Comput Syst Appl 2016; 25: 77–82.
18. Li CY, Ding YR and Xu WB. Research of protein structure prediction based on multiple-layer quantum-behaved particle swarm optimization and toy model. Comput Appl Chem 2010; 27: 1574–1578.
19. Junni LZ and Jun SL. A new sequential importance sampling method and its application to the two-dimensional hydrophobic–hydrophilic model. J Chem Phys 2010; 117: 3492–3498.
20. Ramychitra D and Ajeeh A. A multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction. Gene Rep 2017; 8: 100–106.
21. Emerson A and Amala A. Protein contact maps: a binary depiction of protein 3D structures. Physica A 2017; 465: 782–791.
22. Rashid S, Saraswathi S, Kloczkowski A, et al. Protein secondary structure prediction using a small training set (compact model) combined with a complex-valued neural network approach. BMC Bioinf 2016; 17: 1–18.
23. Du TC, Liao L, Wu H, et al. Prediction of residue-residue contact matrix for protein–protein interaction with Fisher score features and deep learning. Methods 2016; 110: 97–105.
24. Hou CX. The research of protein structure prediction algorithms based on improved particle swarm. Dalian University, 2014. pp.43–48.