Protein Secondary Structure Prediction via Pigeon-Inspired Optimization*

Wei Zheng, Hemeng Sun, and Haibin Duan, *Senior Member, IEEE*

Abstract - **Proteins are the essential elements in all creatures' life process. The prevailing experimental method to detect protein structure is time consuming. According to Anfinsen's theory, the primary structure is the key to shaping the three-dimensional structure of the protein. Thus, the prediction of proteins' structure avoiding complex experiments is theoretically feasible. Algorithms are applied to make the protein structure prediction while they have some unsatisfactory blemish. For instance, relatively high Gibbs free energy or long iterating progress. In this paper, a new algorithm is introduced to solve this problem. Its name is "Pigeon-Inspired Optimization(PIO) Algorithm". PIO can work out an accurate prediction in relatively short iterations. The advantages and feasibility of this algorithm will be demonstrated through the experiments, comparing to Particle Swarm Optimization.**

I. Introduction

On a chemical level, proteins are polypeptides chains, playing a crucial role in the life process of all creatures. Protein engineering which is an indispensable part of bioinformatics becomes the forefront development field of modern biotechnology. While structure determination is taxing and frustrating, the protein secondary structure prediction is priceless for its value in the experimental design.

The sequence of amino acids as well as the folding structure is the main factor determines the function of different proteins.

Each protein has its own unique ammo acid composition and sequence, and it plays the valid biological function only when it is folded into the correct structure. Therefore, careful researches about protein folding(folding code) is essential to the computational biology[1].

Scientists mainly use Nuclear Magnetic Resonance and X-ray crystallography to figure out the space structure of the protein. However, these experimental process is devastating due to the enormous time cost. Thus, there is a discrepancy between the former progress and the emergence of protein sequences. Hence, it is necessary to solve the prediction problem through computation instead of experiments. In 1950s, Anfinsen[2] reckoned that the protein folding problem equals to the amino acids sequence problem, because the primary structure independently determined the three-dimension structure of the proteins[3]. Additionally, on the basis of kinetic energy law and thermodynamics basic law, it is found that the most stable molecule's structure has the lowest Gibbs free energy. As a result, the problem of protein structure prediction is transferred to a mathematical problem to search for the lowest Gibbs free energy of the protein by computation with limited time and energy.

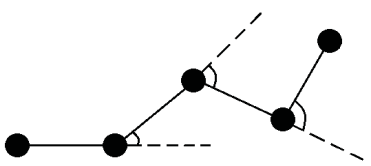
When the problem of protein structure prediction is converted into an optimization problem, a fitness function of protein representing the real energy situation with respect to the relative location of component units of protein is needed. The prediction can be accomplished by using a specific method to discover the least value of fitness function. In this paper, PIO algorithm is introduced to solve this problem.

II. Adopted Protein Structure Model

The protein structure can keep stable when the sum of the energy (of every molecule) remains in the lowest. Thus, an appropriate energy function is needed in order to simulate the properties of protein to search for the minimum value of the Gibbs free energy function. The solution will be the most steadfast protein structure correspondingly.

In this paper, AB Off-Lattice Model is selected as the protein structure model due to its' simplicity and feasibility. Frank H. Stillinger[4] first proposed AB Off-Lattice Model in 1993 based on the HP Lattice Model. There are two types of the amino acids in this model. The hydrophobic one is expressed as A while the hydrophilic one is presented as B. Additionally, there exists several regulations to be followed when constructing protein structure in this model as follows: Firstly, the bond length between the adjacent amino acids is regard as the same. Secondly, all amino acids are regard as independent spheres in the three-dimension space. Thirdly, torsion angles of two adjacent bonds is ranging from - π to π .

The planar structure model is presented in picture below(Fig.1) according to the above rules. The characteristic of protein structure can be represented by $n-2$ variables $\theta_2,...,\theta_{n-1}$.



Wei Zheng is with School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing, 100191 P. R. China. (phone: +86-10-8233-8672; e-mail: zhengweihao2@163.com).

Hemeng Sun is with School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing, 100191, P. R. China.(phone: +86-10-8233-8672; e-mail:605242027@qq.com).

Haibin Duan is with Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering Beihang University (BUAA), Beijing, 100191, P. R. China (phone: +86-10-8231-7318; e-mail: hbduan@buaa.edu.cn).

978-1-4673-8318-9/16/$31.00©2016 IEEE

Fig. 1 The planar sketch of the protein structure according to AB Off-Lattice Model

The Gibbs free energy function according to AB Off-Lattice Model is presented 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 the notation V_1 reflects the energy of the backbone. According to Stillinger's study, the sequence of amino acids is the sole impact factor of this part. Where the notation V_2 reflects the energy of the two adjacent amino acids and is determined by both sequence and the distance between each other. The difference between amino acids is expressed through the notation ξ_i. To represented amino hydrophobic acid, $\xi_i = 1$; to show the hydrophilic one, $\xi_i = -1$. The notation r_{ij} is the distance between amino acid i and amino acid j that is defined as follows:

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

Since V_1 is only related to θ_i, it can be defined as

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

The variable V_2 can be defined as

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

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

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

Once the amino acids sequence is fixed, all of the torsion angles are settled as the input of the Gibbs free energy function. Arming with the above free energy function, our task is simplified to find the best solution of the fitness function(Gibbs free energy function). The solution reflects the main characters of the protein structure and our mission is completed.

III. Pigeon-Inspired Optimization Algorithm

A. Behavior of Pigeons

Pigeons are very common creatures(see Fig. 2), which can be easily found in most countries, cities and regions. Ordinary as they are, pigeons once were important messenger in militaries due to their special homing strategy. For instance, pigeons played significant roles in the Australian, German, French, American, and United Kingdom forces, during the World War I and II. Pigeons have the unique homing ability using a special combination of the geomagnetic field, landmarks as well as the sun to migrate from one place to another.

Figure. 2. Picture of pigeons

Guilford [5] reckons that, during the different periods of the journey, navigational tools used by pigeons are not the same. His group has developed a mathematical model to predict the actual process of which strategy pigeons are used during different parts of their journey. According to Guilford, pigeons mainly adopt compass-like tools when they just start their voyage. While at the time they are close to the destination, they switch their strategy to landmarks instead of persisting on compass-like tools.

According to further investigation, tiny magnetic particles in pigeons' beak contributes a lot to their superior ability to discern different magnetic fields and finally results in their outstanding homing skills. Studies indicate that signals from magnetite particles are delivered from the pigeons' nose to their brain crossing the complicated trigeminal nerve[6].

Meanwhile, there are clear evidence shows that the sun participates in pigeons' navigation. To be exact, pigeons can distinguish discrepancy in altitude between the sun and the land. Thus they can use the sun as a compass[7].

In addition, it is found that landmarks such as buildings, valleys and rivers play an indispensable role in pigeons' homing strategy. The birds may follow these features to find their way around instead of rushing toward the destination recklessly.

The homing strategy of pigeons can be formulated in a new optimization algorithm that can be associated with the objective function. Its name is PIO[8]-[10]. Details will be elaborated as follows.

B. Pigeon-Inspired Optimization Algorithm

First, two operators in order to idealize several homing characteristics of pigeons is proposed, under several rules.

1) Map and Compass Operator

By using magnetic signal to sketch a map in their tiny brains, pigeons are able to sense the earth field. In this strategy, the altitude of the sun plays a role of the compass in the adjustment of the direction. When pigeons are near to their destination, they depend less on this strategy.

2)Landmark Operator

Pigeons will be divided to two groups according to the familiarity of the landmarks when they get close to them, the birds that are familiar to the landmark and the bird are unversed to the landmark. The former will head toward the destination while the latter will follow the former.

a. Mathematical Model of Map and Compass Operator

In the mathematical model, imaginary pigeon without mass and volumes in a D-dimension search space replace the

pigeons naturally for simplicity. The velocity and the position of different pigeons are updated according to the equations:

$$V_i^t = V_i^{t-1} + rand * (X_* - X_i^{t-1}) \quad (6)$$

$$X_i^t = X_i^{t-1} * (1 - e^{-R*t}) + V_i^t \quad (7)$$

where the notation t is the iteration number and *rand* is a random number from 0 to 1. R ranges from 0 to 1, represents the map and compass operator, and X_* is the global best position at current that is refreshed every iteration by comparing different fitness values of every pigeon.

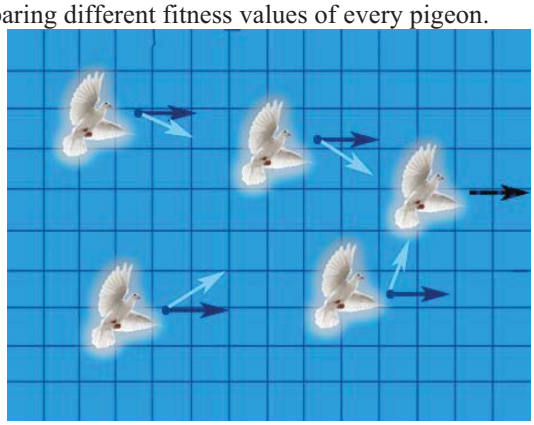


Fig3. Map and compass operator model

As depicted in Fig.3, the positions of pigeons are determined by calculating the sum of two vectors decided by the map and compass operators. As can be seen in the picture, the unique pigeon's position is the most suitable one in this iteration after comparing all these positions, while other pigeons amend its flying direction following the special pigeon, which are sketched by the tilting vectors Meanwhile, the horizontal vectors represent its original flying direction($X_i^{t-1} * (1 - e^{-R*t})$). Obviously, the following flying direction is the sum of the black arrow and the blue arrow under vector addition rules.

In the iterating process, X_i^t relies more on X_i^{t-1} rather than V_i^t as iteration number t increasing, which shows pigeons rely less on the map and compass strategy.

Similarity can be found between the updating procedure of PIO and PSO. Actually, pigeon algorithm can be seen as the development of the standard particle swarm optimization.

b. Mathematical Model of Landmark Operator

To simulate the homing behavior of pigeons when they are close to the destination, half of the number of pigeons $pigeonnum_i$ in every iteration will be eliminated because they are unfamiliar with the landmarks and far from the destination. Then, the centre of group members' position $xcenter_i$ will replace the destination and other pigeons may directly fly to the center. In this operator, the update of variables are given as follows:

$$pigeonnum_i = pigeonnum_{i-1} / 2 \quad (8)$$

$$xcenter_i = xsum_i / pigeonnum \quad (9)$$

$$X_i = X_{i-1} + rand * (xcenter_i - X_{i-1}) \quad (10)$$

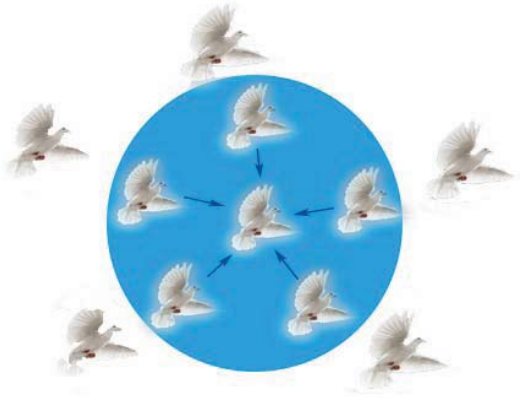


Fig. 4 Landmark operator model

As depicted in Fig. 4, the center of all positions(right in the center of the circle) is defined as the destination in every iteration. Then, pigeons(pigeons that are outside of the circle) that are unversed and distant from the destination will fly after pigeons(pigeons that are inside the circle) that are close to their destination. In this way, two pigeons will share the same position.

In this searching process, only a few generations are needed to be updated due to the short distance between pigeons and their destination. The number of pigeons will soon decrease to 1, whose position is just the destination we search for.

To solve a real problem, the parameter R will range from 0 to 1. In the actual implementation, we usually adjust the range of the search to restrict the position of pigeons. The process of protein secondary structure prediction via PIO method is described as follows:

Step 1: Input Gibbs free energy function of AB Off-Lattice Model as the fitness function.

Step 2: Initializations of the basic parameters: Set the length of the protein chain D, the map and compass factor R and the total chains number N,. The number of iteration Nc for two operators will be input as well. Whether the value of R is appropriate will have an essential impact on the performance. To get a desired performance, adjustment of R is indispensable.

Step 3: Initialization of Position and Velocity: Set a randomized velocity and position for initial protein chain. Then, calculate the current best position in the group by comparing each protein structure's fitness.

Step 4: Use map and compass operator. First, refresh position and velocity of all chains according to (6) and (7). If the position is illegal, fix its position to the near boundary in this iteration. Then, search for the new best position of the group by comparing different chains' fitness.

Step 5: If $Nc > Nc1_{max}$, move toward step 6. Otherwise, return to step 4.

Step 6: Use landmark operator: Rank all chains according to their Gibbs free energy(fitness function). Updated the parameters according to (8), (9) and (10). The half of potential chains left behind share the position with other chains. Then, refresh the best position of the group and the best fitness function's value.

Step 7: If $Nc > Nc2_{max}$, quit the program. Deal with the

results as you need. If not, go back to Step 6.

The detailed procedure can also be illustrated with Fig.5.

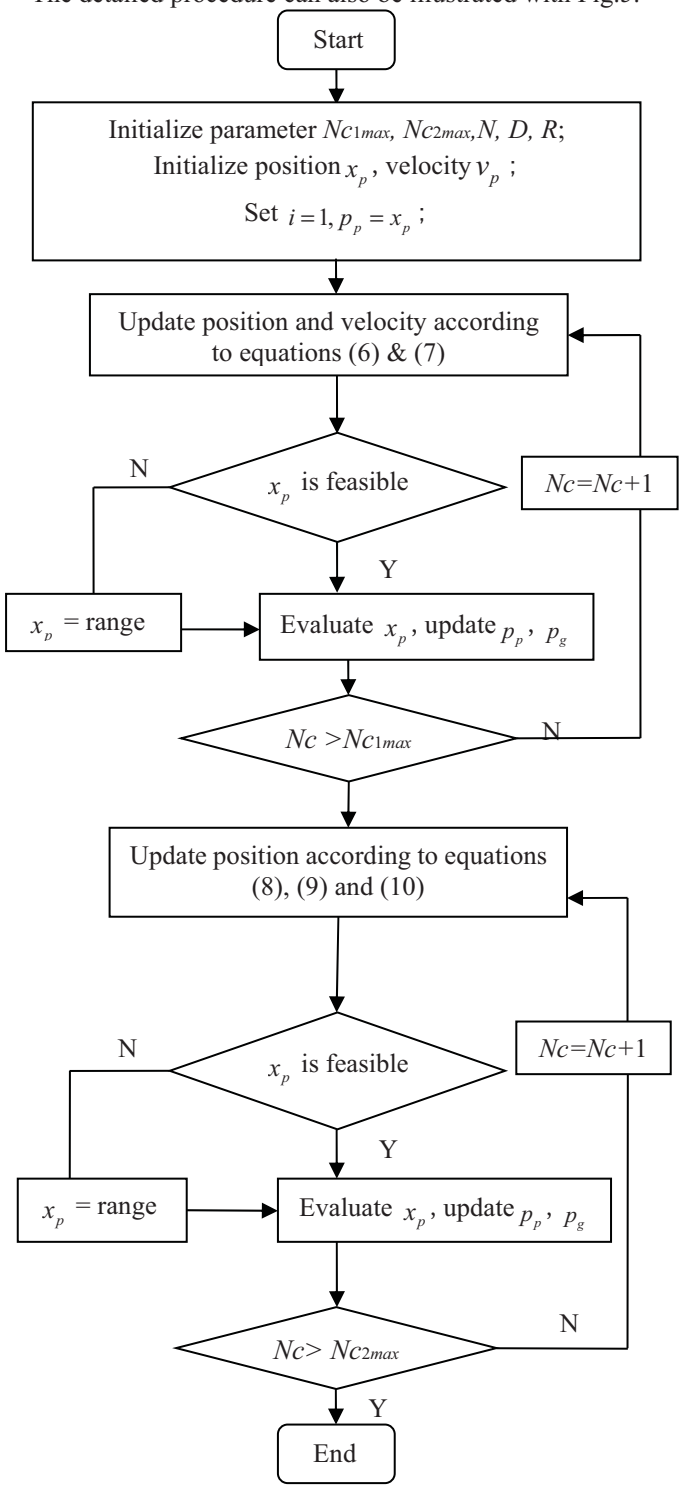


Fig.5. Procedure of protein secondary structure prediction via PIO method

IV. Experimental Results

In this section, a series of experiments using Matlab is conducted for the purpose of evaluating the feasibility and advantages of the PIO algorithm. Our proposed PIO is compared with PSO. In PIO algorithm, parameters are selected as follows: $Nc1_{max}$=250, $Nc2_{max}$=50, R=0.2, N=500. The experimental results are illustrated in Fig.6- Fig.8. From the diagram, no doubt our PIO algorithm can work out the same good result like PSO and with a faster convergence speed in fewer iterations.

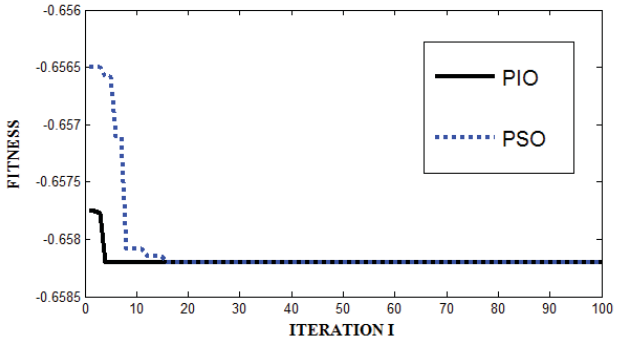


Fig. 6 The performance of PIO and PSO applying to the amino acids sequence AAA.

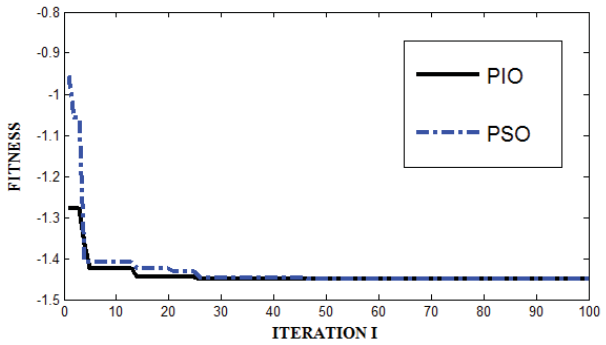


Fig. 7 The performance of PIO and PSO applying to the amino acids sequence AABA.

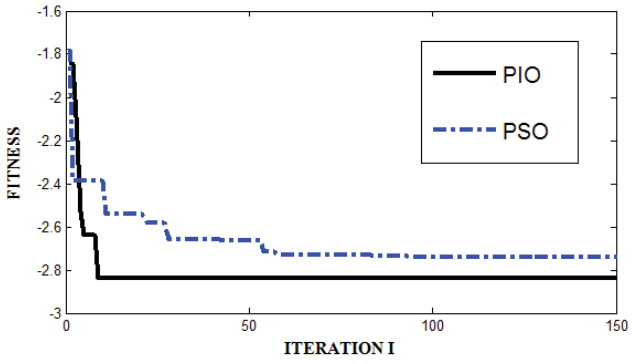


Fig. 8 The performance of PIO and PSO applying to the amino acids sequence AAAAA.

The following table shows the comparison of PIO, PSO method on the same protein sequence. Each test is made over 30 independent runs. The mean value of the fitness function are illustrated. It can be found that PIO performs close to PSO algorithm. In most cases, the difference is negligible.

SEQUENCE	PIO	PSO	ORIGINAL [11]
AAA	-0.65820465	-0.65820466	-0.65821
AAB	0.032226563	0.032226563	0.03223
ABA	-0.65820465	-0.65820466	-0.65821
ABB	0.032226563	0.032226563	0.3223
BAB	-0.03027344	-0.03027344	-0.03027
BBB	-0.03027344	-0.03027344	-0.03027
AAAA	-1.67600581	-1.676327	-1.67633

AAAB	-0.58520399	-0.58527277	-0.58527
AABA	-1.45053611	-1.4509772	-1.45098
AABB	0.067204501	0.067204136	0.06720
ABAB	-0.64932595	-0.64937535	-0.64938
ABBA	-0.03596413	-0.03617095	-0.03617
ABBB	0.00470431	0.004704136	0.00470
BAAB	0.06171753	0.061717167	0.06172
BABB	-0.00078266	-0.00078283	-0.00078
BBBB	-0.13966472	-0.13973795	-0.13974
AAAAA	-2.82109733	-2.82346902	-2.84828
AAAAB	-1.57606353	-1.5894433	-1.58944
AAABA	-2.41189786	-2.39229021	-2.44493
AAABB	-0.54548922	-0.54687776	-0.54688
AABAA	-2.50856773	-2.5316965	-2.53170
AABAB	-1.33326606	-1.3453983	-1.34774
AABBA	-0.91051678	-0.92662111	-0.92662
AABBB	0.040171356	0.040170229	0.04017
ABAAB	-1.3657604	-1.3764666	-1.37647
ABABA	-2.18908857	-2.2202007	-2.22020
ABABB	-2.17712232	-2.2202007	-0.61080
ABBAB	0.003855049	0.012116345	-0.00565
ABBBA	-0.38898748	-0.39803978	-0.39804
ABBBB	-0.06166152	-0.05960381	-0.06596
BAAAB	-0.51937971	-0.52107587	-0.52108
BAABB	0.09620954	0.096206698	0.09621
BABAB	-0.64645063	-0.64802496	-0.64803
BABBB	-0.17458391	-0.18265725	-0.18266
BBABB	-0.23502489	-0.24020359	-0.24020
BBBBB	-0.44711549	-0.45266354	-0.45266

TABLE 1COMPARISON BETWEEN PIO'S RESULTS, PSO'S RESULTS AND ORIGINAL RESULTS.

V. CONCLUSION

This paper applied a novel algorithm, combine with 2D AB Off-Lattice mode, to protein structure prediction problem. The experimental results clearly show that:

1) PIO is a new algorithm that holds merits in solving several constrained linear problems as well as nonlinear ones.

2) PIO performs close to PSO algorithm(the difference is negligible in most cases) in searching for the best solutions of complex problems, and seems to have a faster convergence speed especially in a long sequence. Thus it has great value when facing the problem to detect protein structure. PIO can obtain good solutions in much fewer iterations when it comes to the field of long protein sequences.

3) Applying PIO algorithm to predict the long-chain protein structure is still challenging due to the locally optimal solution.

4) As new algorithm, PIO has potential possibility to be improved to adapt to other real-life problems besides protein predilection and be optimized to avoid locally optimal solution.

Our further work will concentrate on adjustment of this nascent algorithm in order to avoid the local optimum solutions as well as get a better performance facing the challenge of long-chain protein structure prediction. Meanwhile, using different models is on the schedule to improve the consistency of the predilection and the real situation.

REFERENCES

[1] T. L. Chiu 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 and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence space of proteins". *Macromolecules*, 1989, 22: 3986-3997.

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

[5] T. Guilford, S. Roberts, and D. Biro, "Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models". *Journal of Theoretical Biology*, vol. 227, no. 1, pp. 25-38, Mar. 2004.

[6] C. V. Mora, C. V. Davison, J. M. Wild, and M. M. Walker, "Magnetoreception and its trigeminal mediation in the homing pigeon". *Nature*, vol. 432, no. 7016, pp. 508 -511, Nov. 2004.

[7] A. Whiten, "Operant study of sun altitude and pigeon navigation". *Nature* vol. 237, pp. 405-406, Jun. 1972.

[8] H. B. Duan, and P. X. Qiao, "Pigeon-Inspired Optimization: A New Swarm Intelligence Optimizer for Air Robot Path Planning". *International Journal of Intelligent Computing and Cybern etics*, 2014, vol.7, No.1, pp.24-37

[9] H. B. Duan, and X. H. Wang. "Echo State Networks with Orthogonal Pigeon-Inspired Optimization for Image Restoration". *IEEE Transactions on Neural Networks and Learning Systems*, 2016, in press, DOI:10.1109/TNNLS.2015.2479117

[10] B. Zhang, and H. B. Duan. "Three-Dimensional Path Planning for Uninhabited Combat Aerial Vehicle Based on Predator-Prey Pigeon-Inspired Optimization in Dynamic Environment". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, in press, DOI:10.1109/TCBB.2015.2443789

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