Protein Tertiary Structure Prediction Using Artificial Bee Colony Algorithm

Hesham Awadh A. Bahamish
School of Computer Science
Universiti Sains Malaysia
11800 Penang Malaysia
hesham@cs.usm.my

Rosni Abdullah
School of Computer Science
Universiti Sains Malaysia
11800 Penang Malaysia
rosni@cs.usm.my

Rosalina Abdul Salam
School of Computer Science
Universiti Sains Malaysia
11800 Penang Malaysia
rosalina@cs.usm.my

Abstract

Proteins are essential for the biological processes in the human body. They can only perform their functions when they fold into their tertiary structure. Protein structure can be determined experimentally and computationally. Experimental methods are time consuming and high-priced and it is not always feasible to identify the protein structure experimentally. In order to predict the protein structure using computational methods, the problem is formulated as an optimization problem and the goal is to find the lowest free energy conformation. In this paper, Artificial Bee Colony algorithm (ABC) is a swarm intelligence based optimization algorithm inspired by the behaviour of honey bee foraging. This algorithm is adapted to search the protein conformational search space to find the lowest free energy conformation. Interestingly, the algorithm was able to find the lowest free energy conformation for a test protein (i.e. Met enkephaline) using ECEPP/2 force fields.

1. Introduction

Proteins play a vital role in the biological processes of the human body. They are considered as the building blocks of the human body. They perform many biological functions such as the Haemoglobin protein which is responsible for the oxygen transportation in the blood. Significantly, a protein can only be able to perform its biological function when it folds into its tertiary structure. This state is called the biological active state or the native state.

The protein structure prediction problem is regarded as a grand challenge and is one of the great puzzling problems in computational biology [1]. It is how to get the structure of the protein given only its sequence.

This problem can be solved experimentally using experimental methods such as NMR and X-ray Crystallography. Experimental methods are the main source of information about protein structure and they can generate more accurate results. However, they are also time consuming where the determination of the structure of a single protein can take months and they are expensive, laborious and need special instruments as well. Moreover and due to some limitations in the experimental methods, it is not always feasible to determine the protein structure experimentally which results in creating a big gap between the number of protein sequences and known protein tertiary structures. In order to bridge this gap, other methods are much needed to determine the protein structure.

Scientists from many fields have worked to develop theoretical and computational methods which can help provide cost effective solutions for the protein structure prediction problem. Accordingly, the best existing alternative is using computational methods which can offer cost effective solutions.

Computational methods can be traditionally divided into three approaches: Homology Modelling, Threading and Ab initio. In Homology Modelling and Fold Recognition methods, the prediction is performed using the similarities between the target protein sequence and the sequences of already solved proteins structures. So, these methods are limited to predict the structure of proteins which belong to protein families with known structures. On the contrary, Ab initio methods are not limited to protein families with at least one known structure. They are based on the Anfensin [3] hypothesis which states that the tertiary structure of the protein is the conformation with the lowest free energy.

To predict the protein structure using Ab initio method, the problem is formulated as an optimization problem with the aim to find the lowest free energy conformation. In order to perform that, protein conformation must be represented in a proper representation. This representation is ranged from all atoms representation to simplified representation. Then, an energy function is used to calculate the conformation energy and a conformational search algorithm is utilized to search the conformation

978-0-7695-3648-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AMS.2009.47

search space to find the lowest free energy conformation.

Conformational search algorithms explore the protein conformational search space with a major goal to find the lowest free energy conformation [2]. Searching the protein conformational search space is a grand challenge in protein tertiary structure prediction due to the large number of possible conformations and the local minima problem. In general, if a protein has n atoms, the degree of freedom is $3n$-6. And if a protein with 100 amino acids and each amino acid has 20 atoms, the number of degree of freedom is equal to ([(100*20)*3]-6=5994) [3].

If we consider the torsion angles representation of the protein, take for instance 5 angles per amino acid and consider five values for each angle, the number of possible conformations is 25^{100}. It is impractical to test all the feasible conformations to find lowest free energy conformation. Therefore, success in the prediction of the protein tertiary structure is dependent on the efficiency of search method over different conformations without testing all conformational possibilities [4].

Recently, researchers are initiated to study the behaviour of social insects in an attempt to use the Swarm Intelligence concepts to develop algorithms that have the ability to search the solution search space of the problem in a way similar to the foraging search by colony of social insects.

A new field in computational science is now emerging which is based on the inspiration of nature and biology to propose computational algorithms that can be applied to solve a wide diversity of complex optimization problems. These algorithms can, under suitable conditions, outperform the existing conventional algorithms [5].

Swarm intelligence which is a new active research area [5] belongs to this kind of algorithms. Swarm intelligence algorithms based on social insects such as ants and bees are successful in solving combinatorial optimization problems and begin to show their power and effectiveness in many applications [6]. This success is due to the computational advantages [7]. Theoretically, every possible problem can be solved using a swarm-based system [32].

Using the principles of honey bees colony, the difficult combinatorial optimization problems such as protein tertiary structure prediction can be solved. Artificial Bee Colony algorithm (ABC) is a swarm based algorithm proposed by Karaboga [8]. It simulates the behaviour of honey bee foraging. In this paper, ABC algorithm is adapted to search the protein conformational search space in order to find the lowest free energy conformation.

The layout of this paper is organized as follows: An overview of honey bees in nature has been given in section 2. Description of honey bees foraging is expounded in Section 3. Section 4 is devoted to present the Artificial Bee Colony algorithm. The adaptation of Artificial Bee Colony algorithm to the protein conformational search is described in Section 5. Experiments and results are shown in Sections 6 and 7. Finally, the conclusion is presented in Section 8.

2. Honey bees in nature

Honey bees are the most beneficial and the most well studied insects. They live in hives around the world in very well organized colonies. Honey bees colonies are characterized by the division of labour where specific bees do specific jobs. There are no idle bees, the work in the hive is load balanced. Also, it is characterized by the communication on the individual and group level and cooperative behaviour.

The honey bees colony contains around 10,000 to 60,000 bees [9]. The honey bees colony may contain one or more than one queen. In the first case, it is called monogynous while in the second one it is called polygynous. Besides the queen, the colony contains drones, workers and broods. The queen specializes in egg laying. It lies around 1500-2000 eggs per day and in some cases it may lay 3000 eggs per day. The drones have only one job to do, which is mating with the queen. The drone is haploid, in that, it has only the half number of chromosomes. The workers take care of the broods and forage for nectar. The broods are the children of the colony and they arise from fertilised or unfertilised eggs. When it grows, the fertilised egg becomes a worker or a queen and the unfertilised one becomes a drone (Figure 1).

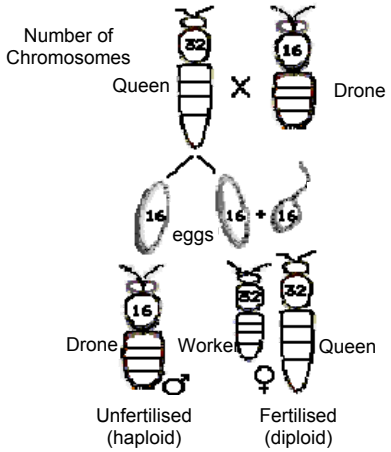

Figure 1. Honey bees genetics [10]

Honey bees colony can be established in two ways [11]. The first way is called the independent

founding. In this way, the colony is established starting with one or more reproductive females who start building the hive and laying the eggs. The second way is called swarming. In this way, a single queen or more with a group of workers leave the original colony when it becomes too large, so they start searching for another place to build a new hive for the colony.

Two activities in the life of honey bees colony attract the computer scientists, the foraging and the process of reproduction. In the next section, the honey bees foraging behaviour will be explained.

3. Honeybees foraging

A Honeybee colony must accumulate sufficient food in summer to consume it during the winter season where food sources become scarce [12]. Foraging for food is done by the foraging bees which are the adult worker bees and they represent around 25% of workers [12]. The honeybees colony coordinates its foraging activity in an efficient way. It sends the foragers in multiple directions simultaneously to cover a large search area [13, 14]. Honeybees has the ability to find the food sources even when they are far away from the hive and can return back to the hive without going astray. Honeybees colony concentrates its search and selects the most profitable nectar sources in the field among the available sources and adjusts the searching pattern precisely [6].

A forager bee starts foraging processes by searching for a food source without any guidance from other bees. In this case it is called a scout. It searches randomly for a food source and moves from one source to another. When it finds one, it collects an amount of nectar, evaluates the food source based on the quality of food, and the distance from the hive and the amount of energy consumed during the foraging. A honeybee has an efficient memory which enables it to memorise the location of the food source. When it returns back and arrives the hive, it unloads the nectar. After that, the forager bee has to take one decision out of the following three decisions:

1) Perform the waggle dance to recruit more foragers to the same food source.
2) Abandon the food source so no more bees forage it.
3) Return to foraging directly.

If it decides to share the information about the food source with other bees, it performs a waggle dance in the dance floor. The dance floor is an area near the hive entrance. Bees attending in the dance floor can understand the waggle dance and gain the needed information to start foraging from the food source. Based on the quality of the food sources, more bees will forage from the high quality food sources. A forager bee may abandon the food source if its quality becomes low, or it may return to forage without telling other bees about its source.

The goal of the honeybees colony in foraging is to visit the rich food sources in order to gain maximum level of food.

4. The artificial bee colony algorithm

Karaboga [8] examined the foraging behaviour of honeybee colony and proposed an algorithm which simulates this behaviour to solve multidimensional and multimodal optimization problems [8]. This algorithm is called Artificial Bee Colony (ABC). The algorithm was used after that to solve numerical function optimization problems [15], constrained optimization problems [16], and for training feed-forward neural networks [17, 18]. The results showed the power of the proposed algorithm and its robustness.

The artificial colony consists of three types of artificial bees. The employed bees, the onlooker bees and scouts. The colony is divided into two equal parts. The first part contains the employed bees and the second part contains the onlooker bees. The number of food sources is equal to the number of employed bees. Each food source is considered as a possible solution to the optimization problem and the nectar amount is the fitness of the solution.

The algorithm starts by generating SN randomly distributed solutions (food sources). Where SN denotes to the size of artificial colony. Each food source is a D-dimensional vector, where D is the number of optimization variables.

The employed bee modifies the solution in its memory and calculates its fitness. The modification is done using equation (1). In equation (1) $\kappa \in \{1,2,...,SN\}$ and $j \in \{1,2,...,D\}$ are randomly chosen indexes. κ has to be different from i. ϕ_{ij} is a random number between [-1,1]. It controls the production of a neighbourhood food source position around x_{ij}. The employed bee saves the new solution if its fitness is better than the old one. After all employed bees complete the search process, they share the information about their solutions with the onlookers. An onlooker chooses a food source based on probability related to the fitness of the solution. This probability is calculated using equation (2), fit_i is the fitness value of the solution i.

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (1)$$

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (2)$$

The onlooker bee modifies the chosen solution and calculates its fitness. As in employed bees, the onlooker bee saves the new solution if its fitness is better than the old one or it keeps the old one.
The food source which is not improved for a number of iteration (limit) is abandoned. A new food source is created and replaced the abandoned food source.

5. Protein conformational space search using artificial bee colony algorithm

This section is devoted to describe how the Artificial Bee Colony algorithm was adapted to solve the protein conformational search problem in order to find the lowest free energy conformation.

5.1 Protein conformation representation

Each amino acid consists of two parts: the main chain and the side chain (Figure 2). The main chain torsion angles are: φ, ψ and ω. The side chain torsion angles are χn.

As the overall structure of proteins can be described by their backbone [19, 20] and side chain torsion angles, the tertiary structure of a protein can be obtained by rotating the torsion angles around the rotating bonds [21]. So, the protein conformation is represented as a sequence of the torsion angles [22]. This representation is a common protein conformation representation and it is widely used in protein conformational search algorithms [23-25].

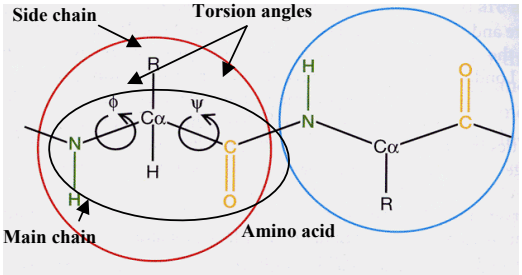

Figure 2. Amino acid [26]

In the torsion angles representation, each conformation is represented as an array of real values. These values are the values of the amino acid torsion angles. The length of the array represents the number of torsion angles of the protein. Generating conformations is done by changing the values of the torsion angles randomly.

5.2 Energy function

The protein energy function is the objective function and the torsion angles are the variables. The conformation energy is calculated using ECEPP/2 force fields which it is implemented as a part of the SMMP (Simple Molecular Mechanics for Proteins) [27-29].

5.3 The algorithm

This section describes the adaptation of the Artificial Bee Colony algorithm to search the protein conformational space to find the lowest fee energy conformation (see Figure 3.).

Each food source represents a protein conformation. *N* food sources (conformations) are initialized randomly. Then, the food sources are evaluated using energy function. The food sources are improved using one of four types of improvements adapted from [30]. In the first type, the conformation is improved by changing the all torsion angles (main and side chain) randomly one by one and calculate the energy of the conformation. Then minimize the lowest conformation energy using minimization procedure in the SMMP package. In the second type, the side chain torsion angles are changed randomly and the main chain angles remain fixed. Then, the lowest conformation is minimized. The Third type is similar to the second type but for main chain torsion angles only. The Fourth type is mixed between the Second type and the Third type where the side chain angles are changed then the main chain angles.

After the improvement of the conformations, they are sorted and the best *n*/2 are chosen to be the employed bees and the rest are the onlooker bees. Onlooker bees choose employed bees based on their energy values and generate new conformations by using hpilod crossover [9] then apply the improvement to the conformation. Onlooker bee saves the new conformation if its energy is lower than the old energy unless it keeps the old one.

Similarly, employed bees search in their neighbour and keep the new conformations if their energies are lower than the old ones.

Checking for duplicated conformations is performed every *m* iterations. The duplicated conformation is replaced by generating a random conformation.

These steps are repeated until the maximum iteration number is met.

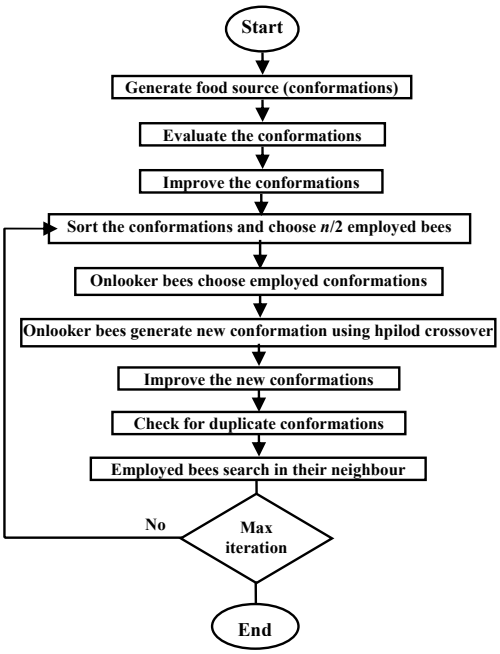

Figure 3. ABC flowchart

6. Experiments

The algorithm was implemented using visual C++ on AMD Athlon (tm) 64 processor 3000+, 2.0 GHz and 1.25 GB of RAM. The machine runs windows XP professional operation system. Consequently, the SMMP package was converted from FORTRAN code into C++ code with the necessary modifications. ECEPP/2 force field was used to calculate the energy.

The algorithm was applied to find the lowest free energy conformation of Met-enkephaline, i.e. a small protein which is extensively used to test the conformational search methods. It consists of 5 amino acids with 24 torsion angles.

Four types of experiments were performed. Each experiment used one type of improvement.

The number of employed bees is set to 10 and the number of onlooker bees is set to 10. The number of maximum iterations is set to 1000.

7. Results

One hundred independent runs were performed for each experiment. The results of the experiments are given in Table 1.

The lowest free energy conformation found in all four types of experiments was -12.910121 kcal/mol which is the same result reported in [23] and is lower than the result reported in [31]. From Table 1, we can see that first type of improvement gained the best success rate (84%) followed by the third type (%80), forth type (77%) and second type (%65). This is because the first type considers all torsion angles in the same time so it covers the conformational search space efficiently. Whereas the third type considers the main chain torsion angles only. The forth type considers main chain and side chain torsion angles separately in the improvement. The second type considers only the side chain torsion angles in the improvement.

In terms of computation time, the second type gained the best average time because it considers small number of torsion angles (9 side chain torsion angles).

Table 1. Results of the 4 experiments

	Success rate	Best	Worst	Average	Average time
1	84%	-12.910121	-10.91010	-12.64999	623.988s
2	65%	-12.910121	-10.00296	-12.40344	588.308s
3	80%	-12.910121	-10.36007	-12.61972	600.166s
4	77%	-12.910121	-10.91010	-12.59192	711.965s

8. Conclusion

This paper adapted the swarm intelligence based algorithm, i.e. the Artificial Bee Colony algorithm to search the protein conformational search space to find the lowest free energy conformation. The results indicated that the algorithm was able to find the lowest free energy conformation of -12.910121 kcal/mol usign ECEPP/2 force field. Best results were gained using first type of improvement. Further work is needed to compare the performance of the alogrithm on larger proteins and also to improve the performace of the algorithm by parallelzing and comparing the performace of the algorithm with other existing algorithms for protein conforamtional search.

9. Acknowledgement

This research is supported by FRGS research grant for project "Parallel Conformational Search Algorithm for Protein Tertiary Structure Prediction Using Honey Bee Colony Optimization" [203/PKOMP/671184].

10. References

1.Chiu, T.-L. and R. Goldstein, *Optimizing energy potentials for success in protein tertiary structure prediction.* Folding and Design, 1998. **3**(3): p. 223-228.

2.Zhang, H., *Protein Tertiary Structures: Prediction from Amino Acid Sequences*, in *Encyclopedia of Life Sciences*. 2002.

3.Schulze-Kremer, S., *Genetic Algorithms and Protein Folding*, in *Protein Structure Prediction Methods and*

Protocols, D. Webster, Editor. 2000, Southern Cross Molecular Ltd. : Bath, UK. p. 175-222.
4.Zhou, Y. and R. Abagyan, *Efficient Stochastic Global Optimization for Protein Structure Prediction*, in *Rigidity Theory and Applications*. 2002. p. 345-356.
5.Yang, X.-S., *Engineering Optimizations via Nature-Inspired Virtual Bee Algorithms*, in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach*. 2005. p. 317-323.
6.Lucic, P., *Modelling Transportation Problems Using Concepts of Swarm Intelligence and Soft Computing*, in *Faculty of the Virginia Polytechnic Instisute and State University*. 2002: Virginia.
7.Hyeong Soo, C., *Converging Marriage in Honey-Bees Optimization and Application to Stochastic Dynamic Programming.* Journal of Global Optimization, 2006. **35**(3): p. 423-441.
8.Karaboga, D., *An Idea Based On Honey Bee Swarm For Numerical Optimization*, in *Technical Report-TR06*. 2005, Erciyes University, Engineering Faculty, Computer Engineering Department.
9.Abbass, H.A. *MBO: Marriage in Honey Bees Optimization -A Haplometrosis Polygynous Swarming Approach*. in *Congress on Evolutionary Computation*. 2001. Seoul, Korea.
10.http://members.aol.com/queenb95/genetics.html#anchor173808.
11. Dietz, A. *Bee Genetics and Breeding*. in *Evolution*. 1986: Academic Press Inc.
12. Sunil, N. and T. Craig, *On Honey Bees and Dynamic Server Allocation in Internet Hosting Centers*. 2004, Sage Publications, Inc. p. 223-240.
13. Pham, D.T., et al. *The Bees Algorithm, A Novel Tool for Coplex Optimisation Problems*. in *2nd International Virtual Conference on Intelligent Production Machines and Systems (IPROMS 2006)*. 2006: Oxford: Elsevier.
14. Seeley, T.D., *The Wisdom of the Hive The Social Physiology of Honey Bee Colonies* 1995: Harvard University Press.
15. Karaboga, D. and B. Basturk, *A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm.* Journal of Global Optimization, 2007. **39**(3): p. 459-471.
16. Karaboga, D. and B. Basturk, *Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems*, in *Foundations of Fuzzy Logic and Soft Computing*. 2007. p. 789-798.
17. Karaboga, D. and B. Akay. *Artificial Bee Colony (ABC) Algorithm on Training Artificial Neural Networks*. in *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*. 2007.
18. Karaboga, D., B. Akay, and C. Ozturk, *Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks*, in *Modeling Decisions for Artificial Intelligence*. 2007. p. 318-329.
19. Betancourt, M.R. and J. Skolnick, *Local Propensities and Statistical Potentials of Backbone Dihedral Angles in Proteins*. Journal of Molecular Biology, 2004. **342**(2): p. 635-649.
20. Dayalan, S., S. Bevinakoppa, and H. Schroder, *A dihedral angle database of short sub-sequences for protein structure prediction*, in *Proceedings of the second conference on Asia-Pacific bioinformatics - Volume 29*. 2004, Australian Computer Society, Inc.: Dunedin, New Zealand.
21. Garduno-Juarez, R. Morales, and L. B**.**, *A genetic algorithm with conformational memories for structure prediction of polypeptides.* Journal of biomolecular structure & dynamics, 2003. **21**(1): p. 65-87.
22. Vengadesan, K. and N. Gautham, *A New Conformational Search Technique and Its Applications*. 2006.
23. Zhan, L., J.Z.Y. Chen, and W.-K. Liu, *Conformational Study of Met-Enkephalin Based on the ECEPP Force Fields*. 2006. p. 2399-2404.
24. L. B. Morales, R.G.-J.J.M.A.-A.F.J.R.-C., *A parallel tabu search for conformational energy optimization of oligopeptides*. 2000. p. 147-156.
25. Guan, X., et al. *Protein structure prediction using hybrid AI methods*. in *Artificial Intelligence for Applications, 1994., Proceedings of the Tenth Conference on*. 1994.
26. Mount, D.W., *Bioinformatics: Sequence and Genome Analysis*. 2004, NY: Cold Spring Harbor Laboratory Press.
27. Eisenmenger, F., et al., *An enhanced version of SMMP—open-source software package for simulation of proteins.* Computer Physics Communications, 2006. **174**(5): p. 422-429.
28. Eisenmenger, F., et al., *[SMMP] A modern package for simulation of proteins*. Computer Physics Communications, 2001. **138**: p. 192-212.
29. *http://www.smmp05.net*. 2007.
30. Meirovitch, H., et al., *A Simple and Effective Procedure for Conformational Search of Macropolecules: Application to Met- and Leu-Enkephalin.* Journal of physical chemistry, 1994. **98**: p. 3.
31. Jooyoung Lee, H.A.S.S.R., *New Optimization Method for Conformational Energy Calculations on Polypeptides: Conformational Space Annealing*. 1997. p. 1222-1232.