Protein Conformational Search Using Bees 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 perform many biological functions in the human body. The structure of the protein determines its function. In order to predict the protein structure computationally, protein must be represented in a proper representation. To this end, an energy function is used to calculate its energy and a conformational search algorithm is used to search the conformational search space to find the lowest free energy conformation. In this paper, the Bees Algorithm, i.e. a Swarm Intelligence based algorithm inspired by the foraging behaviour of honey bees colony, is adapted to search the protein conformational search space. The algorithm was able to find the lowest free energy conformation of Met-enkephaline using ECEPP/2 force fields.

1. Introduction

Proteins are natural substances consist of any number of the 20 naturally occurring amino acid types. Proteins play a crucial role in the biological processes inside the human body. Proteins can perform their functions only when they fold into their tertiary structure (biologically active or the native state). The protein structure prediction problem, which is regarded as a grand challenge and one of great unsolved problems in computational biology [2] is how to get the structure of the protein given its sequence.

Protein structure can be determined using experimental methods such as Nuclear Magnetic Resonance (NMR) and X-ray crystallography. Although they produce accurate structures, they are time consuming and expensive. Moreover, due to some limitations in the experimental methods, it is not always feasible to determine the protein structure experimentally. There is a big gap between the number of protein sequences and known protein tertiary structures. In order to bridge this gap, other methods are needed to determine the protein structure. Scientists from many fields work to develop theoretical and computational methods which help to provide cost effective solutions for the protein structure prediction problem.

Traditionally, computational protein structure prediction methods are divided into three areas. They are: The Homology Modelling, Fold Recognition and Ab initio. As Homology Modelling and Fold Recognition methods are based on the similarities between the target protein sequence and the sequences of already solved proteins structures, they are limited to predict the structure of proteins belong to protein families with known structures. Conversely, Ab intio methods are not limited to protein families with at least one known structure. Ab initio methods are the only methods that can be used to model any protein sequence.

Ab initio methods are based on the thermodynamics hypothesis formulated by Anfinsen [3]. Anfinsen proposed that the tertiary structure of a protein in its physiological environment is the conformation with lowest free energy [3].

In order to predict the protein tertiary structure using Ab initio method, the problem is formulated as an optimization problem. To solve this problem, protein must be represented in a proper representation. Its energy is then calculated using an energy function suitable for the representation and a conformation search algorithm is used to search protein conformation search space.

Conformational search algorithms explore the protein conformational search space and look for the lowest free energy conformation [4]. A major obstacle to predict the protein tertiary structure using computational methods is the challenge of searching the protein conformational search space [5] due to the large number of possible conformations and the local

978-0-7695-3136-6/08 $25.00 © 2008 IEEE
DOI 10.1109/AMS.2008.65

minima problem. In general, if a protein has n atoms, the degree of freedom is $3n$-6. 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) [6]. If we consider the torsion angles representation of the protein, take 5 angels per amino acid and consider five values for each angle, the number of possible conformations is 25^{100}. It is impossible to test all the feasible conformations to find the 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 [7].

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. Using the principles of honey bees colony, the difficult combinatorial optimization problems such as protein tertiary structure prediction can be solved.

Bees Algorithm [1] is a Swarm Intelligence based algorithm inspired by the foraging behaviour of honey bees colony. It was used to train the Learning Vector Quantisation (LVQ) neural network for control chart pattern recognition [8]. It gained better results than the standard LVQ training algorithm. It was used also to train Multi-layered Perceptron (MLP) network to recognise different patterns in control charts [9]. The algorithm was used to the optimization of neural networks for wood defect detection. It was used to identify the defects in plywood veneer [10].

In this paper, the Bees 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. While Section 4 is devoted to present the Bees Algorithm. The adaptation of Bees Algorithm to the protein conformational search is described in Section 5. Experimental results are shown in Section 6 and conclusion in Section 7.

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 [11]. 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 fertilise or unfertilised eggs. When it grows, the fertilised egg becomes a worker or a queen and the unfertilised one becomes a drone (Figure 1).

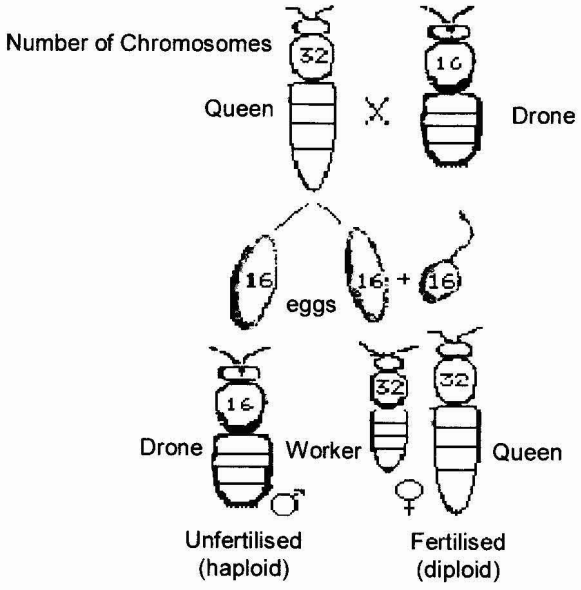

Figure 1. Honey bees genetics [12]

Honey bees colony can be established in two ways [13]. 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 lay 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. They 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. Honey bees foraging

A honey bee colony must accumulate sufficient food in summer to consume it during the winter season where food sources become scarce [14]. Foraging for food is done by the forager bees which are the adult worker bees and they represent 25% of workers [14]. The honey bees colony coordinates its foraging activity in an efficient way. It sends the foragers in multiple directions simultaneously to cover a large search area [1, 15]. Honey bees 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. Honey bees colony concentrates its search and selects the most profitable nectar sources in the field among the available sources and adjusts the searching pattern precisely [16].

Forager bee starts foraging processes by deciding to start 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 food 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. Honey bee has an efficient memory which enables it to memorise the location of the food source. Then, it returns back to the hive. When it arrives, 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. Forager bee may abandon the food source if its quality becomes low, or it may return to forage without telling the other bees about its source.

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

Algorithms which are inspired by the foraging behaviour of honey bees are applied to solve problems in networks routing like BeeHive [17] a fault tolerant, scalable routing algorithm and BeeAdHoc [18] an energy efficient routing algorithm for mobile ad hoc networks. Moreover, honey bee was used in solving the Travelling Salesman Problem [16].

4. The bees algorithm

The pseudo code of the basic Bees Algorithm [1] is shown in Figure 2. The algorithm parameters are:

- number of scout bees *n*.
- number of site selected out of n visited sites *m*.
- number of best sites out of m selected sites *e*.
- number of bees recruited for the other *m-e* selected sites *nsp*.
- initial size of patches *ngh*.

In step 1, the algorithm starts by initializing *n* scout bees randomly. The sites visited by the scout bees are evaluated using a fitness function in step 2. In steps 4 and 5 and based on the fitness value, good bees are selected and the sites visited by them are chosen for neighbourhood search. In step 6, more bees will be assigned for the chosen good sites. The best bee in each patch is chosen to form the next bee population in step 7. In step 8, the remaining bees in the population are assigned randomly searching for new potential solutions. These steps are repeated until the stopping criterion is met.

```
1.  Initialise population with random solutions.
2.  Evaluate fitness of the population.
3.  While (stopping criterion not met)
    // forming new bee population.
4.  Select elite bees.
5.  Select sites for neighbourhood search.
6.  Recruit bees around selected sites and evaluate fitness.
7.  Select the fittest bee from each site.
8.  Assign remaining bees to search randomly and
    evaluate their fitness.
9.  End While.
```

Figure 2. Pseudo code of bees algorithm [1]

5. Bees algorithm for protein conformational search

This section is devoted to describe how the Bees Algorithm was adapted to the protein conformational search problem in order to find the conformation with lowest energy.

5.1 Protein conformation representation

Each amino acid consists of two parts: the main chain and the side chain (Figure 3). The main chain torsion angles are: φ, ψ and ω. The side chain torsion angles are χ_1-χ_8.

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] .

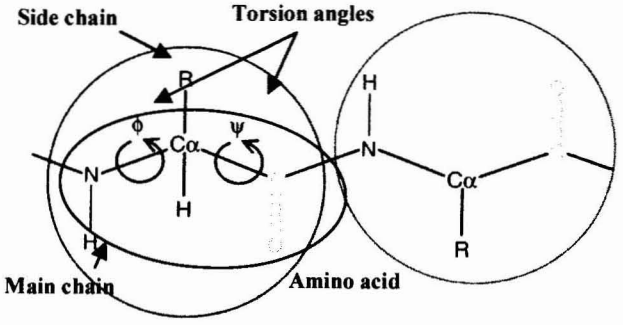

Figure 3. Amino acid [26]

Figure 4 below shows 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.

φ ψ ω	φ ψ ω	φ ψ ω	φ ψ ω	φ ψ ω
X_1-X_8	X_1-X_8	X_1-X_8	X_1-X_8	X_1-X_8
aa1	aa2	aa3	aa4	aa5

Figure 4 Torsion angles representation

5.2. Energy function

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 Bees Algorithm to search the protein conformational space to find the lowest fee energy conformation.

Each food source represents a protein conformation. A number of food sources (conformations) are initialized randomly. Scout bees start searching for food sources. They evaluate the food sources using the energy function. The food sources are sorted based on the energy value. Conformations with low energy values are chosen to the neighbourhood search. More bees will be assigned to these conformations. The rest of the scouts will search the conformational space randomly by generating random conformations. The best conformations found from each patch and the random conformations form the population for the next round of search. These processes continue until there is no improvement in the found solution or when its maximum number of iterations is reached.

6. Experiment

The proposed algorithm was implemented using visual C++. 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. The lowest free energy conformation found was -12.91 kcal/mol which is the same result reported in [23] and is lower than the result reported in [30] (-11.71 kcal/mol). The torsion angles of the lowest free energy conformation are listed in Table 1. The lowest conformation was visualized (Figure 5) using the TINKER software tools for molecular Design version 4.2 of June 2004 [31].

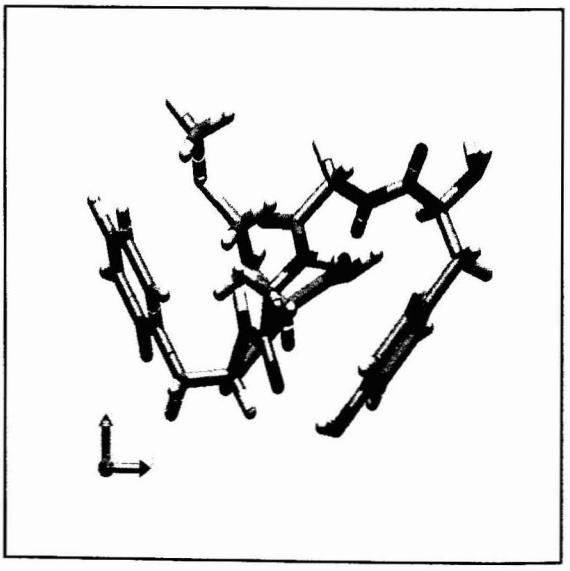

Figure 5. The lowest free energy conformation of Met-enkephaline

	Torsion	This work & [23]	[30]
Tyr1	χ_1	-172.6	-173.2
	χ_2	-101.3	79.4
	χ_6	14.1	166.4
	φ	-85.8	-83.5
	ψ	156.2	155.8
	ω	-176.9	-177.2
Gly2	φ	-154.5	-154.3
	ψ	83.7	86.0
	ω	168.6	168.5
Gly3	φ	83.7	82.9
	ψ	-73.9	-75.1
	ω	-170.1	-170.0
Phe4	χ_1	58.8	58.8
	χ_2	-85.4	94.6
	φ	-137.0	-136.9
	ψ	19.3	19.1
	ω	174.1	-174.1
Met5	χ_1	52.8	52.9
	χ_2	175.3	175.3
	χ_3	-179.8	-179.8
	χ_4	61.4	-58.6
	φ	-163.4	-163.5
	ψ	160.4	161.2
	ω	-179.7	-179.8

Table 1. Torsion angles of the lowest free energy conformation

7. Conclusion

In this paper, the swarm intelligence based algorithm, i.e. the Bees Algorithm, was adapted to search the protein conformational search space to find the lowest free energy conformation. The algorithm is inspired by the foraging behavoiur of honey bees colony. The result indicates that, the algorithm was able to find the lowest free energy conformation of -12.91 kcal/mol usign ECEPP/2 force field. Although, the work in this paper reinforce the ability of the proposed algorithm to find the lowest free energy conformation. Further work is needed to compare the performance of the alogrithm on larger proteins and also to compare the performace of the algorithm with other existing algorithms for protein conforamtional search.

8. References

1. 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.
2. 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.
3. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-230.
4. Zhang, H., *Protein Tertiary Structures: Prediction from Amino Acid Sequences*, in *Encyclopedia of Life Sciences*. 2002.
5. Chan, H.S. and K.A. Dill, *The protein folding problem.* Physics Today, 1993: p. 24-32.
6. 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.
7. Zhou, Y. and R. Abagyan, *Efficient Stochastic Global Optimization for Protein Structure Prediction*, in *Rigidity Theory and Applications*. 2002. p. 345-356.
8. Pham, D.T., et al. *Application of the Bees Algorithm to the Training of Learning Vector Quantisation Networks for Control Chart Pattern Recognition.* in *Information and Communication Technologies (ICTTA'06).* 2006. Syria.
9. Pham, D.T., et al. *Optimisation of the Weights of Multi-Layered Perceptrons Using the Bees Algorithm.* in *IMS 2006: 5th International Symposium on Intelligent Manufacturing Systems.* 2006. Sakara, Turkey.
10. Pham, D.T., et al. *Optimising Neural Networks for Identification of Wood Defects Using the Bees Algorithm.* in *Industrial Informatics, 2006 IEEE International Conference on.* 2006.
11. Abbass, H.A. *MBO: marriage in honey bees optimization-a Haplometrosis polygynous swarming approach.* 2001. Seoul, Korea.
12. 2007 [cited; Available from: http://members.aol.com/queenb95/genetics.html#anchor173808.
13. Dietz, A. *Bee Genetics and Breeding.* in *Evolution.* 1986: Academic Press Inc.
14. Sunil, N. and T. Craig, *On Honey Bees and Dynamic Server Allocation in Internet Hosting Centers.* 2004, Sage Publications, Inc. p. 223-240.
15. Seeley, T.D., *The Wisdom of the Hive The Social Physiology of Honey Bee Colonies* 1995: Harvard University Press.

16. 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.
17. Wedde, H.F., M. Farooq, and Y. Zhang, *BeeHive: An Efficient Fault-Tolerant Routing Algorithm Inspired by Honey Bee Behavior*, in *Ant Colony, Optimization and Swarm Intelligence*. 2004. p. 83-94.
18. Wedde, H.F., et al., *BeeAdHoc: an energy efficient routing algorithm for mobile ad hoc networks inspired by bee behavior*, in *Proceedings of the 2005 conference on Genetic and evolutionary computation*. 2005, ACM Press: Washington DC, USA.
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. R, G.-J. and M. LB., *A genetic algorithm with conformational memories for structure prediction of polypeptides.* Journal of biomolecular structure & dynamics, 2003. **21**(1): p. 65-87.
22. K. Vengadesan, N.G., *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 [cited.
30. Jooyoung Lee, H.A.S.S.R., *New optimization method for conformational energy calculations on polypeptides: Conformational space annealing*. 1997. p. 1222-1232.
31. *THINKER. Software Tool For Molecular Design. Version 4.2* June 2004.