Proceedings of the 3rd IMT-GT Regional Conference on Mathematics, Statistics and Applications
Universiti Sains Malaysia

SWARM INTELLIGENCE BASED PROTEIN CONFORMATIONAL SEARCH ALGORITHM

[1]Hesham Awadh Abdallah Bahamish, [2]Rosni Abdullah, [3]Rosalina Abdul Salam
[1,2,3] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Pulau Pinang
e-mail: [1]hesham@cs.usm.my, [2]rosni@cs.usm.my, [3]rosalina@cs.usm.my

Abstract. *There is no doubt of the role that proteins play in the biological processes inside the human body. Proteins can perform their function only when they fold into their tertiary structure. The thermodynamics hypothesis formulated by Anfinsen stated that the tertiary structure of a protein in its physiological environment is the conformation with the lowest free energy. To predict the tertiary structure of the protein using computational methods, protein must be represented in a suitable representation. An efficient energy function must be used to calculate the protein energy, and then a conformational search algorithm must be applied to find the lowest free energy conformation. Swarm intelligence belongs to a class of algorithms inspired by nature and biology. In this paper the Marriage in honey bee optimization (MBO), a swarm intelligence based algorithm inspired by the process of reproduction in honey bees is adopted to search the protein conformational search space to find the lowest energy conformation of Met-enkephaline. The lowest free energy conformation found was -12.4286 kcal/mol.*

1 Introduction

Proteins are important and essential for the life of the human body. They are the building blocks of the human body which perform many biological functions. Protein can be described in four structural levels: primary, secondary, tertiary and quaternary. The function of the protein is related to its tertiary structure. There is a real need to identify the functions of all proteins [2], therefore the study and prediction of the protein tertiary structure has a huge importance and it is one the great unsolved problems in computational biophysics[3].

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, the protein structure can not be always determined experimentally. With these limitations, it is not possible to determine all the protein structure experimentally [4]. There is a big gap between the number of protein sequences and known protein tertiary structures. To bridge this gap other ways to determine the structure of protein need to be explored. Scientists from many fields work to develop theoretical and computational methods which provide a cost effective solution to the protein structure prediction problem.

Traditionally, computational protein structure prediction methods are divided into three areas depending on the similarity of the target protein to proteins of known structure: The Homology Modelling, Fold Recognition and Ab initio. Ab initio is based on the thermodynamics hypothesis formulated by Anfinsen [5] "The three dimensional structure (tertiary structure) of a native protein in its physiological environment is one in which free energy of the whole system is the lowest".

Based on Anfinsen hypothesis the protein structure prediction problem is modeled as an optimization problem. To solve this problem, protein should be represented in a proper representation, suitable energy function to this representation is used to calculate the energy and a conformational search algorithm is used to search the protein conformational space.

Protein conformational search methods explore the protein conformational search space and look for the lowest free energy conformation [6]. A major obstacle to predict the protein tertiary structure using computational methods is the challenge of searching the protein conformational search space [7] 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. 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) [8]. If we consider the torsion angles representation of the protein and take 5 angels per amino acid and choose five values for each angle, the number of possible conformations is 25^{100}. It is impossible to test all possible conformations to find the lowest energy conformation. Therefore success in prediction of the protein tertiary structure depends on the efficiency of search method over different conformations and the estimation of the energy of these conformations. For this reason an important aim of any protein structure prediction methodology is to search a vast conformational space efficiently [9] and find the global minimum energy conformation without testing all conformational possibilities [10].

Many optimization algorithms are used to find the global minimum energy conformation. These algorithms can be categorized based on properties of the problem that are used and types of guarantees that the methods provide for the final solution into deterministic (e.g. Systemic search, build up method, branch and bound, diffusion equation and packet annealing) and stochastic algorithms (e.g. Genetic algorithm, simulated annealing, molecular dynamics). They can be classified into zero-order and first order algorithms depending on whether a local energy minimization step is performed after each iteration [10]. Energy minimization means derive the associated minimum energy conformation.

Swarm intelligence belongs to a class of algorithms inspired by nature and biology [11]. It is a new active research area [1] in the field of computer science [12]. Swarm Intelligence models the behaviour of social insects or animal societies and uses these models to inspire algorithms to solve real world and search problems [13]. Algorithms based on social insects such as ants and honey bees have been applied to solve many problems and showed their power and effectiveness [14].

In the honey bees colony life, two activities attracted the computer scientists, the foraging behaviour and the process of reproduction (marriage). Marriage in honeybees Optimization (MBO) is an algorithm inspired by the marriage process in honeybee colony. It was proposed by Abbass [1, 15-17] and applied to solve a special group of propositional satisfiability problems (3-SAT) [1, 15-18]. It showed good results for combinatorial optimization problems [11]. MBO algorithm has been also used to solve integrated partitioning/scheduling problem in codesign [19] and clustering problem [20, 21].

This paper presents the adaptation of MBO algorithm for protein conformational search. It is organized as follow: in section 2, the overview of honey bees in nature. In section 3, the process of marriage in honey bees is described. Section 4 describes the MBO algorithm. The MBO algorithm for protein conformational search described in section 5. Experimental results are showed in section 6 and conclusion in section 7.

2 Honey bees in nature

Honey bees are social insects which live around the world in hives in very organized colonies. They are the most beneficial insects. Honey bees are one of the most well studied social insects. It is 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 colony of honey bees may contain one or more than one queen. In the first case it is called monogynous and the second polygynous. Besides the queen, the colony contains drones, workers and broods. The queen specializes in egg laying. It lays around 1500-2000 eggs and in some circumstances it may lay 3000 eggs per day. The drones have only one job to do, mating with the queen. The drone is haploid. It only has 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). The number of bees in the colony is from 10,000 to 60,000 bees [1].

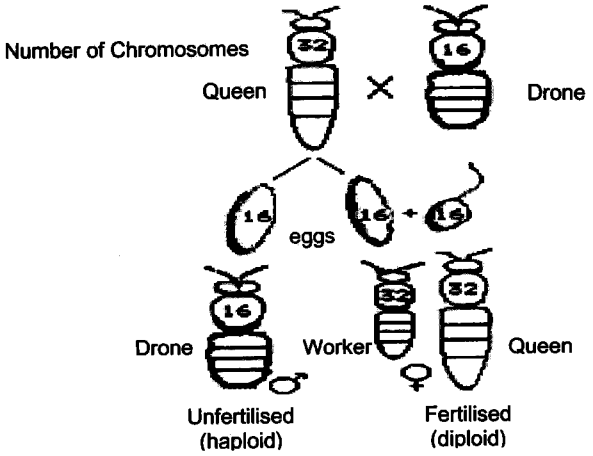

Figure 1. Honey bees genetics [22]

Honey bees colony can be established in two ways [23]. 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.

3 Marriage in honey bees

Marriage process in honey bee is difficult to study because the queen mates with the drones away from the hive which makes it hard to be observed by the scientist [1]. The marriage process in the honey bees starts when the queen perform a waggle dance and then flies in a journey called the mating flight. The drones follow the queen. The queen flies in a high speed, so the strongest and the best drones which can reach the queen participate in the mating in the air. The number of participated drones is around 7 to 20 drones [1]. The drone sperm reaches to the queen seprmatheca and accumulates there. The drone dies immediately after the mating. The queen returns to the hive after the mating flight and starts breeding the broods using a random mixture of accumulated sperm [24]. Queen uses accumulated sperm to fertilize the eggs for all its live which may span from 4 to 5 years.

4 The Marriage in honey bees Optimization

MBO is a new swarm intelligence based optimization algorithm to combinatorial optimization problems. It is inspired by the organization of bee and the behaviour of marriage in honey bee colony. It is considered as an evolutionary methaherustic algorithm which combines concepts and advantages from general methaheuristic and dedicated heuristic [19, 25].

Marriage in honey Bees Optimization (MBO algorithm) consists of two stages: The mating flight stage and broods breeding stage. In the mating flight stage, the queen mating flight is considered as a set of transitions in the search space.

The algorithm starts by initializing a set of workers. Workers are presented as specific heuristics to the problem. They are local search algorithms that improve the current solution. Then the queen genotype which is considered as a complete solution to the problem is generated randomly and evaluated using the problem fitness function. The genome can be represented as an array. The length of the array is the number of problem variables. Depending on the problem, the value of array's elements can be binary or float values. The queen genotype is improved using a randomly selected worker. This step is done for all the queens. Then for a predefined mating flights number, each queen starts its mating flights starting with a randomly initial energy and speed in [0.5,1]. It moves between different states according to its speed and energy and mates with the generated drones probabilistically according to the equation 1, where Prob (Q, D) is the probability of adding sperm of drone D to the spermatheca (set of partial solutions) of queen Q. $?(f)$ is the absolute difference between the fitness of the queen and the drone. $S(t)$ is the speed of the queen at the time t. The drone genotype is generated independently of the queen. If the drone mates successfully, its genotype (sperm) is added to the queen spermatheca and accumulated there.

$$\text{Prob}(Q, D) = e^{\frac{-\Delta(f)}{S(t)}} \quad (1)$$

After each move in the space, speed and energy of the queen are reduced using the following equations:

$$s(t+1) = \alpha * s(t) \text{ and } E(t+1) = E(t) - \gamma \quad (2)$$

α is a number]0,1[and γ is the amount of energy reduction after each transition and calculated using the equation (3) where M is the size of spermatheca (the maximum number of drones).

$$\gamma = \frac{0.5 * E(t)}{M} \quad (3)$$

The queen returns to the hive when its energy reaches zero or when its spermatheca is full. After the returning of all queens from the mating flight, the second stage, the broods breeding starts by selecting a random sperm from the queen spermatheca and crossover it with the queen genome to generate the broods. Since the drone in nature is hapilod (has only half the number of chromosomes), in the artificial model of the MBO algorithm, half of the drone genes are marked at random. So the kind of crossover used is called hapilod crossover [1]. Figure 2 explains the hapilod crossover assuming using binary representation of the solution.

Genotype	1	1	1	0	0	1	0	0
Genotype–marker	u	m	m	u	u	u	m	m

where, u and m represent an unmarked and a marked gene respectively. Therefore, the drones sperm is

	1	*	*	0	0	1	*	*
Queen	0	1	0	0	0	0	1	0
brood	1	1	0	0	0	1	1	0

Figure 2. The Hapilod Crossover [17]

Broods are mutated and improved using randomly selected worker. The worker fitness is updated based on the amount of improvement achieved on broods. After that, the less quality queens are replaced with better broods and a new mating flight starts. Figure 3 presents a generic MBO algorithm.

```
initialize workers
randomly generate the queens
apply local search to get a good queen
for a pre-defined maximum number of mating-flights
    for each queen in the queen list
        initialize energy, speed and position
        the queen moves between states
            and probabilistically chooses drones
        if a drone is selected, then
            add its sperm to the queen's spermatheca
        end if
        update the queen's internal energy and speed
    end for each
    generate broods by crossover and mutation
    use workers to improve the broods
    update workers' fitness
    while the best brood is better than the worst queen
        replace the least-fittest queen with the best brood
        remove the best brood from the brood list
    end while
end for
```

Figure 3. A generic MBO algorithm [1]

5 MBO for protein conformational search

This section describes the application of the MBO algorithm to the protein conformational search problem to find the conformation with the lowest energy.

5.1 Protein conformation representation

Protein conformation can be described by the torsion angles of the protein amino acids. Each amino acid consists of two parts: the main chain and the side chain. In the main chain, the torsion angles named: φ, ψ and ω. In the side chain torsion, the torsion angles named χ_1-χ_8. Figure 4 shows the torsion angles representation. Each conformation is represented as an array of real values which are the values of the torsion angles of the amino acid. The length of the array represents the number of torsion angles of the protein. The generation of the 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
χ_1-χ_8	χ_1-χ_8	χ_1-χ_8	χ_1-χ_8	χ_1-χ_8
aa1	aa2	aa3	aa4	aa5

Figure 4. Torsion angles representation of main chain

5.2 Energy function

Conformation energy is calculated using ECEPP/3 force fields which is a part of the SMMP (Simple Molecular Mechanics for Proteins) [26, 27].

5.3 The algorithm

In this algorithm the single queen single worker colony will be used [15]. The algorithm starts by setting the parameters: the queen spermatheca size (the number of drones), the number of broods and the number of mating flights. The queen is then initialized by generating a random conformation by giving random values to the torsion angles. The queen conformation is evaluated using the energy function. The queen speed and energy is initialized randomly in [0.5, 1]. Queen conformation is improved using the minimization function. A number of mating flights are performed by the queen. In each flight the queen moves between the different state in the conformational search space based on its speed and energy. A drone conformation is generated randomly. If the drone is better than the queen, it is accepted. If it is worse, it is accepted based on:

$$If\ exp(-\Delta(f)/T) > r$$
$$drone\ is\ accepted$$

Where: $-\Delta(f)$ is the absolute difference between the fitness of the queen and the drone, r is a randomly generated number in]0,1[, T is the speed of the queen.

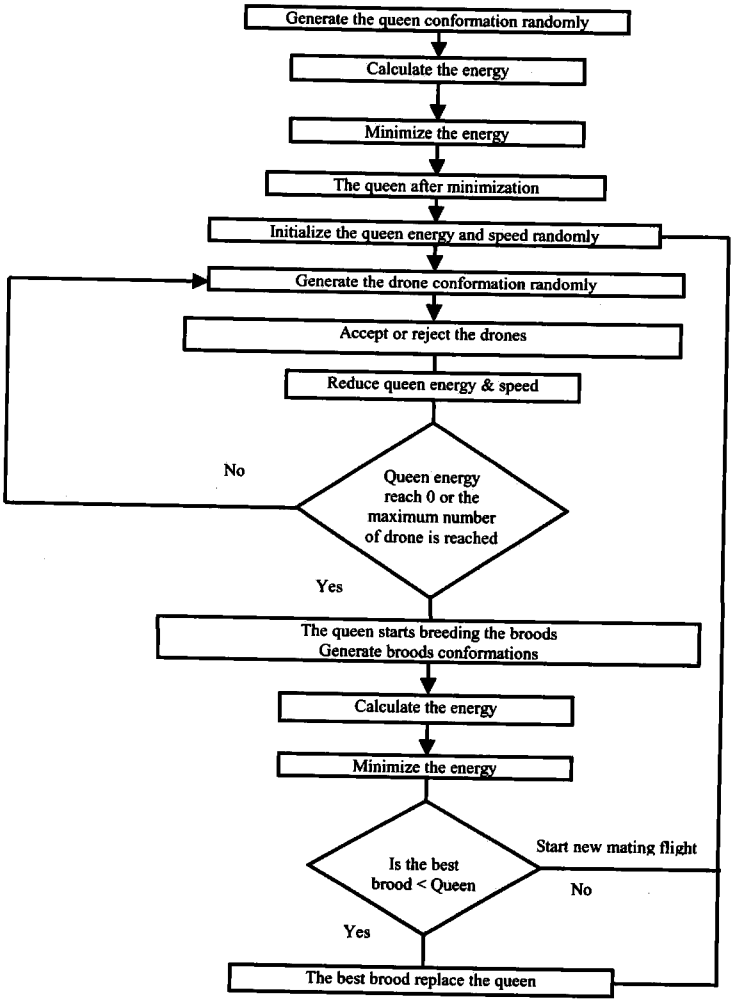

Figure 5. The flowchart of MBO algorithm for protein conformational search

When the queen energy reaches zero or when the number of participant drones reaches the predefined number. The queen starts to generate the broods conformations by selecting a random drone form its spermatcha. Then it masks half of its genes randomly and uses the halipod crossover. The energy of the generated brood is calculated and minimized. The best brood replaces the queen if its conformation energy is lower than the queen conformation energy. Then a new mating flight starts. Figure 5 shows the flowchart of the algorithm.

6 Experimental results

The algorithm was implemented using visual C++. The SMMP package were converted from Fortran code into C++ code with the necessary modifications and integrated with the code of the algorithm. ECEPP/3 force fields was used to calculate the energy.

The algorithm was applied to find the lowest conformation of Met-enkephaline a small protein which is extensively used to test the conformational search methods. It consists of 5 amino acids with 24 torsion angles. The algorithm parameters are set as following: the number of mating flights is 10, the number of drone is 20 and the number of broods is 100. 50 independent runs were performed. The lowest free energy conformation found was -12.4286 kcal/mol which is the same result reported in [28] using basin paving method. Among the 50 independent runs the algorithm was able to find the lowest free energy conformation in 32 runs, and found the values -11.70 kcal/mol in 6 runs, -11.69 kcal/mol in 7 runs and -11.27 kcal/mol in 5 runs. The torsion angles of the lowest free energy conformation are listed in Table 1.The lowest free energy conformation was visualized (Figure 6) using the TINKER software tools for molecular Design version 4.2 of June 2004.

	Torsion	value
Tyr1	χ_1	-137.2
	χ_2	-100.7
	χ_6	13.7
	φ	-83.1
	ψ	155.8
	ω	-177.1
Gly2	φ	-154.2
	ψ	85.8
	ω	168.5
Gly3	φ	83.0
	ψ	-75.0
	ω	-170.0
Phe4	χ_1	58.9
	χ_2	-86.5
	φ	-136.8
	ψ	19.1
	ω	174.1
Met5	χ_1	52.9
	χ_2	175.3
	χ_3	-179.9
	χ_4	61.4
	φ	-163.4
	ψ	160.8
	ω	-179.8

Table 1. Torsion angles of the lowest free energy conformation

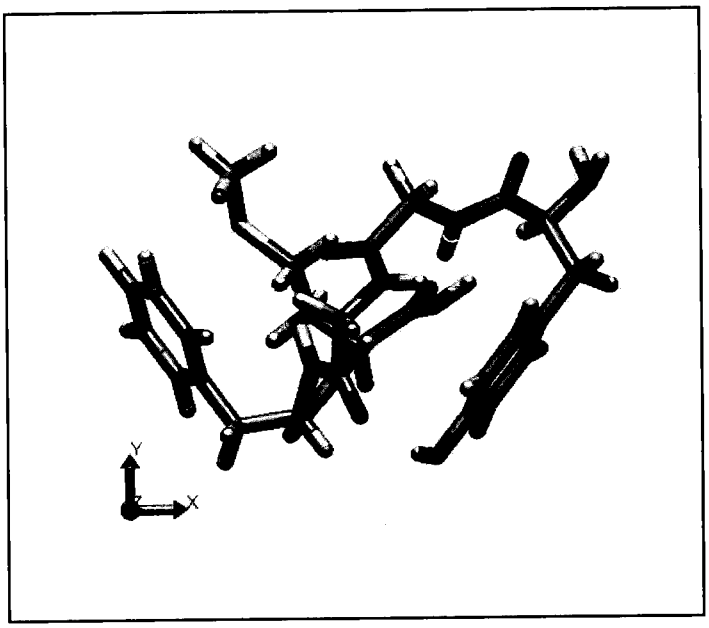

Figure 6. The lowest free energy conformation of Met-enkephaline

7 Conclusion

In this paper the swarm intelligence based algorithm MBO algorithm was adapted to search the protein conformational search to find the lowest free energy conformation. The algorithm is inspired by the marriage process in honey bees. The algorithm was able to find the lowest free energy conformation of -12.4286 kcal/mol using ECEPP/3 force fields. Further work is needed to compare the performance of the algorithm on larger proteins and to also compare the performace of the algorithm with other existing algorithms.

8 References

1. Abbass, H.A. *MBO: marriage in honey bees optimization-a Haplometrosis polygynous swarming approach.* 2001. Seoul, Korea.
2. Skolnick, J. and A. Kolinski, *Computational studies of protein folding.* Computing in Science & Engineering, 2001. **3**(3): p. 40-50.
3. 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.
4. Zhang, Z., *An Overview of Protein Structure Prediction: From Homology to Ab Initio.* 2002.
5. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-230.
6. Zhang, H., *Protein Tertiary Structures: Prediction from Amino Acid Sequences*, in *Encyclopedia of Life Sciences.* 2002.
7. Chan, H.S. and K.A. Dill, *The protein folding problem.* Physics Today, 1993: p. 24-32.
8. 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.
9. Eyrich, V.A., et al., *Protein tertiary structure prediction using a branch and bound algorithm.* 1999. p. 41-57.
10. Zhou, Y. and R. Abagyan, *Efficient Stochastic Global Optimization for Protein Structure Prediction*, in *Rigidity Theory and Applications.* 2002. p. 345-356.
11. 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.
12. Merkle, D. and M. Middendorf, *Swarm Intelligence*, in *Search Methodologies Introductory Tutorials in Optimization and Decision Support Techniques*, E.K. Burke and G. Kendall, Editors. 2005, Springer. p. 620.
13. Bonabeau, E., M. Dorigo, and G. Theraulaz, *Swarm intelligence: from natural to artificial systems.* 1999: Oxford University Press, Inc.
14. 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.
15. Abbass, H.A., *A Single Queen Single Worker Honey Bees Approach to 3-SAT*, in *The Genetic and Evolutionary Computation Conference.* 2001: San Francisco, USA.
16. Abbass, H.A., *A monogamous MBO approach to satisfiability*, in *International Conference on Computational Intelligence for Modeling, Control and Automation, CIMCA'2001.* 2001: Las Vegas,NV, USA.
17. Abbass, H.A., *An agent based approach to 3-SAT using marriage in honey-bees optimization.* International Journal of Knowledge-Based Intelligent Engineering Systems (KES), 2002. **6**(2): p. 1-8.
18. Abbass, H.A. and J. Teo, *A True Annealing Approach to the Marriage in Honey-Bees Optimization Algorithm.* International Journal of Computational Intelligence and Applications, 2003. **3**(2): p. 199 - 211.
19. Koudil, M., et al., *Using artificial bees to solve partitioning and scheduling problems in codesign.* Applied Mathematics and Computation, 2007. **186**(2): p. 1710-1722.
20. Fathian, M. and B. Amiri, *A honeybee-mating approach for cluster analysis.* The International Journal of Advanced Manufacturing Technology, 2007.
21. Fathian, M., B. Amiri, and A. Maroosi, *Application of honey-bee mating optimization algorithm on clustering.* Applied Mathematics and Computation, 2007. **190**(2): p. 1502-1513.
22. 2007 [cited; Available from: http://members.aol.com/queenb95/genetics.html#anchor173808.
23. Dietz, A. *Bee Genetics and Breeding.* in *Evolution.* 1986: Academic Press Inc.
24. Page, R.E., R.B. Kimsey, and H.H. Laidlaw, *Migration and dispersal of spermatozoa in spermathecae of queen honeybees (Apis mellifera L.).* Cellular and Molecular Life Sciences (CMLS), 1984. **40**(2): p. 182-184.
25. Benatchba, K., L. Admane, and M. Koudil, *Using Bees to Solve a Data-Mining Problem Expressed as a Max-Sat One*, in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach.* 2005. p. 212-220.
26. Eisenmenger, F., et al., *[SMMP] A modern package for simulation of proteins.* Computer Physics Communications, 2001. **138**: p. 192-212.
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. 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.