International Journal of Computational Intelligence and Informatics, Vol. 4: No. 2, July - September 2014

An Experimental Analysis of Evolutionary and Swarm Intelligence Algorithms for 3D HP Structure Prediction

V Veeralakshmi
Department of Computer Science
Bharathiar University
India Coimbatore-046

D Ramyachitra
Department of Computer Science
Bharathiar University
India Coimbatore-046

***Abstract*- Predicting the structure of protein has been the focus of the scientific research, but it has challenging in bioinformatics due to the computational complexity. The protein structure is determined by the experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. These methods cannot always be applied. So, the computational methods are frequently used to predict the structure with lowest free energy conformations. The lowest free energy is calculated based on the hydro polar and hydrophilic interactions. Many of the computational algorithms are used to solve the protein structure problem. In this comparative study the evolutionary algorithm, Genetic Algorithm and swarm intelligence algorithms Ant colony optimization (ACO) and artificial bee colony (ABC) algorithms are used and comparison is based on its energy value. The lowest energy value can easily predict the well known structures.**

Keywords- Protein Structure Prediction, HP Model, Evolutionary Algorithm and Swarm Intelligence Algorithms.

I. INTRODUCTION

In the molecular biology the DNA and the protein sequence have many secrets. The protein structure prediction problem (PSP) is that of computationally predicting the three dimensional structure of protein from the sequence of amino acids. The total interaction energy amongst the amino acids in the sequence is minimized [1]. The structural properties of the protein structure are used to find out the medicine and drug development for the cancer disease and many other diseases. The computational approaches to protein structure prediction are very attractive.

The protein folding is used to fold the protein sequences from native to non native ones. The protein folding problem describes the amino acid sequence of a protein dictates its structure which determines its mechanism of action. The difficulty in solving protein structure prediction problems stems from two major sources: (1) finding good measures for the quality of candidate structures and (2) given such measures, determining optimal or close-to-optimal structures for a given amino-acid sequence [2]. There is a large number of existing search algorithms that attempt to solve the PSP problem by exploring feasible lattice-based structures called conformations.

The optimal conformation in the HP model is the one that has the maximum number of H–H contacts which gives the lowest energy value. The assigned energy value is -1 [3]. The total free energy conformation is based on the HP model, becomes the sum of the energy released by all pairs of non-consecutive hydrophobic amino acid.

In this paper, an analysis is made to find out the minimum energy value by using the evolutionary and swarm intelligence algorithms. This paper uses the protein dataset for comparison of those algorithms. The remaining section of this paper is organized as follows. Section 2 describes the literature review, Section 3 describes the methodology and Section 4 describes our experimental result. And finally Section 5 gives the Conclusion and Future work.

II. LITERATURE REVIEW

Camelia Chira et al., proposed to address the hydrophobic - polar model of the protein folding problem based on hill-climbing genetic operators. The crossover and mutation are applied using a steepest-ascent hill-climbing approach [4]. The evolutionary algorithm with hill-climbing operators is successfully applied to the protein structure prediction problem for a set of difficult bi dimensional instances from lattice models.

Xiaolong Zhang et al., investigates the genetic tabu search algorithm to develop an efficient optimization algorithm. The crossover and mutation operators can improve the local search capability and variable population size strategy can maintain the diversity of the population, and the ranking selection strategy [5].

Thang N. Bui et al., proposed an efficient genetic algorithm for the protein folding problem used the HP model in the two-dimensional square lattice. A special feature of this algorithm is its usage of secondary structures that the algorithm evolves, as building blocks for the conformation. The algorithm performs very well against existing evolutionary algorithms and Monte Carlo algorithms [6].

Stefka Fidanova and Ivan Lirkov develop an ant algorithm for 3D HP protein folding problem. The components of an algorithm contribute to its performance and the performance is affected by the heuristic function and selectivity of pheromone updating. The aim is to achieve more realistic folding [7].

Alena Shmygelska et al., investigate a new algorithm, dubbed ACO-HPPFP-3, and are based on very simple structure components. The run-time required by ACO-HPPFP-3 for finding best known energy conformations scales worse with sequence length than PERM in 3D [8].

C. Vargas et al., proposed a parallel artificial bee colony algorithm approaches for protein structure prediction using 3dhp-sc model [9]. Two parallel approaches for the ABC are: master-slave and hybrid-hierarchical relations. The parallel models achieve good level of efficiency, and the hybrid hierarchical approach improved the quality of solutions.

Karaboga et al., presented the Artificial Bee Colony (ABC) algorithm for constrained optimization problems. The performances of the Artificial Bee Colony (ABC) algorithm is used for solving constrained optimization problems and produce the best results [10].

III. METHODOLOGY

A. Swarm Intelligence Algorithms

a*) Ant Colony Optimization (ACO) Algorithm*

Ant Colony Optimization (ACO) is a population-based stochastic search method for solving a wide range of combinatorial optimization problems .Ant colony optimization (ACO) algorithms were introduced by Marco Dorigo (Dorigo et al., 1996) [11]. An ACO algorithm is a simple iterative first improvement procedure that is based on this long-range mutation [12].

ACO algorithm is inspired by social behavior of ant colonies. ACO is an iterative construction search method in which a population of simple agents ('ants') repeatedly constructs candidate solutions to a given problem.

The protein sequences are initialized as an input for an ant colony algorithm to find out the best solution. In the beginning the whole sequences are given and ants construct the feasible solution. The pheromone are updated from the starting node, the next node is picked based on the transition rule. At each iteration ants produce the set of possible moves. The pheromone level updates have two stages.

- Local Update
- Global Update

In the Local update the pheromone information can be changed dynamically. After all ants are completed their tours the global update is performed and this will reduce the other ants to have the same solutions.

(1) Begin
(2) Initialize
(3) While stopping criterion not satisfied do
(4) Position each ant in a starting node
(5) Repeat
(6) For each ant do
(7) Choose next node by applying the state transition rule
(8) Apply local pheromone update
(9) End for
(10) Until every ant has built a solution
(11) Update best solution
(12) Apply global pheromone update
(13) End While
(14) End

Figure 1. Pseudo code for Ant Colony Optimization (ACO)

ACO is used to solve as the search for a minimum cost path in a graph, and to use artificial ants to search for good paths. Fig. 1 describes the pseudo code for ant colony optimization algorithm. An ACO algorithm iteratively undergoes three phases:

- Construction phase
- Local search phase
- Pheromone update phase

During the construction phase, the amino acid from left end sequence adds the next amino acid based on the pheromone level. In each ant path the intensity is updated. The starting position of the sequence is selected randomly and the end of the each iteration, the pheromone values is updated to construct the solution. The best solution is identified after all iterations are completed and the best energy path is updated by the global pheromone updating rule.

B. Artificial Bee Colony (ABC) Algorithm

Artificial bee colony (ABC) algorithm is one of the most recently introduced swarm-based algorithms. In ABC, the position of a food source represents a possible solution to the problem and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution [13]. The main algorithm of ABC is relatively simple and its implementation is, therefore, straightforward for solving optimization problems [14] and ABC has been found to be very effective and to produce very good results at a low computational cost. Fig. 2. describes the pseudo code for artificial bee colony algorithm.

(1) Begin
(2) Init Population ()
(3) While remain iterations do
(4) Select sites for the local search
(5) Recruit bees for the selected sites and to evaluate fitness
(6) Select the bee with the best fitness
(7) Assign the remaining bees to looking for randomly
(8) Evaluate the fitness of remaining bees
(9) Update Optimum ()
(10) End While
(11) Return Best Solution
(12) End

Figure 2. Pseudo code for Artificial Bee Colony (ABC)

In the ABC algorithm, a colony of bees is divided (C´esar Manuel Vargas Ben´ıtez and Heitor Silv´erio Lopes 2010) [15] into three groups:

- Employed bees
- Onlookers and
- Scout

In ABC, the protein sequences are used to produce the better optimum result. In the initial population the total sequences are given, the employed bees are associated with extracting food source, which is currently amino acid. It carried out the information from one amino acid to another amino acid based on the nectar amount. It engages the selected amino acid and evolves the fitness. To select the bee, based on the best fitness in order to evaluate the fitness of the remaining bees and the best solution is optimized.

Onlooker bees choose their next amino acid depending on the information given by the employed bees. The neighborhood amino acid is determined and its fitness value is computed. The scout bees are used to search the new solutions, randomly. The performance of the ABC depends on the max cycle number and the process is terminated to find the best solution.

C. Evolutionary Algorithm

a) *Genetic Algorithm*

GAs belongs to a specific class of evolutionary algorithms. Unger and Moult first applied GA to PSP. GA is to maintain a population of solutions. The size of the population is being maintained by replacing chromosomes in the population with solutions created by the genetic operators, based on the operations. Michalewicz (1996) [16] point toward that genetic algorithm is composed of five basic procedures:

1. Demonstrating problem characteristics or solutions in genetic forms.

2. Creating a number of initial solutions.

3. Establishment the design of fitness function.

4. Utilizing genetic operator to produce offspring, and the most commonly seen genetic operators are the following three: Selection, Crossover and Mutation. Fig. 3 describes the pseudo code for genetic algorithm. GA produce the better results than monto carlo algorithm.

(i) Choose the initial population of individuals

(ii) Evaluate the fitness of each individual in population

(iii) Repeat until termination condition satisfied:

Selection:

Select the individuals with greater fitness for reproduction

Crossover:

Breed new individuals through crossover

Mutation:

Apply probabilistic mutation on new individuals

Form a new population with these off springs.

(iv) Terminate

Figure 3. Pseudo code for Genetic Algorithm (GA)

In GA, the amino acids are used to find the minimum energy value. Choose the initialize population of the individual amino acid and the fitness is evaluated. In the selection process the greater fitness for production is selected. After that the cross over or the mutation is used to produce the new off springs with the minimum free energy value. To terminate the conditions until the new best optimal results are found.

IV. EXPERIMENTAL RESULTS

The protein instances used in our experiments are taken from the literature (as shown in Table I). The first six proteins 4RXN, 1ENH, 4PTI, 2IGD, 1YPA, and 1CTF are taken from [17] and the next three proteins 3MX7, 3NBM, and 3MQO from [18]. For the following experiment results, All experiments were performed on PCs with 2 GHz Intel(R) CPU 2020M and 2 GB RAM, running windows 7 (our reference machine). The java programming language is used for analyzing the protein sequences and find out the energy values.

In this paper we calculate the experimental measures by using the performance factors such as the classification accuracy and execution time. And also we find out the accuracy measure and error rate to determine the best algorithm for the ecoli protein dataset. The performance factors for these classification algorithms are listed in Table I and the accuracy measure by class for the classifier algorithms is depicted in Table II.

TABLE 1. THE PROTEINS USED IN OUR EXPERIMENTS

ID	LENGTH	SEQUENCES
4RXN	54	MKKYTCTVCGYIYNPEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDQFEEVE
1ENH	54	RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI
4PTI	58	RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
2IGD	61	MTPAVTTYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATKTFTVTE
1YPA	64	MKTEWPELVGKAVAAAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAQVPRVG
1CTF	74	AAEEKTEFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKEGVSKDDAEALKKALEEA GAEVEVK
3MX7	90	MTDLVAVWDVALSDGVHKIEFEHGTTSGKRVVYVDGKEEIRKEWMFKLVGKETFYVGAAKTKATIN IDAISGFAYEYTLEINGKSLKKM
3NBM	108	SNASKELKVLVLCAGSGTSAQLANAINEGANLTEVRVIANSGAYGAHYDIMGVYDLIILAPQVRSYYR EMKVDAERLGIQIVATRGMEYIHLTKSPSKALQFVLEHYQ
3MQO	120	PAIDYKTAFHLAPIGLVLSRDRVIEDCNDELAAIFRCARADLIGRSFEVLYPSSDEFERIGERISPVMIAH GSYADDRIMKRAGGELFWCHVTGRALDRTAPLAAGVWTFEDLSATRRVA

In the HP model [15], when two non- consecutive hydrophobic amino acids become topologically neighbors, they release a certain amount of energy, which for simplicity is shown as −1. The total free energy (E) of a conformation based on the HP model becomes the sum of the contributions of all pairs of non-consecutive hydrophobic amino acids as shown in Eq. 1.

$$E = \sum_{i<j-1} c_{ij} . e_{ij} \quad (1)$$

Here, cij = 1 if amino acids i and j are non-consecutive neighbours on the lattice, otherwise 0; and eij = −1 if ith and jth amino acids are hydrophobic, otherwise 0.

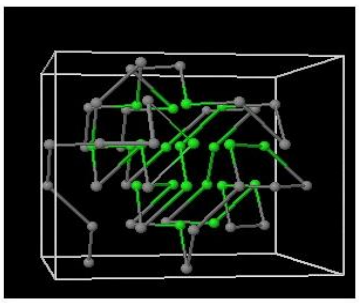
From the experimental results, the Table II shows the energy value is achieved by different algorithms for protein structure prediction in 3D HP model. The bold faces indicated the lowest energy value is obtained by the algorithms. The structure of the 9 protein sequences can be clearly seen in Fig. 4.

In Table II, the energy values are produced and it's shown that GA produces the minimum energy value with the minimum time complexity. In GA, the sequence is searched from the amino acid and when the threshold values are increased, the population generation is decreased and also the execution time is reduced.

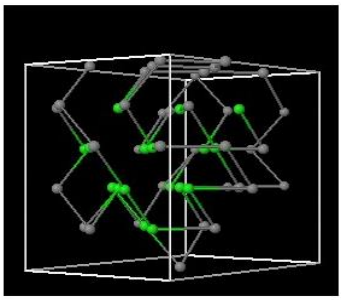
The energy value minimized is used to stable the structure of the protein from the amino acid. The energy value is given in the negative sign. In protein structures consists of the H core that hides the hydrophobic from the water surface. H-core formation is the main objective of HP based PSP. Energy minimization will bring the conformation to the nearest local minimum. In the minimum energy value, the energy is minimized then only the 3D HP protein structure is most stable because in the energy minimization the H-core becomes more strong and each amino acid is bonded tightly.

TABLE 2. THE ENERGY VALUES IS ACHIEVED BY GENETIC ALGORITHM, ANT COLONY OPTIMIZATION, AND ARTIFICIAL BEE COLONY ALGORITHM

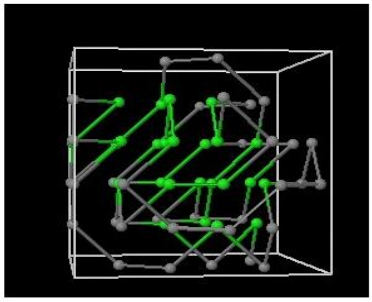
ID	LENGTH	ENERGY(-ve)		
		ACO	*ABC*	*GA*
4RXN	54	8	9	9
1ENH	54	19	23	18
4PTI	58	20	27	**29**
2IGD	61	29	33	**36**
1YPA	64	13	39	39
1CTF	74	31	33	**35**
3MX7	90	40	42	40
3NBM	108	43	49	**52**
3MQO	120	42	53	**58**



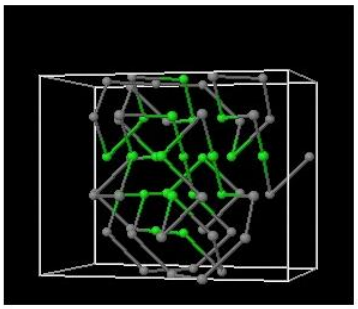
4RHN, LB-FE =-62



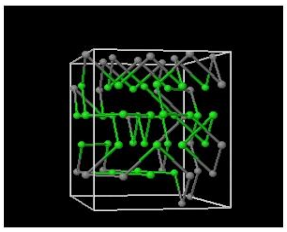
1EHN, LB-FE=-48



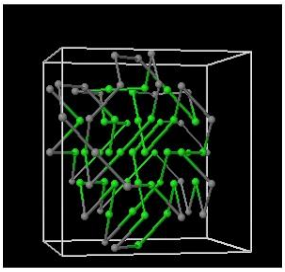
4PTI, LB-FE= -76



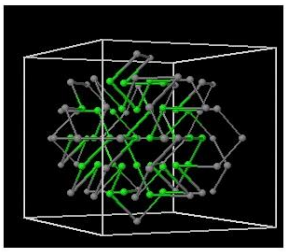
2IGD, LB-FE= -64



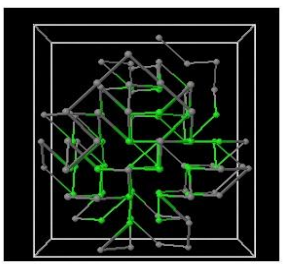
1YPA, LB-FE= -110



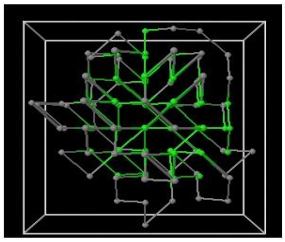
1CTF, LB-FE= -127



3MX7, LB-FE=-124



3NBM, LB-FE= -183



3MQO, LB-FE= -220

Figure 4. Results of the 3D HP structures of the each protein. (a) The green color denotes the hydrophobic (H) and the grey color denotes hydrophilic (P) amino acids. (b) The lower bound energy is given by the each protein.

V. CONCLUSION AND FUTURE WORK

In this paper we analyzed the performance of 3 algorithms namely Ant Colony Optimization, Artificial Bee Colony Algorithm and Genetic Algorithm. We used the real protein datasets for calculating the performance by using HP energy model for predicts the protein 3D structure. From the results, it is observed that the Genetic algorithm performs better than other algorithms. In Future the Evolutionary Algorithm can be experimented on other real protein datasets and HP instances also. And in future we can modify the Evolutionary and Swarm Intelligence Algorithms to obtain more effective results.

REFERENCES

[1] Mahmood A Rashid, Md Tamjidul Hoque, Hakim Newton M.A., Duc Nghia Pham, Abdul Sattar, "A New Genetic Algorithm for Simplified Protein Structure Prediction", Springer Berlin/Heidelberg, vol. 7691, pp. 107-119, 2012.

[2] Krasnogor N, Hart WE, Smith J, Pelta DA, "Protein structure prediction with evolutionary algorithms", Proceeding of the Genetic and Evolutionary Computation conference, pp. 1596-1601, 2009.

[3] Cheng-Jian Lin and Shih-Chieh Su, "Using An Efficient Artificial Bee Colony Algorithm For Protein Structure Prediction On Lattice Models", International Journal Of Innovative Computing, Information And Control , vol. 8, pp. 2049-2064, 2012.

[4] Camelia Chira, Dragos Horvath, "Dumitru DumitrescuEvolutionary Computation, Machine Learning and Data Mining in Bioinformatics", Lecture Notes in Computer Science, vol. 6023, pp. 38-49, 2010.

[5] Xiaolong Zhang, Ting Wang, Huiping Luo, Jack Y Yang, Youping Deng, Jinshan Tang, Mary Qu Yang, "3D Protein structure prediction with genetic tabu search algorithm", BMC Systems Biology, vol. 4, pp. 1-9, 2010.

[6] Thang N. Bui and Gnanasekaran Sundarraj, "An Efficient Genetic Algorithm for Predicting Protein Tertiary Structures in the 2D HP Model", GECCO '05 Proceedings of the 7th annual conference on Genetic and Evolutionary computation, pp. 385-392, 2005.

[7] Stefka Fidanova, Ivan Lirkov, "Ant Colony System Approach for Protein Folding", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 887–891, 2008.

[8] Alena Shmygelska, and Holger H Hoos, "An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem", BMC Bioinformatics, vol. 6, pp. 1-22, 2005.

[9] Vargas Benitez and Lopes, "Parallel artificial bee colony algorithm approaches for protein structure prediction using the 3dhp-sc model", Intelligent Distributed Computing, vol. 4, pp. 255-264, 2010.

[10] Karaboga N, Cetinkaya MB, "A novel and efficient algorithm for adaptive filtering: Artificial bee colony algorithm". Turk Journal of Electronic Engineering Computer Science, vol. 19, pp. 175–190, 2012.

[11] Dorigo, Maniezzo and Colorni, "Ant system: Optimization by a colony of cooperating agents", IEEE Transaction on Systems, Man, and Cybernetics – Part B, vol. 26, pp.29–41, 1996.

[12] Bai Li, YaLi , LigangGong ,"Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model", Engineering Applications of Artificial Intelligence, vol. 27, pp. 70–79, 2014.

[13] C´esar Manuel Vargas Ben´ıtez and Heitor Silv´erio Lopes, "Parallel Artificial Bee Colony Algorithm Approaches for Protein Structure Prediction Using the 3DHP-SC Model", Springer-Verlag Berlin Heidelberg 2010.

[14] Karaboga N, Cetinkaya, "A novel and efficient algorithm for adaptive filtering: Artificial bee colony algorithm", Turk Journal of Electronic Engineering Computer Science, vol.19, pp.175–190, 2012.

[15] Bor-Wen Cheng, Chun-Lang Chang, "A study on flowshop scheduling problem combining Tayuchi experimental design and genetic algorithm", Expert System with Applications, vol. 32, pp. 415-421, 2007.

[16] D. Ullah and K. Steinh¨ofel, "A hybrid approach to protein folding problem integrating constraint programming with local search", BMC Bioinformatics, vol. 11, 2010.

[17] S. Shatabda, M. H.Newton, and A. Sattar, "Mixed heuristic local search for protein structure prediction," in AAAI Conference on Artificial Intelligence, pp. 876–882, 2013.

[18] Mahmood A Rashid, Md Tamjidul Hoque, Hakim Newton M.A., Duc Nghia Pham, Abdul Sattar, "A New Genetic Algorithm for Simplified Protein Structure Prediction", Springer Berlin Heidelberg, vol. 7, pp. 107-119, 2012.