VOL. 10, NO. 11, JUNE 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ACO-METAHEURISTIC FOR 3D-HP PROTEIN FOLDING OPTIMIZATION

N. Thilagavathi and T. Amudha
Department of Computer Applications, Bharathiar University, Coimbatore, India
E-Mail: thilagamca86@gmail.com

ABSTRACT

Protein Folding is a broad research field in computational Biology, Molecular Biology and Bioinformatics. Protein Folding Optimization is one of the NP-hard problems. Bio-inspired metaheuristics plays a major role in solving the protein folding optimization which can mimic the insect's problem solving abilities like foraging, nest building and mating. In this paper Ant Colony Optimization (ACO) - Metaheuristic was applied to solve 3D-HP protein folding optimization. The 3D structure of a protein is also called as final native structure, which is responsible for functioning of a particular protein. Misfolded or unfolded protein is responsible for several neurodegenerative diseases. The instances for 3D-HP protein folding were taken from the HP benchmarks. The energy minimization is the major objective function to obtain the best 3D structure of protein. Various energy functions are used in this work to obtain different energy values.

Keywords: protein folding optimization, ACO-Metaheuristics, 3D-HP protein folding, alternative energy functions.

1. INTRODUCTION

The Protein sequence is formed by a combination of 20 amino acids. The combinations of amino acids are linked to each other with poly peptide bond to form a primary structure of a protein. Proteins have four levels of structures similar to primary, secondary, tertiary and quaternary. The tertiary (3D) structure of protein is called the conformation structure and has the optimal free energy possible. To find the native conformation structure of protein is a global optimization process and also it takes lots of time to optimize the structure. Each protein has its own ability to fold into a globular state or structure automatically. Sequence of amino acids (Primary Structure) altered in to 2D and 3D structure is called as 'protein folding'. Most of the proteins fold up into unique 3-Dimensional structures in few milliseconds. Protein Folding is the general procedure in all living organic cells. Structure of protein is very important for permanence and standard functioning [12]. The most stable tertiary structure of a protein is biologically active only in its 3D structure. Protein misfolding or unfolding is a common incident in living cells. Because of protein misfolding or unfolding many diseases will occur like Alzheimer's disease, Huntington's disease, Parkinson's disease and the prion diseases [1].

Bio-inspired algorithms are mainly developed to optimize the combinatorial optimization problems which mimic the behaviors of social insects. Ant Colony Optimization (ACO) Metaheuristic is a population based search method for solving combinatorial optimization problems. ACO imitates the foraging behavior of real ants to solve optimization problems. The foraging behavior is the concept of indirect communication between the members of population through pheromones [13]. The pheromones have capability to evaporate; the solution is calculated depending upon the strength of pheromones. Ant's population is initialized as number of folds and each ant give unique fold to the given HP sequence. The best solution (fold) is selected among all feasible solution through local and global pheromone updations. After completion of every iteration the pheromone values are updated locally and global pheromone updation is the best solution among all local solutions.

2. 3D-HP PROTEINFOLDING

The computational techniques and strategies are currently used in the development of 'in silico' methods for the 3-D HP Protein Folding Problem. In HP Lattice model the 20 amino acids are converted into 2 types as H (Hydrophobic) and P (Hydrophilic) which depends upon the characteristics of amino acids. Amino acids have two types of characteristics, water repellent (Hydrophobic) and water attraction (Hydrophilic). 3D-HP model consists of 3D cubic lattice model. The 3D cubic lattice model was designed by a common back bone or side chain of a Hydrophobic or Hydrophilic. The 3D-HP model have 6 possible directions as L (Left), R (Right), U (Up), D (Down), F (Forward) and B (Backward) to move a residue in lattice. In lattice model the hydrophobic (H) interactions are driving force for protein folding. In lattice, each sequence pattern is a self-avoiding walk.A self-avoiding walk is a sequence of moves on a lattice or a grid that does not visit the same position more than once. Links between H-H monomers are constructive [14]. The native conformation of a HP sequence is defined as the set of conformations with the biggest possible number of H-H contacts. The energy values are calculated based on the number of H-H contacts. The energy value should be minimized to get the best 3D structure. Figure-1 shows the diagrammatic representation of 3D-HP lattice model [9, 11].

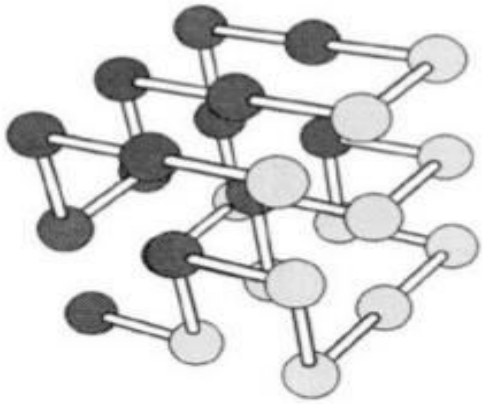

Figure-1. 3D-HP lattice model [16].

3. ACO-METAHEURISTIC FOR PROTEIN FOLDING OPTIMIZATION

Ant Colony Optimization (ACO) is one of thebio-inspired algorithm which is used for solving hard combinatorial optimization problems. ACO takes inspiration from the foraging behavior of real ant colony. In the real world, ants (initially) walk at random, and finding food return to their colony while laying down pheromone trails. ACO is based on the indirect communication of a colony of simple agents, called (artificial) ants, mediated bypheromone trails. The ACO algorithm, pheromone trails serve as distributed, numerical information which the ants use to probabilistically construct solutions to the problem being solved. The ants adapt during the algorithm's execution to reflect their search experience. At each step, ants compute a set of possible moves and select rest of the travel around [10]. The transition probability is based on the heuristic information and pheromone trail level of the move. In the starting stage, the pheromone level is set to a small positive constant value and then the ants update pheromone values after completing the construction phase. The pheromone updations have two stages, Local Update and Global Update.

a) Local update

The Pheromone values of the visited paths are updated locally, while ants construct their solutions by applying the local update rule. The main aim of the local update rule is to make better use of thepheromone in sequence by dynamically changing the desirability of edges.

$$\tau_{ij} \leftarrow 1 - \rho\tau_{ij} + \rho\tau_0 \quad (1)$$

Where $'\tau_{ij}'$ is an amount of the pheromone on the arc (i, j) of the 3D cubic lattice model, $'\rho'$ is the persistence of the trail and the term (1- ρ) can be interpreted as trail evaporation. $'\tau_0'$is the pheromone initialization value, set as a small positive constant value [10].

b) Global update

The Global updating is performed after all ants have completed their tours. The pheromone levels will be reduced and this will reduce the possibility, that the other ants will select the same solution and as a result the search will be more diversified [10].

$$\tau_{ij} \leftarrow 1 - \rho\tau_{ij} + \Delta\tau_{ij} \quad (2)$$

Where

$$\Delta\tau_{ij} = \begin{cases} -E_{gb} \text{ if}(i,j) \in \text{bestsolution} \\ 0 \qquad\qquad \text{Otherwise} \end{cases} \quad (3)$$

$'E_{gb}$ is the global energy of the best folding. The global update rule is used to provide a greater amount of pheromone on the paths of the best solution.

c) Probability to select next move

In 3D-HP Lattice model, there are six possible moves. During the construction phase, ants fold a protein from the left end of the sequence adding one amino acid at a time based on the two sources of information; pheromone matrix value and heuristic information. The heuristic values are calculated for each residue before it is placed on the lattice. The Maximum heuristic value for each position is 6. In 3D-HP model the positions are represented as L (Left), R (Right), U (Up), D (Down), F (Forward) and B (Backward) [2].

$$P_{ij} \leftarrow \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{k\in Unused}\tau_{ik}^{\alpha}\eta_{ik}^{\beta}} \quad (4)$$

The $'\tau_{ij}'$ is the intensity of the pheromone deposited by each ant on the path (i, j). The 'i' is the start position and 'j' is the next movable position, $'\alpha'$ is the intensity control parameter, $'\eta_{ij}'$ is the heuristic information equal to the number of new H-H contacts if the position 'j' is chosen, $'\beta'$ is the heuristic parameter. When the next amino acid is polar then the probability should be$P_{ij} = 0$.

The Ant Colony Optimization Pseudo code starts with initialization. The initialization has size of population, pheromone value,$'\alpha'$,$'\beta'$ and $'\rho'$ values. The $'\alpha'$ is intensity control parameter, $'\beta'$ is the heuristic parameter and $'\rho'$ is the persistence of the trail. The stopping criterion for this folding process is to occupy the unique lattice point by each residue in the given sequence. The starting point in lattice for folding is selected randomly. Each ant gives unique fold with different energy values. At the end of each iteration, the pheromone values are updated locally. The number of iterations will vary based upon the best possible solution. The best optimal solution is identified after completion of all iterations. The best fold path is updated by the global pheromone updating rule. The energy value for best folding is calculated, after all iterations are completed [9]. The native structure have finalized, which structure has the minimal energy value.

Pseudocode of ACO-Metaheuristic

```
[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 position 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
```

4. ALTERNATIVE ENERGY FUNCTIONS

The alternative energy functions are used to calculate the energy values in different method. Three letter acronyms have been assigned to evaluate the energy functions are D85, K99 and I09. The acronyms describe the first letter of author and followed by the published year. The main aim of these alternative formulations of the energy function is to provide a further fine-grained discrimination, as a means of guiding metaheuristics in a more efficient way in the process of finding possible solutions to the new problem [7].

a) Energy function D85

The alternative energy function D85 is calculated from the native structure of a protein also called as 'Free Energy (FE)' function. The energy value is taken as -1 when both S_i and S_j are H in non-consecutive order and also form a topological contact, otherwise it taken as 0[7].

$$E_{D85}(c) = \sum_{s_i,s_j} e(s_i,s_j) \quad (5)$$

In equation 5, 'c' is conformation state of a protein sequence and $e(s_i,s_j)$ is expressed as

$$e(s_i,s_j) = \begin{cases} -1, & \text{if } s_i \text{ and } s_j \text{ are both H and} \\ & \text{form a topoogical contact} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

b) Energy function K99

The alternative energy function K99 is also called as 'distance - dependent' energy function [6], as given in equation (7). The energy values are calculated depends upon the distance between the two hydrophobic (H) amino acids.

All nonconsecutive topological contacts are considered for this energy calculation.

$$E_{K99}(c) = \sum_{s_i,s_j} e(s_i,s_j) \quad (7)$$

Where

$$e(si, sj) = \begin{cases} -1 \text{ if } s_i \text{ and } s_j \text{ are both H and they form a} \\ \text{topological contact,} \\ -\frac{1}{d(s_i,s_j)^k L_H} \text{ if } s_i \text{ and } s_j \text{ are both H but the lattice} \\ \text{distance between them is } d(s_i,s_j) > 1 \\ 0, \text{ otherwise.} \end{cases} \quad (8)$$

Krasnogor *et al.* [6] suggested to use the values $k = 4$ for the square lattice and $k = 5$ for the cubic and triangular lattices, respectively.

c) Energy function I09

1) H-Compliance

H-compliance (H_C) measures the closeness of H amino acids to the center of a hypothetical rectangle (or cuboid in 3D space) enclosing all H amino acids, which is denoted by the reference point (x_r,y_r,z_r). Where x_s, y_s and z_s denote the lattice coordinates of the 's' amino acid [3,4,5]. **'L_H'** is the number of hydrophobic amino acids in a given sequence.

$$H_c = \frac{\sum_{s|s=H}(x_r-x_s)^2+(y_r-y_s)^2+(z_r-z_s)^2}{L_H} \quad (9)$$

The coordinates (x_r, y_r, z_r) of the "center"are obtained as:

$$x_r = (xhmax - xhmin)/2 \quad (10)$$

$$y_r = (yhmax - yhmin)/2 \quad (11)$$

$$z_r = (zhmax - zhmin)/2 \quad (12)$$

2) P-Compliance

P-compliance (P_C) computes how close P amino acids are to the boundaries of a hypothetical rectangle enclosing all P amino acids. Here **'L_P'** is number of polar amino acids in a given sequence. Such a cuboid is defined by x_{min}, x_{max}, y_{min},y_{max} ,z_{min} and z_{max}[3,4,5].

$$p_c = \frac{\sum_{s|s=p} \min\{|x_{min}-x_s|,|x_{max}-x_s|,|y_{min}-y_s|,|y_{max}-y_s|,|z_{min}-z_s|,|z_{max}-z_s|\}}{L_p} \quad (13)$$

The modified fitness function (E_{I09}) is defined as

$$E_{I09}(c) = \alpha E_{D85} + H_C + P_C \quad (14)$$

E_{D85} is the conventional energy function of the HP model and **'α'**is large enough to ensure this will be the dominant term with high integer constant value. H-compliance and P-compliance values are add with conventional energy function to find out the I09 energy value.

5. IMPLEMENTATION RESULTS AND DISCUSSIONS

Protein Folding Problem was solved by using ACO-Metaheuristic and implementation using NET Framework 3.5 with C# Language. In this research, the

Ant Colony was initialized with five ants, and each ant performed folding process and gave different folding structures. Further, the number of iterations was increased to get better folding structure. The solutions (folding) given by all the five ants were compared and the optimal folding was chosen with the help of minimum energy obtained by the ants.

In 3D-HP protein folding problem, certain constraints to be followed are the folding sequence chain should not break; the folding process should be within the cubic lattice boundary and follow self-avoiding walk. In cubic lattice model the starting position of folding process was selected randomly. Each residue in sequence was added to the folding process. The directions for cubic lattice model should be L (left), R (Right), U (Up), D (Down), F (Forward) and B (Backward). Each residue hassix possible movements; among six solutions one solution was chosen which is optimal. After generating each solution, it was checked to identify whether the movement could lead towards the right direction.

The movements are encoded in several formats, like symbols, numbers or alphabets. Sequences have a number of possibility structures,

Possibility structure = n^6 (15)

n = Sequence length

6 = Possible movements in cubic lattice

Table-1. Parameter settings of ACO.

Parameter	Value	Description
$\propto$	7 - 9	Intensity control parameter
β	0.4 - 0.5	Heuristic parameter
ρ	0.3 - 0.5	Persistence of the trail
m	500 - 1000	Number of ants in ACO
τ_0	0.5	Pheromone initialization

Table-2. HP benchmark sequences for the 3D cubic lattice.

Seq. Id	Sequence	Length	Energy	ACO under FE		
				D85	K99	I09
3d1	HPHP2H2PHP2HPH2P2HPH	20	-11	-10	-9.04	-7.5
3d2	H2P2HP2HP2HP2HP2HP2HP2H2	24	-13	-8	-8.03	-5.64
3d3	P2HP2H2P4H2P4H2P4H2	25	-9	-6	-6.02	-3.69
3d4	P3H2P2H2P5H7P2H2P4H2P2HP2	36	-18	-10	-10.04	-5.92
3d5	P2H3PH3P3HPH2PH2P2HPH4PHP2H5PHPH2P2H2P	46	-32	-21	-21.06	-16.41

In this paper 5 3D-HP benchmark sequences [8, 15] were tested with ACO algorithm. Free Energy (FE) results of 5 different 3D sequences with various energy functions are listed in Table-2. The resultant optimal protein structure obtained by ACO algorithm for the 5 3D-HP protein sequences are given in the below figures. The energy values are calculated with 3 different energy functions, D85, K99 and I09. The energy calculation method used in each function is different from each other. Figure-2, Figure-3, Figure-4, Figure-5 and Figure-6 shows the optimal conformation structure of sequence length 20, 24, 25, 36 and 46 by ACO algorithm, respectively.

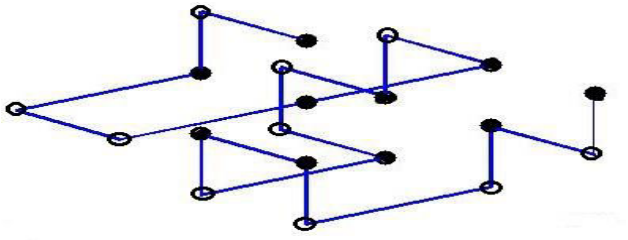

Figure-2. Optimal structure obtained by ACO for 3d1.

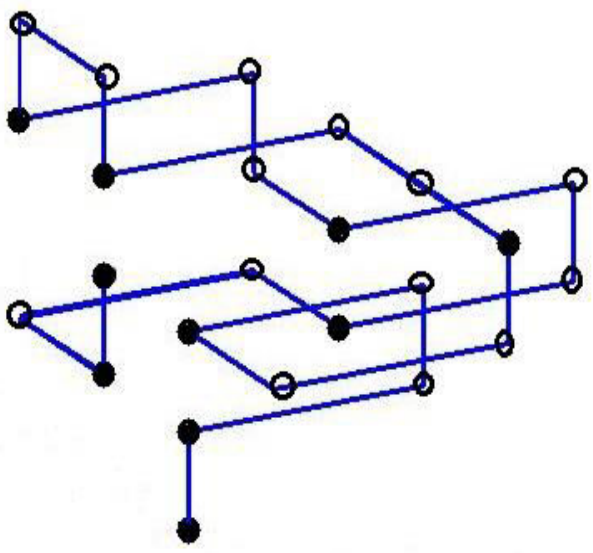

Figure-3. Optimal structure obtained by ACO for 3d2.

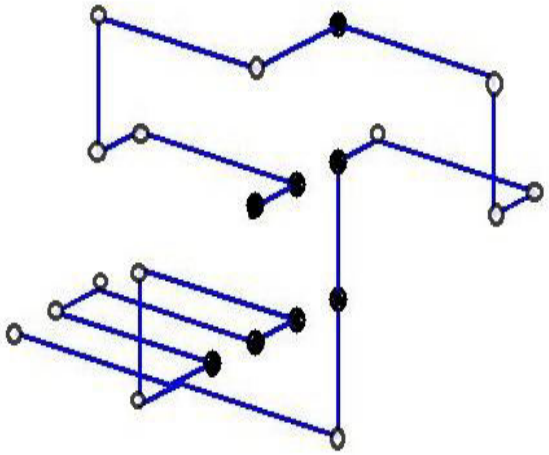

Figure-4. Optimal structure obtained by ACO for 3d3.

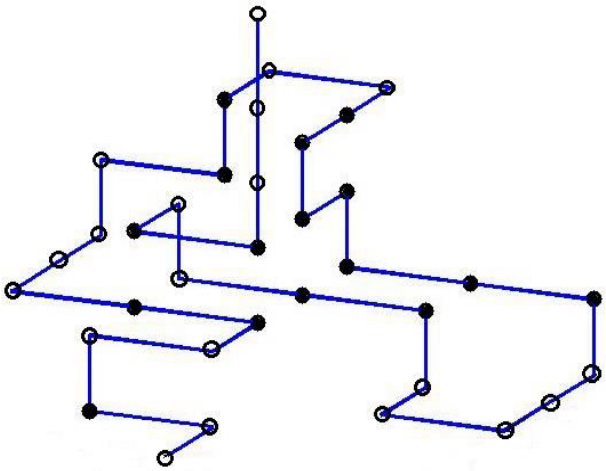

Figure-5. Optimal structure obtained by ACO for 3d4.

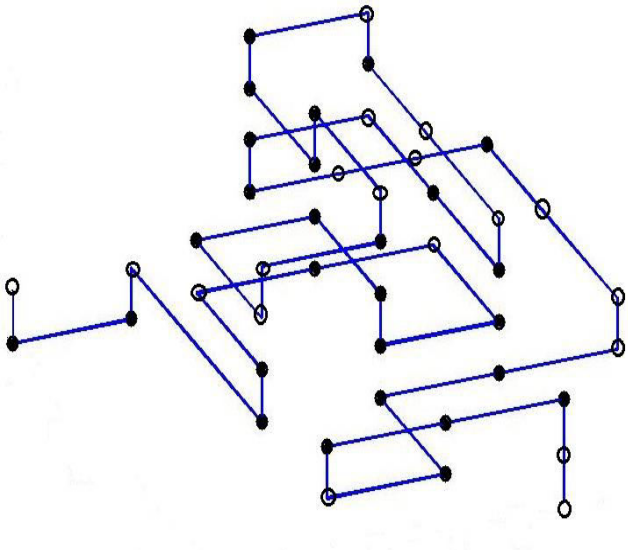

Figure-6. Optimal structure obtained by ACO for 3d5.

6. CONCLUSION AND FUTURE WORK

In this paper, the Ant Colony Optimization (ACO) algorithm was applied to 3D-HP protein folding to solve the problem in an efficient way. The problem instances for testing the algorithm were taken from the standard HP Benchmarks in the literature [8,15]. It was proven that the ACO algorithm could obtain the best-so-far optimal solutions (minimal energy) for most of the problem cases. In future work, Bio-inspired algorithms can be applied to optimally fold the real time proteins. In this research work the ACO was applied for 3D-HP benchmarks; in future, other modified variants can be applied for real proteins to predict the structure and to analyze the protein misfolding diseases.

REFERENCES

[1] Claudio Soto. 2001. "Protein misfolding and disease; protein refolding and therapy", Elsevier Science, 498, 204-207.

[2] D. Chu, M. Till and A. Zomaya. 2005. "Parallel Ant Colony Optimization for 3D Protein Structure Prediction using the HP Lattice Model", Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) 1530-2075/05$20.00 IEEE.

[3] Islam M., Chetty M. and Murshed M. 2011. "Novel local improvement techniques in clustered memetic algorithm for protein structure prediction", In Proc. IEEE Congress on Evolutionary Computation, pp.1003-1011, June.

[4] Islam M. and Chetty M. 2010. "Clustered memetic algorithm for protein structure prediction". In: Proc. IEEE Congress on Evolutionary Computation, pp. 1-8, July.

[5] Islam M. and Chetty M. 2009. "Novel memetic algorithm for protein structure prediction". In Lecture Notes in Computer Science 5866, Nicholson A, Li X (eds.), Springer Berlin/Heidelberg, pp.412-421.

[6] Krasnogor N., Hart W. and Smith J. *et al.* 1999. "Protein Structure prediction with evolutionary algorithms", Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1569-1601, July.

[7] Mario Garza-Fabre, Eduardo Rodriguez-Tello and Gregorio Toscano-Pulido. 2013. "Comparative Analysis of Different Evaluation Functions for Protein Structure Prediction under the HP Model", Journal Of Computer Science and Technology Vol. 28, No. 5, pp. 868–889, DOI 10.1007/s11390-013-1384-7, September.

[8] Mario Garza-Fabre, Gregorio Toscano-Pulido and Eduardo Rodriguez-Tello. 2003. "Benchmark Sequences for the HP Model of Protein Structure Prediction: 2D square and 3D cubic lattices".

[9] Shmygelska A. and Hoos H. 2005. "An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem". BMC Bioinformatics, 6: Article No.30.

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

[11] Stefka Fidanova. 2006. "3D HP Protein Folding Problem using Ant Algorithm", BioPS'06, 24-25, III.19-III.26, October.

[12] Thilagavathi N. and T. Amudha. 2014. "An Analytical Study of NP-Hard Protein Folding Problems", IEEE International Conference on Intelligent Computing Applications, 978-1-4799-3966-4 © 2014 IEEE, DOI 10.1109/ICICA.2014.47, March.

[13] Thilagavathi N. and T. Amudha. 2015. "Rank Based Ant Algorithm for 2D-HP Protein Folding". Computational Intelligence in Data Mining - Volume 3, Smart Innovation, Systems and Technologies 33, DOI 10.1007/978-81-322-2202-6_40, © Springer India.

[14] Zhang Y. and Skolnick J. 2004. "Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins", Biophysical Journal, vol. 87, pp. 2647–2655.

[15] http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html www.slideshare.net,"The mechanism of protein folding".