# 3D HP Protein Folding Problem using Ant Algorithm

**Fidanova S.**

*Institute of Parallel Processing – BAS*
*25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria*
*Phone: +359 2 979 66 42*
*E-mail: stefka@parallel.bas.bg*

**Summary:** Predicting the 3D form of protein from their linear sequence is one of the major challenges in modern biology. Even under simplified lattice models, the problem is NP-hard and the standard computational approaches are not powerful enough to search for the correct structure in the huge conformation space. Due to the complexity of the protein folding problem simplified models such as hydrophobic-polar (HP) model has become one of the major tools for studying protein structure. Various optimization methods have been applied on folding problem including Monte Carlo methods, Evolutionary algorithms, ant colony optimization algorithm. In this work we develop an ant algorithm for 3D HP protein folding problem. It is based on very simple design choices in particular with respect to the solution components reinforced in the pheromone matrix. The achieved results are compared favorably with specialized state-of the art methods for this problem. Our empirical results indicate that our rather simple ant algorithm outperforms the existing results for standard benchmark instances from the literature.

**Keywords:** Protein Folding, HP Model, Metaheuristics, Ant Algorithm, Pheromone Information, 3D Cubic Lattice.

## 1. INTRODUCTION

Determining the functionality of a protein molecule from amino acid sequence remains a central problem in computational biology, molecular biology, biochemistry and physics. Even the experimental determination of these conformations is often difficult and time consuming. It is common practice to use models that simplify the search space of possible conformation. These models try to generally reflect different global characteristics of protein structures. In the hydrophobic-polar (HP) model [3] the primary amino acid sequence of a protein (which can be represented as a string over twenty-letter alphabet) is abstracted to a sequence of hydrophobic (H) and polar (P) residues that is represented as a string over the letter H and P. In the model, the amino acid sequence is abstracted to a binary sequence of monomers that are either hydrophobic or polar. The structure is a chain whose monomers are on the vertices of a three dimensional cubic lattice. The free energy of a conformation is defined as the negative number of non-consecutive hydrophobic-hydrophobic contacts. A contact is defined as two nonconsecutive monomers in the chain

occupying adjacent sites in the lattice. In spite of its apparent simplicity, finding optimal structures of the HP model on a cubic lattice is NP-complete problem [1].

Ant Colony Optimization (ACO) is a population-based stochastic search method for solving a wide range of combinatorial optimization problems. ACO is based on the concept of indirect communication between members of a population through interaction with the environment. From the computational point of view, ACO is an iterative construction search method in which a population of simple agents (ants) repeatedly constructs candidate solutions to a given problem. This construction process is probabilistically guided by heuristic information on the given problem instances as well as by a shared memory containing experience gathered by the ants in previous iterations.

This work is an investigation of the HP model in a three dimensional cubic lattice using an ACO as a tool to find the optimal conformation for a given sequence. The achieved results are evaluated and compared with other heuristic methods using 10 sequences of 48 monomers from the literature.

The paper is organized as follows: the problem is described in Section 2. The ACO algorithm is in Section 3. The achieved results are discussed in Section 4. The paper ends with a summary of the conclusions.

## 2. THE PROTEIN FOLDING PROBLEM

The processes involving in folding of proteins are very complex and only partially understood, thus the simplified models like Dill's HP model have become one of the major tools for studying proteins [3]. The HP model is based on the observation that hydrophobic interconnection is the driving force for protein folding. The protein conformations of this sequence are restricted to self-avoiding paths on 3-dimensional sequence lattice. One of the most common approaches to protein structure prediction is based on the thermodynamic hypothesis which states that the native state of the protein is the one with lowest Gibbs free energy. In the HP model, the energy of a conformation is defined as a number of topological contacts between hydrophobic amino acid that are not neighbors in the given sequence. More specifically a conformation $c$ with exactly $n$ such H-H contacts has free energy $E(c) = n(-1)$. The 3D HP protein folding problem can be formally defined as follows. Given an amino acid sequence $s = s_1, s_2, \ldots, s_n$, find an energy minimizing conformation of $s$, i.e. find

$$c^s \in C(s)$$

such that

$$E^s = E(c^s) = \min\{E(c) \mid c \in C\},$$

where $C(s)$ is the set of all valid conformations for $s$.

A number of well-known heuristic optimization methods have been applied to the 3D protein folding problem including Evolutionary Algorithm (EA) [7], Monte Carlo (MC) algorithm [8] and Ant Colony Optimization (ACO) [9]. An early application of EA to protein structure prediction was presented by Unger and Moult [11]. Their EA incorporates characteristics of Monte Carlo methods. Currently among the best known algorithms for the HP protein folding problem is Pruned-Enriched Rosenblum Method (PERM) [6]. Among these methods are the Hydrophobic Zipper (HZ) method [4] and the Constraint-based Hydrophobic Core Construction Method (CHCCM) [12]. The Core-direct chain growth method (CG) [2] biases construction towards finding a good hydrophobic core by using a specifically designed heuristic function.

## 3. ACO ALGORITHM FOR PROTEIN FOLDING PROBLEM

Real ants foraging for food lay down quantities of pheromone (chemical cues) marking the path that they follow. An isolated ant moves essentially at random but an ant encountering a previously laid pheromone will detect it and decide to follow it with high probability and therefore reinforce it with a future quantity of pheromone.

The ACO algorithm uses a colony of artificial ants that behave as co-operative agents in a mathematical space where they are allowed to search and reinforce path ways (solutions) in order to find the optimal ones. The problem is represented by graph and the ants walk on the graph to construct solutions. After initialization of the pheromone trails, ants construct feasible solutions and the pheromone trails are updated. At each step ants compute a set of feasible moves and select the best one (according to some probabilistic rules) to carry out the rest of the tour. The transition probability is based on the heuristic information and pheromone trail level of the move. The higher the value of the pheromone and the heuristic information, the more profitable is to select this move and resume the

search. In the beginning, the initial pheromone level is set to a small positive constant value $\tau_0$ and then ants update this value after completing the construction stage. ACO algorithms adopt different criteria to update the pheromone level. In our implementation Ant Colony System (ACS) approach is used [5]. In ACS the pheromone updating consists of two stages: local update and global update. While ants build their solutions, at the same time they locally update the pheromone level of the visited paths by applying the local update rule as follows:

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

Where $\tau_{ij}$ is an amount of the pheromone on the arc $(i, j)$ of the 3D cube lattice, $\rho$ is a persistence of the trail and the term $(1 - \rho)$ can be interpreted as trail evaporation. Using this rule, ants will search in a wide neighborhood of the best previous solution. AS shown in the formula, the pheromone level on the paths is highly related to the value of evaporation parameter $\rho$. The pheromone level will be reduced and this will reduce the chance that the other ants will select the same solution and consequently the search will be more diversified. When all ants have completed their solutions, the pheromone level is updated by applying the global updating rule only on the paths that belong to the best solution since the beginning of the trials as follows:

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

Where $\Delta\tau_{ij} = \begin{cases} -E_{gb} & if\ (i,j) \in \text{best solution} \\ 0 & \text{otherwise} \end{cases}$

The $E_{gb}$ is the free energy of the best folding. This global updating rule is intended to provide a greater amount of pheromone on the paths of the best solution, thus intensify the search around this solution.

There are six possible positions on the 3D lattice for every amino acid. They are the neighbor positions of the precedence amino acid. Since conformations are rotationally invariant, the position of the first two amino acids can be fixed without loss of generality. 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, which represents previous search experience, and heuristic information. The

transition probability to select the position of the next amino acid is given as:

$$P_{ij} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{s \in \text{allowed}} \tau_{is}^{\alpha}\eta_{is}^{\beta}} & \text{if } j \in \text{allowed} \\ 0 & \text{otherwise} \end{cases}$$

Where $\tau_{ij}$ is the intensity measure of the pheromone deposited by each ant on the path (*i, j*), $\alpha$ 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 and the *allowed* is the set of free neighbor positions. Thus the higher the value of $\tau_{ij}$ and $\eta_{ij}$, the more profitable is to put the next amino acid on the position *j*. When the next amino acid is polar, the probability is $P_{ij} = 0$. In this case the position is chosen randomly between allowed positions. When the set of allowed positions is empty, the ant does some steps back and after that it continues construction of the solution.

## 4. EXPERIMENTAL RESULTS

Ten standard benchmark instances of length 48 for 3D HP protein folding shown in Table 1 have been widely used in the literature [2, 6, 8, 9, 10]. Experiments on these standard benchmark instances were conducted by performing a number of independent runs for each problem instance, 20 runs. The following parameter settings are used for all experiments: $\alpha = \beta = 1$, $\rho = 0.5$. Furthermore, all pheromone values were initialized to $\tau_{ij} = 0.5$ and a population of 5 ants were used. The algorithm was terminated after 200 iterations. All experiments were performed on IBM ThinkPad Centrino 1.8GHz CPU, 512 KB RAM running SuSe Linux.

In Table 2 the achieved results by various heuristics have compared. For every of the benchmark instances the best found result by various methods is reported.

Table 1. 3D standard benchmark instances

| Name | HP string |
|---|---|
| S2-1 | HPHHPPHHHHPHHHPPHHPPHPHHPHPHHPPHHPPPHPPPPPPPPHHP |
| S2-2 | HHHHPHHPHHHHHPPHPPHHPPHPPPPPPHPPHPPPHPPHPHHPPHHHPH |
| S2-3 | PHPHHPHHHHHHPPHPHPPHPHHPHPHPPPHPPHHPPHHPPHPHPPHP |
| S2-4 | PHPHHPPHPHHHPPHHPHHPPPHHHHHPPHPHHPHPHPPPPHPPHPHP |
| S2-5 | PPHPPPHPHHHHPPHHHHPHHPHHHPPHPHPHPPHPPPPPPHHPHHPH |
| S2-6 | HHHPPPHHPHPHHPHHPHHPHPPPPPPPHPHPPHPPPHPPHHHHHHPH |
| S2-7 | PHPPPPHPHHHPHPHHHHPHHPHHPPPHPHPPPHHHPPHHPPHHPPPH |
| S2-8 | PHHPHHHPHHHHPPHHHPPPPPPHPHHPPHHPHPPPHHPHPHPHHPPP |
| S2-9 | PHPHPPPPHPHPHPPHPHHHHHHPPHHHPHPPHPHHPPHPHHHPPPPH |
| S2-10 | PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHHPPHH |

Table 2. Performance comparison of various algorithms for the 3D HP protein folding

| Name | HZ | CHCC | CG | CI | PERM | ACO-H | ACO-F |
|---|---|---|---|---|---|---|---|
| S2-1 | 31 | 32 | 32 | 32 | 32 | 32 | **48**(35.15) |
| S2-2 | 32 | 34 | 34 | 33 | 34 | 34 | **49**(36) |
| S2-3 | 31 | 34 | 34 | 32 | 34 | 34 | **43**(32.6) |
| S2-4 | 30 | 33 | 33 | 32 | 33 | 33 | **43**(30.6) |
| S2-5 | 30 | 32 | 32 | 32 | 32 | 32 | **43**(35.15) |
| S2-6 | 29 | 32 | 32 | 30 | 32 | 32 | **43**(32.75) |
| S2-7 | 29 | 32 | 32 | 30 | 32 | 32 | **42**(33.8) |
| S2-8 | 29 | 31 | 31 | 30 | 31 | 31 | **42**(32.95) |
| S2-9 | 31 | 34 | 33 | 32 | 34 | 34 | **46**(34.44) |
| S2-10 | 33 | 33 | 33 | 32 | 33 | 33 | **46**(36.45) |

They are the solution quality obtained by hydrophobic zipper (HZ) [4] algorithm, the constrain-based hydrophobic core construction (CHCC) [12] method, the core-directed chain growth (CG) [2] algorithm, the contact interactions (CI) [10] algorithm, the pruned-enriched Rosenbluth method (PERM) [6], the ACO algorithm of Hoos (ACO-H) [9] and the ACO algorithm presented in this paper (ACO-F). For ACO-F the best found results is reported in bold and in the brackets is the average result over 20 runs. In the majority of the cases our average results are better than the best found results by other methods. And for all of the cases our best result is better then the best result of other methods. In ACO-H a local search

procedure is used to improve the results. ACO-F is without local search procedure. The main differences between these two ACO implementations are the location of the polar amino acids, the construction of the heuristic information and the pheromone updating.

## 5. CONCLUSION

In this work is shown that ACO can be successfully applied to the 3D protein folding problem. Our ACO algorithm outperforms other methods find in the literature. We have shown that the components of the ACO algorithm contribute to its performance. In particular, the performance is affected by the heuristic function and selectivity of pheromone updating. The obtained results are encouraging and the ability of the developed algorithm to generate rapidly high-quality solutions can be seen.

## REFERENCES

1. Berger B., T. Leighton, Protein Folding in the Hydrophobic-hydrophilic (HP) Model is NP-complete, *Computational Biology*, 1998, 5, 27-40.
2. Beutler T., K. Dill, A Fast conformational Method: A New Algorithm for Protein Folding Simulations, *Protein Sci*., 1996, 5, 147-153.
3. Dill K., K. Lau, A Lattice Statistical Mechanics Model of the Conformational Sequence Spaces of Proteins, *Macromolecules*, 1989, 22, 3986-3997.
4. Dill K., K. M. Fiebig, H. S. Chan, Cooperativity in Protein-folding Kinetics, *Nat. Acad. Sci*., USA, 1993, 1942-1946.
5. Dorigo M., L. M. Gambardella, Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem, *IEEE Transactions on Evolutionary Computing*, 1997, 1, 53-66.
6. Hsu H. P., V. Mehra, W. Nadler, P. Grassbergen, Growth Algorithm for Lattice Heteropolymers at Low Temperature, *Chemical Physics*, 2003, 118, 444-451.
7. Krasnogor N., D. Pelta , P. M. Lopez, P. Mocciola, E. de la Cana, Genetic Algorithms for the Protein Folding Problem: A Critical View, *Engineering of intelligent systems*, Alpaydin C. editor, ICSC Academic press, 1998, 353-360.
8. Liang F., W. H. Wong, Evolutionary Monte Carlo for Protein Folding Simulations, *Chemical Physics*, 2001, 115(7), 444-451.

9. Shmygelska A., H. H. Hoos, An Ant Colony Optimization Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem, *BMC Bioinformatics*, 2005, 6(30).
10. Toma L., S. Toma, Contact Interaction Method: A New Algorithm for Protein Folding Simulations, *Protein Sci,* 1996, 5, 147-153.
11. Unger R., J. Moult, Genetic Algorithms for Protein Folding Simulations, *Molecular Biology*, 1993, 231, 75-81.
12. Yue K., K. Dill, Forces of Tertiary Structural Organization in Globular Proteins, *Nat. Acad. Sci.*, USA, 1995, 146-150.