Analele Universităţii de Vest, Timişoara
Seria Matematică – Informatică
XLVII, 2, (2009), 33– 46

3D Protein Structure Prediction

Stefka Fidanova and Ivan Lirkov

Abstract. The protein folding problem is a fundamental problem in computational molecular biology and biochemical physics. The high resolution 3D structure of a protein is the key to the understanding and manipulating of its biochemical and cellular functions. All information necessary to fold a protein to its native structure is contained in its amino-acid sequence. Even under simplified 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 have become one of the major tools for studying protein structure. Various optimization methods have been applied on the folding problem including Monte Carlo methods, evolutionary algorithm, 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. Furthermore, we compare our folding results with proteins with known folding.

AMS Subject Classification (2000). 90C59
Keywords. Ant Colony Optimization, metaheuristics, hydrophobic-polar model, protein folding

Revised version of the paper presented at the Workshop on Computational Optimization, IMCSIT 2008, Wisla, Poland (http://www.imcsit.org)

1 Introduction

The number of amino acids and their sequence give a protein its individual characteristics. The number of amino acids in each protein ranges approximately between 20 and 40000, although most proteins are around hundred amino acids in length. Each protein's sequence of amino acids determines how it folds into a unique three dimensional structure that is its minimum energy state. Knowledge of 3D structure of proteins is crucial to pharmacology and medical sciences for the following important reasons. Most drugs work by attaching themselves to a protein so that they can either stabilize the normally folded structure or disrupt the folding pathway, which leads to a harmful protein. Thus, knowing exact 3D shapes will help to design drugs.

Determining the functionality of a protein molecule from amino acid sequence remains a central problem in computational biology, molecular biology, biochemistry, and physics. A system of differential equations is used to describe the forces, which affect the folding. It is very complicate and difficult to be solved. 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. The aim is to find a conformation, which is close to the real one and then to specify it using system of differential equations. So, as closer is the conformation, as less complex is the system of differential equations. Thus the computational time decreases. These models try to generally reflect different global characteristics of protein structures. In the hydrophobic-polar (HP) model [4] the primary amino acid sequence of a protein (which can be represented as a string over a twenty-letter alphabet) is abstracted to a sequence of hydrophobic (H) and polar (P) residues that is represented as a string over the letters H and P. It describes the proteins based on the fact that hydrophobic amino acids tend to be less exposed to the aqueous solvent than the polar ones, thus resulting in the formation of a hydrophobic core in the spatial structure. 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 non-consecutive 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 a NP-complete problem [2].

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. Ants indirectly communicate with each other by depositing pheromone trails on the ground and thereby influencing the decision processes of other ants. 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 metaheuristic methods using 10 sequences of 48 monomers from the literature and with real proteins with known folding.

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

2 The Protein Folding Problem

Efforts to solve the protein folding problem have traditionally been rooted in two schools of thought. One is based on the principles of physics: that is, the thermodynamic hypothesis, according to which the native structure of the protein corresponds to the global minimum of its free energy. The other school of thought is based on the principles of evolution. Thus methods have been developed to map the sequence of one protein (target) to the structure of another protein (template), to model the overall fold of the target based on that of the template and to infer how the target structure will be changed, related to the template, as a result of substitutions [1].

Accordingly methods for protein-structure prediction has been divided into two classes: de novo modelling and comparative modelling. The de novo approaches can be further subdivided, those based exclusively on the physics of the interactions within the polypeptide chain and between the polypeptide and solvent, using heuristic methods [9, 10], and knowledge-based methods that utilize statistical potential based on the analysis of recurrent patterns in known protein structures and sequences. The comparative modelling models structure by copying the coordinates of the templates in the aligned core regions. The variable regions are modelled by taking fragments with similar sequences from a database [1].

The processes involved 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 [4]. The HP model is based on the observation that hydrophobic interconnection is the driving force for protein folding and the hydrophobicity of amino acids is the main force for development of native conformation of small globular proteins. In the HP model, the primary amino acid sequence of a protein is abstracted to a sequence of hydrophobic (H) and polar (P) residues, amino acid components. 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 $E(c) = n \cdot (-1)$. 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 $E^s = E(c^s) = \min_{c \in C(s)} E(c)$, where $C(s)$ is the set of all valid conformations for s. It was proved that this problem is NP-hard [2].

A number of well-known heuristic optimization methods have been applied to the 3D protein folding problem including Evolutionary Algorithm (EA) [9], Monte Carlo (MC) algorithm [10] and Ant Colony Optimization (ACO) algorithm [7]. An early application of EA to protein structure prediction was presented by Unger and Moult [12]. Their EA incorporates characteristics of Monte Carlo methods. Currently among the best known algorithms for the HP protein folding problem is the Pruned-Enriched Rosenblum Method (PERM) [8]. Among these methods are the Hydrophobic Zipper (HZ) method [5] and the Constraint-based Hydrophobic Core Construction Method (CHCCM) [13]. The Core-direct chain Growth method (CG) [3] 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 repetition of the above mechanism represents the auto-catalytic behavior of real ant colony where the more the ants follow a trail, the more attractive that trail becomes.

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 a 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 the 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 τ_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 [6]. 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:

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \rho\tau_0 \tag{3.1}$$

Where τ_{ij} is an amount of the pheromone on the arc (i,j) of the 3D cube lattice, ρ is a persistence of the trail and the term $(1-\rho)$ can be interpreted as trail evaporation. The aim of the local update rule is to make better use of the pheromone information by dynamically changing the desirability of edges. Using this rule, ants will search in a wide neighborhood of the best previous solution. As is shown in the formula, the pheromone level on the paths is highly related to the value of evaporation parameter ρ. 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:

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \Delta\tau_{ij}, \tag{3.2}$$

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

Table 1: Standard benchmark instances

1	HPHHPPHHHHPHHHPPHHPPHPHHPHPHHPPHHPPPHPPPPPPPPHHP
2	HHHHPHHPHHHHHPPHPPHHPPHPPPPPPHPPHPPPHPPHHPPHHHPH
3	PHPHHPHHHHHHPPHPHPPHPHHPHPHPPPHPPHHPPHHPPHPHPPHP
4	PHPHHPPHPHHHPPHHPHHPPPHHHHHHPPHPHHPHPHPPPHPPHPHP
5	PPHPPPHPHHHHPPHHHHPHHPHHHPPHPHPHPPHPPPPPPHHPHHPH
6	HHHPPPHHPHPHHPHHPHHPHPPPPPPPHPHPPHPPPHPPHHHHHHPH
7	PHPPPPHPHHHPHPHHHHPHHPHHPPPHPHPPPHHHPPHHPPHHPPPH
8	PHPHPPPPHPHPHPPHPHHHHHHPPHHHPHPPHPHHPPHPHHHPPPPH
9	PHPHPPPPHPHPHPPHPHHHHHHPPHHHPHPPHPHHPPHPHHHPPPPH
10	PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHHPPHH

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 previous 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:

$$P_{ij} = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{k \in Unused} \tau_{ik}^{\alpha}\eta_{ik}^{\beta}} \tag{3.3}$$

Where τ_{ij} is the intensity of the pheromone deposited by each ant on the path (i, j), α is the intensity control parameter, η_{ij} is the heuristic information equal to the number of new H-H contacts if the position j is chosen, β is the heuristic parameter. Thus the higher the value of τ_{ij} and η_{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. The ACO algorithm is presented on Fig. 1.

Ant Colony Optimization

```
Initialize number of ants;
Initialize the ACO parameters;
while not end-condition do
    for k=0 to number of ants
        ant k starts from random node;
        while solution is not constructed do
            ant k selects a node with a probability;
        end while
    end for
    Local search procedure;
    Update-pheromone-trails;
end while
```

Figure 1: Pseudo-code for ACO

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 [3,7,9,10,12]. Experiments on these standard benchmark instances were conducted by performing 20 independent runs for each problem instance. The following parameter settings are used for all experiment as: $\alpha = \beta = 1$, $\rho = 0.5$. Furthermore, all pheromone values were initialized to $\tau_0 = 0.5$ and a population of 5 ants was used. The algorithm was terminated after 200 iterations. All experiments were performed on IBM ThinkPad Centrino 1.8 GHz CPU, 512 MB RAM running SuSe Linux.

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

We compared the solution quality obtained by: hydrophobic zipper (HZ) algorithm [5], the constrain-based hydrophobic core construction (CHCC) method [14], the core-directed chain growth (CG) algorithm [3], the contact interactions (CI) algorithm [12], the pruned-enriched Rosenbluth method (PERM) [7], the ACO algorithm of Hoos (ACO) [10] and the ACS approach presented in this paper. For ACS the best found result and the average result over 20 runs are reported. 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 than the best result of other methods. In ACO

Table 2: Comparison of 3D protein folding

Benchmark	HZ	CHCC	CG	CI	PERM	ACO	ACS	
							best	average
1	31	32	32	32	32	32	48	35.15
2	32	34	34	33	34	34	49	36
3	31	34	34	32	34	34	43	32.6
4	30	33	33	32	33	33	43	30.6
5	30	32	32	32	32	32	43	35.15
6	29	32	32	30	32	32	43	32.75
7	29	32	32	30	32	32	42	33.8
8	29	31	31	30	31	31	42	32.95
9	31	34	33	32	34	34	46	34.44
10	33	33	33	32	33	33	46	36.45

a local search procedure is used to improve the results. ACS approach is used without local search procedure. Which means that, if we combine our ACS algorithm with local search procedure, we can improve the achieved results. The main differences between ACO and ACS implementations are the location of the polar amino acids, the construction of the heuristic information and the pheromone updating. In ACO the authors put the polar amino acids on same direction as precedence amino acid, which is not the case in the nature. In our ACS we put the polar amino acids in a random way, thus we give to the ants more possibilities in a search process. We start the folding from the left end of the amino-acid sequence as it is done in the nature. In ACO algorithm authors start the folding from random amino acid in the middle of the protein chain. In our algorithm we take into consideration the real folding in the nature and as a result we achieve better folding with respect to the other algorithms. For heuristic information we use the number of new H-H contacts. In ACO algorithm the heuristic information is defined according to the Boltzman distribution as $\eta = e^{-\gamma h}$, where γ is a parameter and h is the number of new H-H contacts. After each construction phase we update the pheromone according ACS approach and they update the pheromone according MMAS approach [11].

For illustration, we compare two real proteins with known folding and the folding achieved by our ACS algorithm, which outperforms others on benchmark tests. Like test problems we choose Hepsidin and c-src Tyrosine Kinase Sh3 Domain (SrcSH3).

The Hepsidin consists of 21 amino acids: GCRFCCNCCPNMSGCGVCCRP. His folding comprises two crossed sheets and unstructured part between them (see Fig. 2).

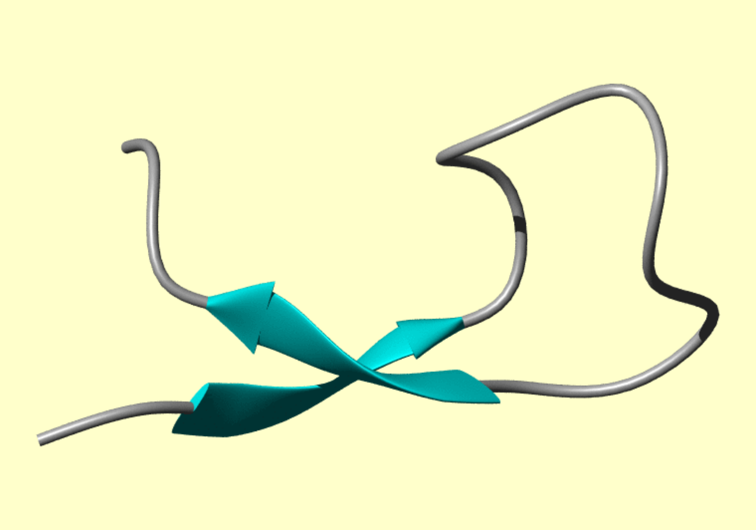

Figure 2: Hepsidin

The HP representation of the Hepsidin is: HPPHPPPPPHPHPHPHHPPPH. By our ACS algorithm we achieve the 3D folding represented on Fig. 3. The nodes represent amino acids and the lines represent their succession. We observe two tense, orthogonally situated parts. One of them consists of 3 amino acids and other consists of 4 amino acids. Between them we observe an unstructured part. Thus we can conclude that there is high similarity between the real Hepsidin folding and this obtained by our algorithm.

The main disadvantage of heuristic methods, as it is mentioned by other authors, is that they achieve good folding for short proteins only. Therefore for long proteins we cut the protein chain on shorter sub-chains. We apply folding algorithm on every subchain and at the end we assemble folded parts to fold entire protein.

The SrcSH3 protein consists of 62 amino acids. Its folding comprises two long parallelly situated sheets like a hairpin inside the protein and short sheets at the beginning and at the end of the protein, which are parallel each other and orthogonal to the hairpin, see Fig. 4.

By our ACS algorithm we achieve a folding represented on Fig. 5. The achieved H-H contacts are 19. We observe that there is not similarity between real folding and this achieved by our algorithm. Thus we prove the conclusion of other works [10], that heuristic methods are good for folding short proteins only. Therefore we decided to cut the HP chain of the SrcSH3 protein to short parts consisting of about 10-11 amino acids. We apply our ACS algorithm

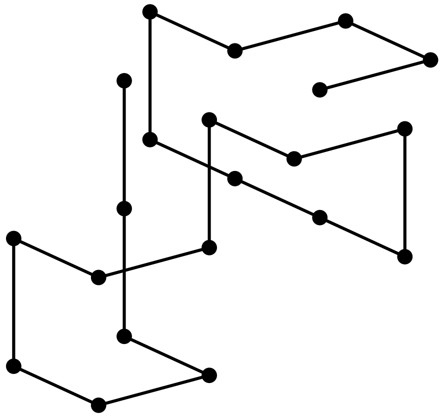

Figure 3: ACS Hepsidin

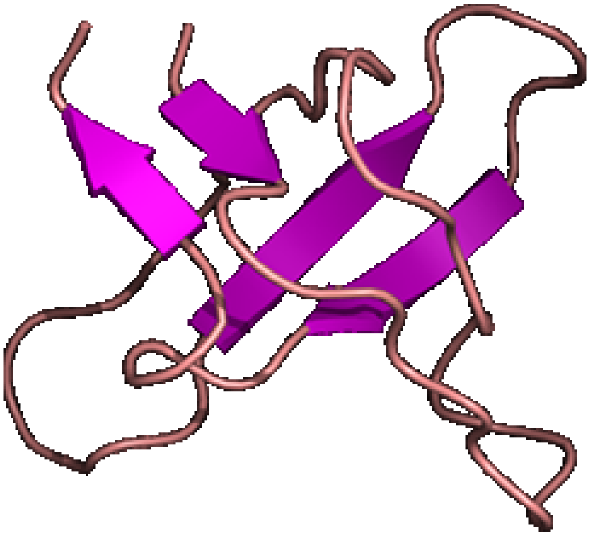

Figure 4: Tyrosine SrcSH3 folding

on every short part and at the end we assemble the folded parts to fold entire protein, see Fig. 6. The achieved H-H contacts are 20. We observe two tense long parallel parts like hairpin. One of them consists of 8 and other 7 amino acids. At one of the ends we observe short tense part orthogonal to the

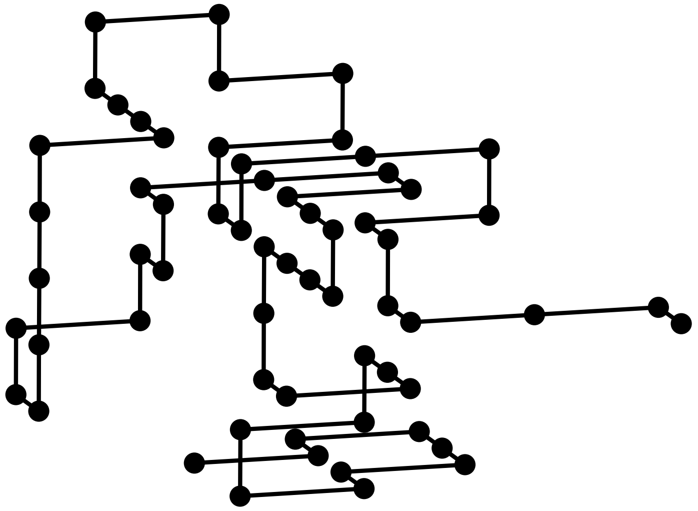

Figure 5: ACS Tyrosine SrcSH3 folding

hairpin. Other protein parts are unstructured. Thus we can conclude that there is a high similarity with the real folding of this protein.

5 Conclusion

Ant Colony System approach can be successfully applied to the 3D protein folding problem. Our algorithm outperforms well known methods from the literature. We have shown that the components of the algorithm contribute to its performance. In particular, the performance is affected by the heuristic function and selectivity of pheromone updating. The folding achieved by our algorithm is very similar to the real protein folding when it is applied on short proteins. When the protein is long, first we cut it on short parts, then we apply the algorithm on every one of the parts separately, finally we assemble the protein parts. Thus the achieved folding has high similarity to the real one. The obtained results are encouraging and the ability of the

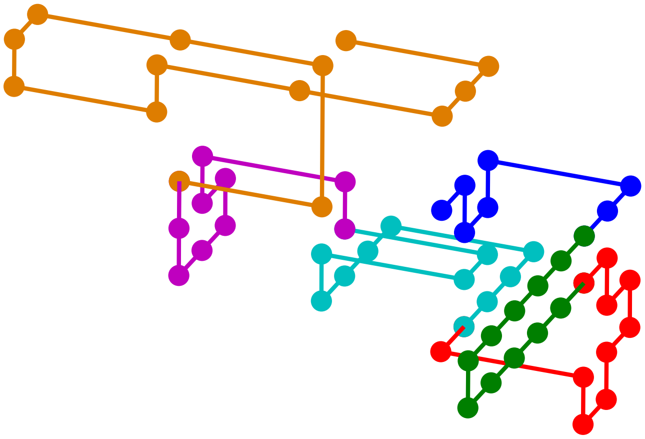

Figure 6: Partially Tyrosine SrcSH3 folding

developed algorithm to generate rapidly high-quality solutions can be seen. In the future we will develop and improve the folding algorithm. The aim is to achieve more realistic folding.

Acknowledgment

The work was supported by the Bulgarian Ministry of Education by the grants "Virtual screening and computer modeling for drug design" and DO-02-147.

References

[1] Balev S., *Solving the Protein Threading Problem by Lagrangian Relaxation*: *Algorithms in Bioinformatics*, Lecture Notes in Computer Sciences **3240** (2004), 182–193.

[2] Berger B. and Leighton T., *Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete*, Computational Biology **5** (1998), 27–40.

[3] Beutler T. and Dill K., *A fast conformational method: A new algorithm for protein folding simulations*, Protein Sci. **5** (1996), 147–153.

[4] Dill K. and Lau K., *A lattice statistical mechanics model of the conformational sequence spaces of proteins*, Macromolecules **22** (1989), 3986–3997.

[5] Dill K., Fiebig K. M., and Chan H. S., *Cooperativity in protein-folding kinetics*, Nat. Acad. Sci., USA (1993), 1942–1946.

[6] Dorigo M. and Gambardella L. M., *Ant colony system: A cooperative learning approach to the traveling salesman problem*, IEEE Transactions on Evolutionary Computing **1** (1997), 53–66.

[7] Hsu H. P., Mehra V., Nadler W., and Grassbergen P., *Growth algorithm for lattice heteropolymers at low temperature*, Chemical Physics **118** (2003), 444–451.

[8] Krasnogor N., Pelta D., Lopez P. M., Mocciola P., and de la Cana E., *Genetic algorithms for the protein folding problem: a critical view*, Engineering of intelligent systems, ICSC Academic press (1998), 353–360.

[9] Liang F. and Wong W. H., *Evolutionary Monte Carlo for protein folding simulations*, Chemical Physics **115(7)** (2001), 444–451.

[10] Shmigelska A. and Hoos H. H., *An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem*, BMC Bioinformatics **6:30** (2005).

[11] Stutzle T. and Hoos H. H., *MAX-MIN Ant System*, Future Generation Computer Systems **16(8)** (2000), 884–914.

[12] Toma L. and Toma S., *Contact interaction method: a new algorithm for protein folding simulations*, Protein Sci. **5** (1996), 147–153.

[13] Unger R. and Moult J., *Genetic algorithms for protein folding simulations*, Molecular Biology **231** (1993), 75–81.

[14] Yue K. and Dill K., *Forces of tertiary structural organization in globular proteins*, Nat. Acad. Sci, USA (1995), 146–150.

Stefka Fidanova

Institute for Parallel Processing
Bulgarian Academy of Sciences
Acad. G. Bonchev, bl. 25A
1113 Sofia
Bulgaria

E-mail: stefka@parallel.bas.bg

Ivan Lirkov

Institute for Parallel Processing
Bulgarian Academy of Sciences
Acad. G. Bonchev, bl. 25A
1113 Sofia
Bulgaria

E-mail: ivan@parallel.bas.bg

Received: 12.01.2009

Accepted: 15.04.2009