An Efficient Ant Colony Optimization Algorithm for Protein Structure Prediction

Dong Do Duc, Phuc Thai Dinh, Vu Thi Ngoc Anh, Nguyen Linh-Trung
AVITECH Institute, University of Engineering and Technology, Vietnam National University Hanoi, Vietnam

Abstract—**Protein structure prediction is considered as one of the most long-standing and challenging problem in bioinformatics. In this paper, we present an efficient ant colony optimization algorithm to predict the protein structure on three-dimensional face-centered cubic lattice coordinates, using the hydrophobic–polar model and the Miyazawa–Jernigan model to calculate the free energy. The reinforcement learning information is expressed in the k-order Markov model, and the heuristic information is determined based on the increase of the total energy. On a set of benchmark proteins, the results show a remarkable efficiency of our algorithm in comparison with several state-of-the-art algorithms.**

I. Introduction

Proteins are essential components of all living cells and play a vital role in biological processes of living organisms. They are sequential chains of amino acid connected by single-peptide bonds, and therefore also known as polypeptides. The three-dimensional (3D) structure of a protein exposes its properties and features. A misfolded protein can cause many dangerous diseases, such as Alzheimer, diabetes, cancer [1]. Analyzing the structure of proteins allow us to understand their features and produce medicines for diseases caused by protein misfolding [2], [3].

Unfortunately, it is complex and difficult to simulate a protein nature into 3D structure [4], [5]. Therefore, protein structure prediction (PSP) remains as a highly challenging problem for both the biological and computational communities. Several *in-vitro* methods were proposed to study proteins at atom-level like, such as X-ray crystallography, nuclear magnetic resonance (NMR). However, these methods is time-consuming and costly, unsuitable for large-scale situations. For this reason, computational methods for predicting the structure of proteins are promising alternatives [6], [7].

So far, there are three computational approaches: homology modeling, threading and *ab initio*. The first two approaches can only be used when compatible labels exist in the Protein Data Bank [8], limiting their applications. Methods in the *ab initio* approach predict the 3D structure of proteins, relying only on its primary amino acid sequence. From a given amino acid sequence, they predict the 3D structure of the protein by finding a unique 3D conformation with minimal interaction energy [4]. The model for solving this problem has been optimized by the search space and the target function.

In practice, the search space is very large and determining interaction energies is a complex and costly task. High-resolution methods can only handle proteins with length below 150 amino acids. That is why the lattice structure is used, wherein every amino acid corresponds to a node in a discretized search space. This simplicity allows developing highly efficient algorithms, especially when applied to longer proteins.

Many methods to apply the lattice structure have been considered [9]–[11], and among them, 3D face-centered cubic lattice (3D-FFC) possesses many advantages over other methods [12], [13] and have been used by many researchers [10], [14]–[16].

There are two popular energy models, aproximating the optimal structure of proteins: Hydrophobic–Polar (HP) energy model [10], [17] and Miyazawa–Jernigan (MJ) energy model [18]. In the HP model, every amino acid is considered a bead labelled as hydrophobic (H) and polar (P), and energy is determined from the physical interactions among H-nodes, whereas P-nodes are seen as neutral. The MJ model considers interactions between specific pairs of amino acids, thus being closer to the realistic model of free energy.

PSP has been classified as an NP-hard problem [19], [20], and so heuristic and metaheuristic algorithms have been proposed to solve it. Many of those are based on population, such as: ant colony optimization (ACO) [21], artificial learning system [22], generic algorithm (GA) [23]–[25], population-based algorithm [26], particle swarm optimization (PSO) [27], firefly algorithm [14]. Recently, Rashid *et al.* has proposed two methods based on the GA: GAplus [15] (HP energy model) and MH-GA [16] (graded energy, strategically mixing the MJ energy with the HP energy). The performance of these algorithms is outstanding in comparison with several the state of the art algorithms.

In this paper, we propose the K-ACO algorithm for PSP, in which the pheromone trail is calculated according to k-order Markov model, which is suitable for 3D structure reception. When using the HP energy model, a local search algorithm is applied to the best solution at each iteration step. Its effectiveness is shown by comparing the simulation study against GAPlus [15], TLS [28] MH-GA [16], Hybrid [29], Local Search [30].

The rest of this paper is organized as follows. In Section II, we briefly provide the background knowledge about the FCC lattice protein representation, the HP and MJ models and some related works. Section III is dedicated for the new algorithm, K-ACO. The simulation study is shown in Section IV. The conclusion is presented in the last section.

II. Problem Statement and Related Works

In this section, we briefly describe PSP from its native amino acid sequence in the FCC lattice representation of proteins, the objective functions (HP and MJ), some related works, and the ACO method.

A. FCC lattice and presentation of protein

The FCC lattice is obtained by discretizing the 3D space, formed around triangles. Each node only has 12 neighbors whose relative coordinates to the current node are $(1,1,0)$, $(1,1,0)$, $(1,1,0)$, $(1,1,0)$, $(0,1,1)$, $(0,1,1)$, $(1,0,1)$, $(1,0,1)$, $(0,1,1)$, $(1,0,1)$, $(0,1,1)$ and $(1,0,1)$. This is illustrated in Fig. 1. Given a primary amino acids sequence, a feasible protein sequence is a sequence where any pair of consecutive amino acids in the primary sequence are neighbors. Compared to other lattices, the FCC lattice is close to the natural structure of proteins, with many advantages [12], [13], such as highest packing density, smaller root mean square deviation values.

B. The energy models

Two energy models frequently used to determine the target function of this problem are the HP and MJ models.

TABLE I: Energy values between every protein pairs

	CYS	**MET**	**PHE**	**ILE**	**LEU**	**VAL**	**TRP**	**TYR**	**ALA**	**GLY**	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	**PRO**
CYS	-1.06	0.19	-0.23	0.16	-0.08	0.06	0.08	0.04	0.0	-0.08	0.19	-0.02	0.05	0.13	0.69	0.03	-0.19	0.24	0.71	0.0
MET	0.19	0.04	-0.42	-0.28	-0.2	-0.14	-0.67	-0.13	0.25	0.19	0.19	0.14	0.46	0.08	0.44	0.65	0.99	0.31	0.0	-0.34
PHE	-0.23	-0.42	-0.44	-0.19	-0.3	-0.22	-0.16	0.0	0.03	0.38	0.31	0.29	0.49	0.18	0.27	0.39	-0.16	0.41	0.44	0.2
ILE	0.16	-0.28	-0.19	-0.22	-0.41	-0.25	0.02	0.11	-0.22	0.25	0.14	0.21	0.36	0.53	0.35	0.59	0.49	0.42	0.36	0.25
LEU	-0.08	-0.2	-0.3	-0.41	-0.27	-0.29	-0.09	0.24	-0.01	0.23	0.2	0.25	0.26	0.3	0.43	0.67	0.16	0.35	0.19	0.42
VAL	0.06	-0.14	-0.22	-0.25	-0.29	-0.29	-0.17	0.02	-0.1	0.16	0.25	0.18	0.24	0.5	0.34	0.58	0.19	0.3	0.44	0.09
TRP	0.08	-0.67	-0.16	0.02	-0.09	-0.17	-0.12	-0.04	-0.09	0.18	0.22	0.34	0.08	0.06	0.29	0.24	-0.12	-0.16	0.22	-0.28
TYR	0.04	-0.13	0.0	0.11	0.24	0.02	-0.04	-0.06	0.09	0.14	0.13	0.09	-0.2	-0.2	-0.1	0.0	-0.34	-0.25	-0.21	-0.33
ALA	0.0	0.25	0.03	-0.22	-0.01	-0.1	-0.09	0.09	-0.13	-0.07	-0.09	-0.06	0.08	0.28	0.26	0.12	0.34	0.43	0.14	0.1
GLY	-0.08	0.19	0.38	0.25	0.23	0.16	0.18	0.14	-0.07	-0.38	-0.26	-0.16	-0.06	-0.14	0.25	-0.22	0.2	-0.04	0.11	-0.11
THR	0.19	0.19	0.31	0.14	0.2	0.25	0.22	0.13	-0.09	-0.26	0.03	-0.08	-0.14	-0.11	0.0	-0.29	-0.19	-0.35	-0.09	-0.07
SER	-0.02	0.14	0.29	0.21	0.25	0.18	0.34	0.09	-0.06	-0.16	-0.08	0.2	-0.14	-0.14	-0.26	-0.31	-0.05	0.17	-0.13	0.01
GLN	0.05	0.46	0.49	0.36	0.26	0.24	0.08	-0.2	0.08	-0.06	-0.14	-0.14	0.29	-0.25	-0.17	-0.17	-0.02	-0.52	-0.38	-0.42
ASN	0.13	0.08	0.18	0.53	0.3	0.5	0.06	-0.2	0.28	-0.14	-0.11	-0.14	-0.25	-0.53	-0.32	-0.3	-0.24	-0.14	-0.33	-0.18
GLU	0.69	0.44	0.27	0.35	0.43	0.34	0.29	-0.1	0.26	0.25	0.0	-0.26	-0.17	-0.32	-0.03	-0.15	-0.45	-0.74	-0.97	-0.1
ASP	0.03	0.65	0.39	0.59	0.67	0.58	0.24	0.0	0.12	-0.22	-0.29	-0.31	-0.17	-0.3	-0.15	0.04	-0.39	-0.72	-0.76	0.04
HIS	-0.19	0.99	-0.16	0.49	0.16	0.19	-0.12	-0.34	0.34	0.2	-0.19	-0.05	-0.02	-0.24	-0.45	-0.39	-0.29	-0.12	0.22	-0.21
ARG	0.24	0.31	0.41	0.42	0.35	0.3	-0.16	-0.25	0.43	-0.04	-0.35	0.17	-0.52	-0.14	-0.74	-0.72	-0.12	0.11	0.75	-0.38
LYS	0.71	0.0	0.44	0.36	0.19	0.44	0.22	-0.21	0.14	0.11	-0.09	-0.13	-0.38	-0.33	-0.97	-0.76	0.22	0.75	0.25	0.11
PRO	0.0	-0.34	0.2	0.25	0.42	0.09	-0.28	-0.33	0.1	-0.11	-0.07	0.01	-0.42	-0.18	-0.1	0.04	-0.21	-0.38	0.11	0.26

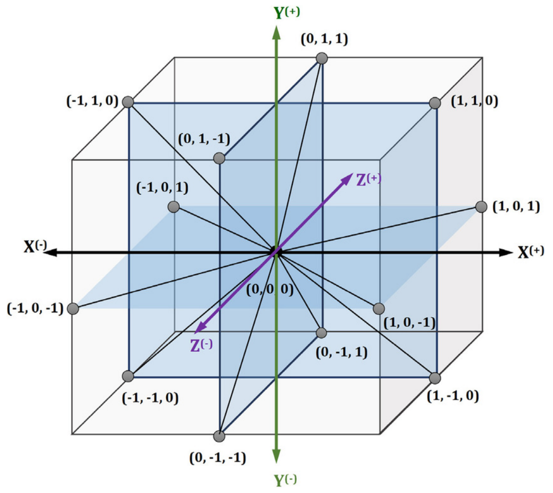

Fig. 1: Basis vectors of 12 neighbors of the origin $(0,0,0)$.

1) HP energy model: The HP energy model proposed by Lau and Dill in 1972 [17]. In this model, the amino acids Gly, Ala, Pro, Val, Leu, Ile, Met, Phe, Tyr, Trp are labeled as hydrophobic (H), others are labeled as polar (P). Two consecutive H-labeled amino acids will create negative energy (-1). The complete HP energy of the model for two amino acids i and j is calculated by

$$E_{\mathrm{HP}} = \sum_{i<j-1} c_{ij} * e_{ij}, \tag{1}$$

where

$$c_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ not consecutive but neighbors,} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

$$e_{ij} = \begin{cases} -1, & \text{if } i \text{ and } j \text{ both hydrophobic,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

2) MJ energy model: Relying on the interactive trend of amino acids, Miyazawa and Jernigan proposed the MJ energy model in 1985 [31]. The complete MJ energy is calculated by

$$E_{\mathrm{MJ}} = \sum_{i<j-1} c_{ij} * e_{ij}, \tag{4}$$

where c_{ij} is determined by Eq. (2) and E_{ij} is taken from Table I.

C. The optimal problem and related algorithms

The optimal problem: for each given protein with the native amino acid sequence of length m, the PSP problem is transformed into finding the representation with optimal E_{HP} or E_{MJ} energy.

Recently, MH-GA [16] has been proven to be the most efficient algorithm to solve the PSP problem by comparing its experimental results with the MJ model against other state-of-the-art algorithms, such as Hybrid algorithm [29], and Local Search [30].

III. The Proposed K-ACO Algorithm

ACO is a stochastic metaheuristic method proposed by Dorigo [32] for the traveling salesman problem (TSP). Many variants have been developed to tackle difficult optimization problems. In this paper, we build a structure graph and transform the original problem into a problem where solutions can be found by sequentially executing a certain procedure on the built structure graph. An ant colony executes the said procedure based on heuristic and reinforcement learning information (i.e., pheromone) in a random manner. When a solution is found, the algorithm appraises it then updates the pheromone to improve the chance of finding better solutions on the next searches, this is repeated till the termination requirement is met. The properties affecting the quality of the algorithm are: (i) a suitable structure graph, (ii) heuristic information, and (iii) how pheromone is stored and updated.

A. Construction graph

Without loss of generality, the first amino acid is placed at the origin $(0,0,0)$ and start there. The 12 neighbors of each node are indexed from 1 to 12. The structure graph for a protein with the length of m has $(m-1)$ columns put in order after the start vertex. There are edges directed from each vertex to all vertices in the next column. The graph is illustrated in Fig. 2. With this, any feasible sequence of length m will correspond to a path on this graph.

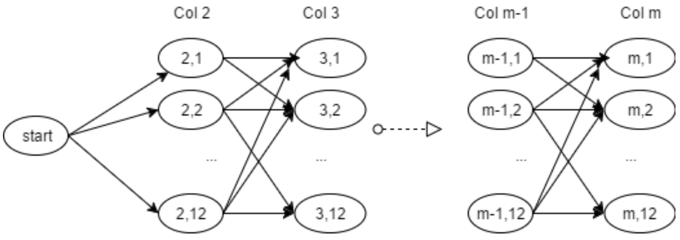

Fig. 2: Construction graph.

B. Randomized procedure to find solution

Each ant will begin at the start vertex and randomly select a vertex on the next column to go. Suppose the ant is on vertex i of column n (or the start vertex), it will select vertex j out of 12 vertices on the next column with the probability $P_{i,j}$ calculated by the following formula:

$$P_{i,j} = \frac{[\tau_{i,j(k)}]^{\alpha}[\eta_{i,j}]^{\beta}}{\sum_{l \in \mathcal{C}_{n+1}} [\tau_{i,l(k)}]^{\alpha}[\eta_{i,l}]^{\beta}}, \tag{5}$$

where $\eta_{i,j}$ is the heuristic information (see III-C), $\tau_{i,j(k)}$ is the pheromone information of the k-degree Markov model (see III-D), $\mathcal{C}_t$ is the set of vertices on column t, α and β are parameters of the ACO system, deciding the impact of heuristic and pheromone information on making decisions.

To ensure self-avoiding walk constraint, we set $P_{i,j} = 0$ when selecting vertex would cause two amino acids to have the same coordinate on the protein representation.

C. Heuristic information

After the first $(i-1)$ amino acids were successfully represented and vector j is the selected direction to go next, let η_{ij} be the heuristic value, E_{ij} be the amount of increased energy, and $E_{\max} = \max(E_{ij})$. Then $\eta_{ij} = E_{\max} - E_{ij} + \epsilon$, where ϵ is a small positive number to ensure η_{ij} always positive. In our implements, we set it to 0.01.

D. Pheromone update

Instead of making choice based only on the pheromone information in the current column, we can also take previously selected vertices into consideration. Let $\tau_{i,j(k)}$ be the pheromone when vertices $(i,j), (i-1, v_{i-1}), \ldots, (i-k+1, v_{i-k+1})$ are selected. This way, the pheromone will give more accurate information during the searches.

After every round of search, we update pheromone using the SMMAS algorithm [33], by

$$\tau_{i,j(k)} = (1-\rho)\tau_{i,j(k)} + \Delta_{ij}, \tag{6}$$

where

$$\Delta_{ij} = \begin{cases} \rho\tau_{\min}, & \text{if } (i,j) \in T, \\ \rho\tau_{\min}, & \text{otherwise.} \end{cases} \tag{7}$$

Above, T is the set of selected vertices in the best solution found in this round.

E. Local Search

At each step of the local search procedure, we first identify the hydrophobic core center (HCC) as the center of the hydrophobic amino acid (H). The coordinates of HCC are determined as follows:

$$x_{\text{HCC}} = \frac{1}{n_H}\sum_{i=1}^{n_H} x_i,\ y_{\text{HCC}} = \frac{1}{n_H}\sum_{i=1}^{n_H} y_i,\ z_{\text{HCC}} = \frac{1}{n_H}\sum_{i=1}^{n_H} z_i, \tag{8}$$

where n_H is the number of amino acids H. Then, we choose an amino acid H to move closer to the HCC so as not to increase the free energy of the protein.

Algorithm 1 Procedure of Local Search

```
while stop conditions not satisfied do
    Calculate the HCC coordinates
    $Move \leftarrow SeclectMove()$
    if Move = Null then
        Break
    AplIyMove()
```

Algorithm 2 Procedure of K-ACO algorithm

```
Initialize pheromone trail matrix and set $A$ of $p$ ants
while stop conditions not satisfied do
    for $a \in A$ do
        Ant a build a solution by random walk procedure
    Update pheromone trail follows SMMAS rule
    Use local search on the best solution
    Update the best solution
Decode solution and save the best solution
```

IV. Simulation

A. Different values of K

E_{MJ} is the average of energy values returned by our algorithm and N_{loops} is the average of the number of loops that our algorithm will be convergent. From Table II, we see that the number of loops needed for convergence increases when K increases. However, the value of E_{MJ} increases significantly when K increases from 1 to 3. Values of E_{MJ} when $K \in \{3,4,5\}$ do not differ much. The larger K, the more running time and memory our algorithm needed to complete. Hence, we choose $K = 3$ as default for the algorithm.

B. HP energy model

The data sets were used are H,F90,S,F180,R (Peter Clote laboratory[1]) and 3MSE, 3MR7, 3MQZ, 3NO6, 3NO3, 3ON7 from Critical Assessment of Protein Structure Prediction competition[2], used in [15].

[1] http://bioinformatics.bc.edu/clotelab/FCCproteinStructure.
[2] http://predictioncenter.org.

TABLE II: The result when trying multiple values of K

K	3NO3		3NO6		3ON7	
	E_{MJ}	N_{loops}	E_{MJ}	N_{loops}	E_{MJ}	N_{loops}
1	-110.29	494	-118.56	456	-120.18	565
2	-128.36	1043	-134.67	1126	-136.8	1247
3	-141.03	2230	-150.13	2371	-154.8	2612
4	-141.99	3104	-150.44	3462	-154.26	3790
5	-141.24	3407	-148.62	3821	-154.34	4207

TABLE III: Results when HP energy model was used

Protein details				State-of-the-art					ACO			
SEQ	size	HS	LBFE	TLS		GA plus						
				best	avg	best	avg	time(s)	best	avg	time(s)	RI(%)
H1	48	24	-69	-68	-66	**-69**	**-69**	1800	**-69**	**-69**	308	0.00
H2	48	24	-69	-68	-65	**-69**	**-69**		**-69**	**-69**	321	0.00
H3	48	24	-72	-69	-66	**-72**	**-72**		**-72**	**-72**	316	0.00
H4	48	24	-71	-70	-65	**-71**	**-71**		**-71**	**-71**	316	0.00
H5	48	24	-70	-68	-65	**-70**	**-70**		**-70**	**-70**	321	0.00
H6	48	24	-70	-69	-66	**-70**	**-69**		**-70**	**-70**	324	1.45
H7	48	24	-70	-69	-66	**-70**	**-70**		**-70**	**-70**	320	0.00
H8	48	24	-69	-67	-64	**-69**	**-69**		**-69**	**-69**	320	0.00
H9	48	24	-71	-68	-66	**-71**	**-71**		**-71**	**-71**	313	0.00
H10	48	24	-68	-68	-65	**-68**	**-68**		**-68**	**-68**	324	0.00
F90_1	90	50	-168	-164	-160	**-168**	**-166**	7200	**-168**	**-166**	584	0.00
F90_2	90	50	-168	-165	-158	**-168**	**-165**		**-167**	**-165**	589	0.00
F90_3	90	50	-167	-165	-159	**-167**	**-164**		-165	-163	596	-0.61
F90_4	90	50	-168	-165	-159	**-168**	**-165**		-167	-163	592	-1.21
F90_5	90	50	-167	-165	-159	**-167**	**-166**		**-167**	**-166**	590	0.00
S1	135	100	-357	-351	-341	-355	-348		**-357**	**-354**	878	1.72
S2	151	100	-360	-355	-343	**-356**	-349		**-356**	**-352**	996	0.86
S3	162	100	-367	-355	-340	**-361**	-348		-359	**-353**	1062	1.44
S4	164	100	-370	-354	-343	**-364**	-352		-360	**-355**	1077	0.85
F180_1	180	100	-378	-338	-326	-351	-341	18000	**-352**	**-343**	1194	0.59
F180_2	180	100	-381	-345	-333	**-362**	**-346**		-350	-343	1185	-0.87
F180_3	180	100	-378	-352	-338	-361	-350		**-363**	**-357**	1189	2.00
R1	200	100	-384	-332	-318	**-355**	**-345**		-353	-341	1341	-1.16
R2	200	100	-383	-337	-324	**-360**	**-346**		-347	-337	1359	-2.60
R3	200	100	-385	-339	-323	**-363**	**-344**		-346	-337	1342	-2.03
3MSE	179	84	-323	-268	-251	**-292**	**-278**		-286	**-278**	1312	0.00
3MR7	189	93	-355	-304	-287	**-330**	-316		-326	**-318**	1324	0.63
3MQZ	215	120	-474	-404	-384	**-427**	-412		-426	**-415**	1547	0.73
3NO6	229	116	-455	-390	-372	**-423**	**-402**	28800	-410	-400	1689	-0.50
3NO3	258	122	-494	-388	-372	-421	-404		**-425**	**-411**	1751	1.73
3ON7	279	146	u/k	-491	-461	**-519**	-490		-510	**-495**	1803	0.00

To evaluate the performance of K-ACO, we use Relative Improvement (RI), defined as

$$\text{RI} = \frac{E_A - E_B}{E_B}, \tag{9}$$

where E_A and E_B are the average energy values achieved by the K-ACO algorithm and by the state-of-the-art one, respectively. K-ACO was compared with two other algorithms: TLS [28] and GA [15]. For each protein, each of the three algorithms were run 50 times. Table III shows the best and the average result of 50 runs for each protein. It can be seen that K-ACO performed better as compared to TLS. However, K-ACO and GA performed similarly; the difference between them always below 3%. K-ACO performed better than GA in 10 protein sequences while GA better than K-ACO in 7 protein sequences. To further compare with GA, we increased the number of loops to $60,000$ and applied this new change for those 7 protein sequences where GA did better. We see that, when increasing the number of loops, K-ACO performance improved and approximately as good as GA, as shown in Table V.

C. MJ energy model

In this section, data in Table IV were used for the MJ energy model. These data were also used in [16].

We run K-ACO on the above dataset and compare the result with other algorithms, namely Hybrid [29], Local search [30] and GA [15]. This is the best and average result taken from 50 runs for each protein sequence. From the column RI in Table VII, we can see that for all proteins sequences, our algorithm improved the average energy.

V. Conclusion

In this paper, we presented the K-ACO algorithm to predict the protein structure on the FCC lattice, using two different energy models– HP and MJ. This algorithm has a simple structure graph, the use of pheromone information in the k-order Markov model is more suitable for the 3D structure prediction and increase the efficiency of the ACO method. The simulation study shows that the proposed algorithm outperforms the state-of-the-art algorithms both in quality and running time. The algorithm can be improved by applying local search techniques according to memetic schemes. In this algorithm, the pheromone trail in the k-order Markov model with $k = 3$ is appropriate. Increasing k costs more memory and time, but the efficiency is not much improved. This technique can be applied to ACO algorithms for other similar problems.

TABLE V: K-ACO vs GA with increased running time

Protein details				GA plus			K-ACO		
SEQ	size	HS	LBFE	best	avg	time(s)	best	avg	time(s)
F90_3	90	50	-167	**-167**	**-164**	7200	-165	**-164**	1763
F90_4	90	50	-168	**-168**	**-165**	7200	-167	**-165**	1782
F180_2	180	100	-381	**-362**	**-346**	18000	-350	**-346**	3496
R1	200	100	-384	**-355**	**-345**	18000	-353	**-345**	4107
R2	200	100	-383	**-360**	**-346**	18000	-348	-340	4092
R3	200	100	-385	**-363**	**-344**	18000	-346	-340	4128
3NO6	229	116	-455	**-423**	-402	28800	-411	**-404**	5092

TABLE IV: Benchmark proteins used in our experiments with MJ model

ID	Length	Protein sequence
4RXN	54	MKKYTCTVCGYIYNPEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDQFEEVEE
1ENH	54	RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI
4PTI	58	RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
2IGD	61	MTPAVTTYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATKTFTVTE
1YPA	64	MKTEWPELVGKAVAAAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAQVPRVG
1R69	69	SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALGVSVDWLLNGTSDSNVR
1CTF	74	AAEEKTEFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKEGVSKDDAEALKKALEEAGAEVEVK
3MX7	90	MTDLVAVWDVALSDGVHKIEFEHGTTSGKRVVYVDGKEEIRKEWMFKLVGKETFYVGAAKTKATINIDAISGFA YEYTLEINGKSLKKYM
3NBM	108	SNASKELKVLVLCAGSGTSAQLANAINEGANLTEVRVIANSGAYGAHYDIMGVYDLIILAPQVRSYYREMKVDAERLGIQIVATRGMEYIHLTKSPSKALQFVLEHYQ
3MQO	120	PAIDYKTAFHLAPIGLVLSRDRVIEDCNDELAAIFRCARADLIGRSFEVLYPSSDEFERIGERISPVMIAHGSYADDRIMKRAGGELFWCHVTGRALDRTAPLAAGVWTFEDLSATRRVA
3MRO	142	SNALSASEERFQLAVSGASAGLWDWNPKTGAMYLSPHFKKIMGYEDHELPDEITGHRESIHPDDRARVLAALKAHLEHRDTYDVEYRVRTRSGDFRWIQSRGQALWNSAGEPYRMVGWIMDVTDRKRDEDALRVSREELRRL
3PNX	160	GMENKKMNLLLFSGDYDKALASLIIANAAREMEIEVTIFCAFWGLLLLRDPEKASQEDKSLYEQAFSSLTPREAEELPLSKMNLGGIGKKMLLEMMKEEKAPKLSDLLSGARKKEVKFYACQLSVEIMGFKKEELFPEVQIMDVKEYLKNALESDLQLFI
3MSE	180	GISPNVLNNMKSYMKHSNIRNIIINIMAHELSVINNHIKYINELFYKLDTNHNGSLSHREIYTVLASVGIKKWDINRILQALDINDRGNITYTEFMAGCYRWKNIESTFLKAAFNKIDKDEDGYISKSDIVSLVHDKVLDNNDIDNFFLSVHSIKKGIPREHIINKISFQEFKDYMLSTF
3MR7	189	SNAERRLCAILAADMAGYSRLMERNETDVLNRQKLYRRELIDPAIAQAGGQIVKTTGDGMLARFDTAQAALRCALEIQQAMQQREEDTPRKERIQYRIGINIGDIVLEDGDIFGDAVNVAARLEAISEPGAICVSDIVHQITQDRVSEPFTDLGLQKVKNITRPIRVWQWVPDADRDQSHDPQPSHVQH
3MQZ	215	SNAMSVQTIERLQDYLLPEWVSIFDIADFSGRMLRIRGDIRPALLRLASRLAELLNESPGPRPWYPHVASHMRRRVNPPPETWLALGPEKRGYKSYAHSGVFIGGRGLSVRFILKDEAIEERKNLGRWMSRSGPAFEQWKKKVGDLRDFGPVHDDPMADPPKVEWDPRVFGERLGSLKSASLDIGFRVTFDTSLAGIVKTIRTFDLLYAEAEKGS
3NO3	238	GKDNTKVIAHRGYWKTEGSAQNSIRSLERASEIGAYGSEFDVHLTADNVLVVYHDNDIQGKHIQSCTYDELKDLQLSNGEKLPTLEQYLKRAKKLKNIRLIFELKSHDTPERNRDAARLSVQMVKRMKLAKRTDYISFNMDACKEFIRLCPKSEVSYLNGELSPMELKELGFTGLDYHYKVLQSHPDWVKDCKVLGMTSNVWTVDDPKLMEEMIDMGVDFITTDLPEETQKILHSRAQ
3NO7	248	MGSDKIHHHHHHENLYFQGMTFSKELREASRPIIDDIYNDGFIQDLLAGKLSNQAVRQYLRADASYLKEFTNIYAMLIPKMSSMEDVKFLVEQIEFMLEGEVEAHEVLADFINEPYEEIVKEKVWPPSGDHYIKHMYFNAFARENAAFTIAAMAPCPYVYAVIGKRAMEDPKLNKESVTSKWFQFYSTEMDELVDVFDQLMDRLTKHCSETEKKEIKENFLQSTIHERHFFNMAYINEKWEYGGNNNE
3ON7	280	GMKLETIDYRAADSAKRFVESLRETGFGVLSNHPIDKELVERIYTEWQAFFNSEAKNEFMFNRETHDGFFPASISETAKGHTVKDIKEYYHVYPWGRIPDSLRANILAYYEKANTLASELLEWIETYSPDEIKAKFSIPLPEMIANSHKTLLRILHYPPMTGDEEMGAIRAAAHEDINLITVLPTANEPGLQVKAKDGSWLDVPSDFGNIIINIGDMLQEASDGYFPSTSHRVINPEGTDKTKSRISLPLFLHPHPSVVLSERYTADSYLMERLRELGVL

TABLE VII: K-ACO vs other algorithms (bold values are the best one in their row)

Protein details			Hybrid		Local search		GA		K-ACO		
SEQ	size	H	best	avg	best	avg	best	avg	best	avg	RI(%)
4RXN	54	27	-32.61	-30.94	-33.33	-31.21	-36.36	-33.6	**-37.98**	**-36.84**	9.64
1ENH	54	19	-35.81	-35.07	-29.03	-28.18	**-38.39**	-35.67	-37.51	**-36.49**	2.3
4PTI	58	32	-32.07	-29.37	-31.16	-28.33	-35.65	-31.01	**-37.2**	**-33.35**	7.55
2IGD	61	25	**-38.64**	-32.54	-32.36	-28.29	-36.49	-33.75	-36.77	**-35.09**	3.97
1YPA	64	38	n/a	n/a	-33.33	-32.15	-40.14	-36.33	**-40.52**	**-38.93**	7.16
1R69	69	30	-34.2	-31.85	-33.35	-32.2	**-40.85**	-36.28	-39.73	**-38.59**	6.37
1CTF	74	42	-38	-35.28	-45.83	-40.94	-51.5	-47.29	**-53.72**	**-51.09**	8.04
3MX7	90	44	n/a	n/a	-44.81	-42.32	-56.32	-50.95	**-58.1**	**-56.04**	9.99
3NBM	108	56	n/a	n/a	-52.44	-49.51	-49.51	-49.9	**-59.71**	**-57.5**	15.23
3MQO	120	68	n/a	n/a	-64.04	-58.84	-62.25	-54.56	**-70.62**	**-67.5**	14.72
3MRO	142	63	n/a	n/a	-87.38	-82.24	-90.05	-82.32	**-101.34**	**-98.2**	19.29
3PNX	160	84	n/a	n/a	-103.04	-96.86	-102.55	-88.06	**-116.31**	**-112.18**	15.82
3MSE	180	83	n/a	n/a	n/a	n/a	-92.61	-84.6	**-110.9**	**-106.44**	25.82
3MR7	189	88	n/a	n/a	n/a	n/a	-93.65	-83.93	**-120.64**	**-115.02**	37.04
3MQZ	215	115	n/a	n/a	n/a	n/a	-104.29	-95.22	**-132.09**	**-126.62**	32.98
3NO3	238	102	n/a	n/a	n/a	n/a	-122.97	-108.7	**-151.84**	**-147.86**	36.03
3NO7	248	112	n/a	n/a	n/a	n/a	-133.95	-117.11	**-163.89**	**-156.01**	33.22
3ON7	280	135	n/a	n/a	n/a	n/a	-116.88	-96.64	**-167.12**	**-160.29**	65.86

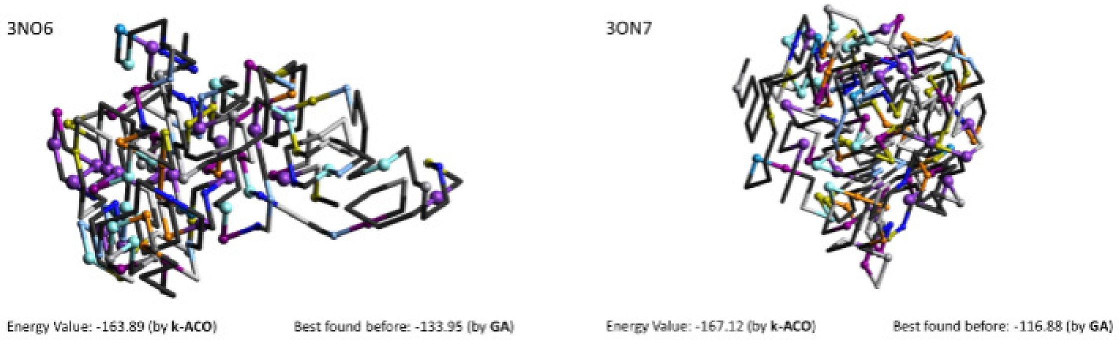
3NO6
Energy Value: -163.89 (by k-ACO)
Best found before: -133.95 (by GA)
3ON7
Energy Value: -167.12 (by k-ACO)
Best found before: -116.88 (by GA)

Fig. 3: New best structure found by K-ACO for two largest datasets.

TABLE VI: Running time of K-ACO and GA

Protein details			K-ACO	GA
SEQ	size	H		
4RXN	54	27	706.97	3600
1ENH	54	19	708.4	
4PTI	58	32	770.32	
2IGD	61	25	798.04	
1YPA	64	38	848.82	
1R69	69	30	916.28	
1CTF	74	42	991.53	
3MX7	90	44	1183.9	
3NBM	108	56	1414.94	
3MQO	120	68	1584.95	
3MRO	142	63	1831.22	
3PNX	160	84	2061.74	
3MSE	180	83	2337.52	7200
3MR7	189	88	2461.5	
3MQZ	215	115	2806.42	
3NO3	238	102	3053.11	
3NO6	248	112	3154.14	
3ON7	280	135	3576.92	

REFERENCES

[1] C. M. Dobson, "Protein folding and misfolding," *Nature*, vol. 426, no. 6968, pp. 884–890, 2003.

[2] A. Breda, N. F. Valadares, O. N. de Souza, and R. C. Garratt, "Protein structure, modelling and applications," in *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*, A. Gruber, A. Durham, and C. Huynh, Eds. Oxford University Press, 2007.

[3] P. Veerapandian, *Structure-based drug design*, 1997, vol. 11, no. 32.

[4] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[5] A. Bruce, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walters, "The shape and structure of proteins," *Molecular Biology of the Cell*, 2002.

[6] C. A. Floudas, "Computational methods in protein structure prediction," *Biotechnology and Bioengineering*, vol. 97, pp. 207–213, 2007.

[7] C. M. Dobson, "Computational biology: protein predictions," pp. 176–177, 2007.

[8] H. Berman, "The protein data bank," *Nucleic Acids Res*, pp. 235–242, 2000.

[9] A. Bechini, "On the characterization and software implementation of general protein lattice models," *PLoS ONE*, 2013.

[10] I. Dotu, M. Cebrian, P. V. Hentenryck, and P. Clote, "On lattice protein structure prediction revisited," *IEEE/ACM Tr. Comp. Biol Bioinfo.*, 2011.

[11] M. Mann and R. Backofen, "Exact methods for lattice protein models," *Bio-Algorithms and Med-Systems*, vol. 10, no. 4, pp. 213–225, 2014.

[12] D. Covell and R. Jernigan, "Conformations of folded proteins in restricted spaces," *Biochemistry*, pp. 3287–94, 1990.

[13] T. C. Hales, "A proof of the kepler conjecture," *The Annals of Mathematics*, vol. 162, no. 3, pp. 1065–1185, 2005.

[14] B. Maher, A. A. Albrecht, M. Loomes, X.-S. Yang, and K. Steinhfel, "A firefly-inspired method for protein structure prediction in lattice models," *Biomolecules*, pp. 56–75, 2014.

[15] M. A. Rashid, F. Khatib, M. T. Hoque, and A. Sattar, "An enhanced genetic algorithm for ab initio protein structure prediction," *IEEE Transactions on Evolutionary Computation*, vol. 20, pp. 627–644, 2016.

[16] M. A. Rashid, S. Iqbal, F. Khatib, M. T. Hoque, and A. Sattar, "Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction," *Comp. Biology and Chemistry*, pp. 162–177, 2016.

[17] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, 1989.

[18] S. Miyazawa and R. L. Jernigan, "Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 623–644, 1996.

[19] R. Unger and J. Moult, "Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications," *Bulletin of Mathematical Biology*, vol. 55, no. 6, pp. 1183–1198, 1993.

[20] M. Paterson and T. Przytycka, "On the complexity of string folding," *Discrete Applied Mathematics*, vol. 71, no. 1-3, pp. 217–230, 1996.

[21] A. Shmygelska and H. H. Hoos, "An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem," *BMC Bioinformatics*, vol. 6, no. 1, p. 30, 2005.

[22] V. Cutello, G. Nicosia, M. Pavone, and J. Timmis, "An immune algorithm for protein structure prediction on lattice models," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 1, pp. 101–117, 2007.

[23] R. Unger and J. Moult, "A genetic algorithm for 3D protein folding simulations," in *5th Intl. Conf. Genetic Algorithms*, 1993, p. 581.

[24] M. T. Hoque, M. Chetty, and A. Sattar, "Protein folding prediction in 3D FCC HP lattice model using genetic algorithm," in *IEEE Congress on Evolutionary Computation*, 2007, pp. 4138–4145.

[25] S. R. D. Torres, D. C. B. Romero, L. F. N. Vasquez, and Y. J. P. Ardila, "A novel ab-initio genetic-based approach for protein folding prediction," in *9th Conf. Genetic and Evolutionary Computation*, 2007, pp. 393–400.

[26] L. Kapsokalivas, X. Gan, A. A. Albrecht, and K. Steinhöfel, "Population-based local search for protein folding simulation in the MJ energy model and cubic lattices," *Comp. Biol. Chem.*, vol. 33, no. 4, pp. 283–294, 2009.

[27] N. Mansour, F. Kanj, and H. Khachfe, "Particle swarm optimization approach for protein structure prediction in the 3D HP model," *Interdisciplinary Sciences, Comp. Life Sciences*, vol. 4, pp. 190–200, 2012.

[28] M. Cebrián, I. Dotú, P. Van Hentenryck, and P. Clote, "Protein structure prediction on the face centered cubic lattice by local search," *23rd Conference on Artificial Intelligence*, vol. 8, pp. 241–246, 2008.

[29] A. D. Ullah and K. Steinhöfel, "A hybrid approach to protein folding problem integrating constraint programming with local search," *BMC Bioinformatics*, vol. 11, no. 1, p. S39, 2010.

[30] S. Shatabda, M. Newton, and A. Sattar, "Mixed heuristic local search for protein structure prediction," in *Conf. Arti. Intel.*, 2013, pp. 876–882.

[31] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, 1985.

[32] M. Dorigo, V. Maniezzo, and A. Colorni, "Positive feedback as a search strategy," Tech. Rep., 1991.

[33] D. Do Duc, H. Q. Dinh, and H. H. Xuan, "On the pheromone update rules of ant colony optimization approaches for the job shop scheduling problem," in *PRIMA Conference*. Springer, 2008, pp. 153–160.