A Novel and Efficient Ant Colony Optimization Algorithm for Protein 3D Structure Prediction

Dong Do Duc
Advanced Institute, University of Engineering and Technology, Vietnam National University Hanoi, Vietnam

***Abstract*—Protein structure prediction (PSP) is considered as one of the most long-standing and challenging problem in bioinformatic. 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 hydrophobic-polar (HP) model and MiyazawaJernigan (MJ) model to calculate the free energy. The reinforcement learning information is expressed in the k-order Markov model, 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 by comparing with the 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], [2]. Analyzing the structure of proteins allow us to understand their features and produce medicines for diseases caused by protein misfolding [3], [4]. Unfortunately, it is very complex and difficult to simulate a protein nature into 3D structure [5]–[7]. Therefore, predicting the protein structure still 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 [8], nuclear magnetic resonance (NMR) [9]. However, these methods is time-consuming and costly, unsuitable for large-scale situations. For this reason, computational methods [10], [11] for predicting the structure of proteins are receiving great attentions. 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 [12], 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 [6]. The model for solving this problem has been optimized by the search space and the target function. In real-time, the search space is very large and determining the energies [OF WHAT] 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 to discretized the 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 [13]–[15], and among them, 3D face-centered cubic lattice (3D-FFC) possesses many advantages over other methods [16] [17] and have been used by many researchers [14] [18] [19] [20].There are two popular energy models for finding the approximately optimal structure of proteins. They are Hydrophobic-Polar (HP) energy model [21] [14] and MiyazawaJernigan (MJ) energy model [22] [23]. In HP-model, every amino acid is considered a bead labelled as hydrophobic (H) and polar (P), energy is determined from the physical interactions between H-nodes, P-nodes are seen as neutral. MJ-model takes interactions between specific pairs of amino acids, thus is closer to the realistic model of free energy. PSP has been classified as an NP-hard problem [24] [25], and so heuristic and metaheuristic algorithms have been proposed to solve it. Many of those are based on population, such as artificial learning system [26], generic algorithm [27] [28] [29], population-based algorithm [30], firefly algorithm [18], particle swarm optimization [31], ant colony optimization (ACO) [32]. Especially, Rashid has been proposed two methods based on genetic algorithm: GAplus [19](HP energy model) and MH-GA [20](graded energystrategically mixes 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 present K-ACOPSP algorithm to tackle 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 [19], TLS [33] MH-GA [20], Hybrid [34], Local Search [35]. The rest of this paper is organized as follow. In section 2, we introduce brief background knowledge about FCC lattice, HP-model, MJ-model and some related works. Section 3 is dedicated for new algorithm k-ACOPSP. The simulation study is shown in section 4. The conclusion is presented in the last section.

II. Problem statement and related works

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

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

A. FCC lattice and presentation of protein

FCC lattice is discretized from three-dimensional space, formed around triangles.Each node only has 12 neighbours with relative coordinates to the current node equal to (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. *Representation of protein on FFC lattice*: Given a

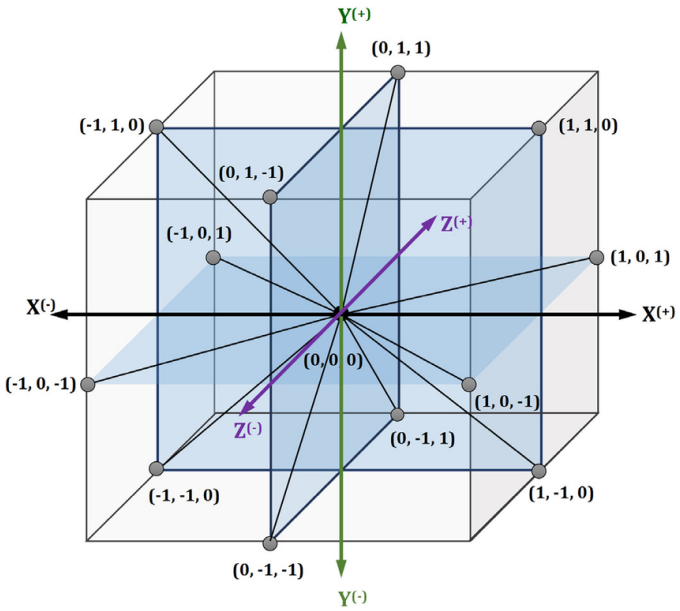

Fig. 1: The 12 basis vectors of the neighbors of the origin (0,0,0)

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 [16] [17] such as highest packing density, root mean square deviation values are smaller.

B. The energy models

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

1) HP energy model: The HP energy model proposed by Lau and Dill (1972) [21].In this model, the amino acids Gly, Ala, Pro, Val, Leu, Ile, Met, Phe, Tyr, Trp are labelled as hydrophobic (H), others are labelled as polar (P). Two consecutive H-labelled amino acids will create negative energy (-1). The complete HP energy of the model is calculated by equation (1):

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

where:

$$c_{ij} = \begin{cases} 1 \text{ if node i and j are not consecutive but are neighbor} \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

$$e_{ij} = \begin{cases} \text{-1: if amino acids i and j are both hydrophobic} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

2) MJ energy model: Relies on the interactive trend of amino acids, Miyazawa and Jernigan proposed the MJ enery model in 1985 [22]. The complete MJ energy is calculated by equation (4):

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

Where: c_{ij} is determined by euqation (2) and E_{ij} is taken from table I.

C. The optimal problem and related algorithms

Optimal problem: for each given protein with the native amino acid sequence of length m, PSP problem is transformed into finding the representation with optimal E_{HP} or E_{MJ} energy. Most recently, MH-GA [20] is proven to be the most efficient algorithm to solve PSP by comparing its experiment result with MJ-model against other state-of-the-art algorithms, such as Hybrid algorithm [34], and Local Search [35].

III. K-ACOPSP ALGORITHM

Ant colony optimization (ACO) is a stochastic metaheuristic method proposed by Dorigo [36] for traveling salesman problem (TSP). Till today, many variants have been developed to tackle difficult optimization problems. In this algorithm, 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 learnings information (pheromone) in a random fashion. When a solution is found, the algorithm appraises it then update the pheromone to improve the chance of finding better solutions on the next searches, this is repeated till the terminate requirement is met. The properties affect the quality of the algorithm are:

- A suitable structure graph.
- Heuristic information.
- How pheremone is stored and updated.

A. Construction graph

Without loss of generality, the first amino acid is put at the origin of the space (0,0,0) and start there. Neighbours 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 arcs directed from each vertex to all vertices in the next column. Illustrated in Fig. 2 With this, any feasible sequence of length m will correspond to a path on this graph.

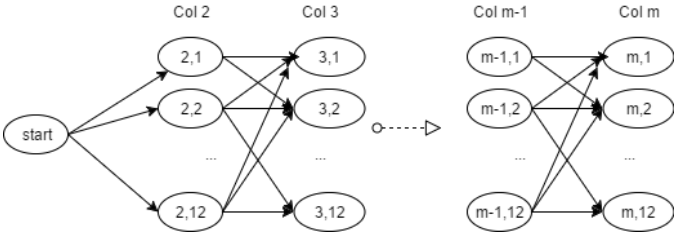

Fig. 2: Construction Graph

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 formula:

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

Where:

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

Note: To ensure self-avoiding walk constraint, we set $P_{i,j} = 0$ when selecting vertex would cause 2 amino acids to have the same coordinate on the protein representation.

C. Heuristic information

After the first $i-1$ amino acid were successfully represented and vector j is the selected direction to go next, then:

- Let η_{ij} be the heuristic value.
- Let E_{ij} be the amount of increased energy.
- Let $E_{max} = MAX(E_{ij})$

Then: $\eta_{ij} = E_{MAX} - E_{ij} + eps$, where eps is a small positive number to ensure η_{ij} is always positive. In our implements, we set it to 0.01.

D. Pheromone update

Instead of making choice only based on pheromone information in the current column, we can also take previously selected vertices into consideration too. Let $\tau_{i,j,v_{i-1},v_{i-2},\cdots,v_{i-k+1}}$ be the pheromone when vertices $(i,j),(i-1,v_(i-1)),\cdots,(i-k+1,v_{i-k+1})$ are selected. This way, the pheromone will give more accurate information during the searches.
Let $\tau_{i,j(k)} = \tau_{i,j,v_{i-1},v_{i-2},\cdots,v_{i-k+1}}$.
After every round of search, we update pheromone with SMMAX [37] formula:

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

where:

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

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:

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

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..

TABLE II: The result when HP energy model were 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

Algorithm 1 Procedure of Local Search

```
while stop conditions not satisfied do
    Caculate the HCC coordinates;
    Move ← SeclectMove();
    if Move = Null then
        Break;
    ApllyMove();
```

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 ∈ 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 study

A. Different values of K

MJ_{energy} is the average of energy values returned by our algorithm and Loops is the average of the number of loops that our algorithm will be convergent.

TABLE III: The result when trying multiple values of K

K	3NO3		3NO6		3ON7	
	MJ_{energy}	$Loops$	MJ_{energy}	$Loops$	MJ_{energy}	$Loops$
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

From the table III we see that the number of loops needed for convergence increases when K increases. Howevr, the value of MJ_{energy} increases significantly when K increases from 1 to 3. Values of MJ_{energy} 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 (taken from Peter Clote laboratory website[1]) and 3MSE, 3MR7, 3MQZ, 3NO6, 3NO3, 3ON7 from Critical Assessment of Protein

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

TABLE IV: The 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 YEYTLE-INGKSLKKYM
3NBM	108	SNASKELKVLVLCAGSGTSAQLANAINEGANLTEVRVIANSGAYGAHYDIMGVYDLIILAPQVRSYYREMKVDA ERLGIQIVATRGMEYIHLTKSPSKALQFVLEHYQ
3MQO	120	PAIDYKTAFHLAPIGLVLSRDRVIEDCNDELAAIFRCARADLIGRSFEVLYPSSDEFERIGERISPVMIAHGSY ADDRIMKRAGGELFWCHVTGRALDRTAPLAAGVWTFEDLSATRRVA
3MRO	142	SNALSASEERFQLAVSGASAGLWDWNPKTGAMYLSPHFKKIMGYEDHELPDEITGHRESIHPDDRARVLAALKA HLEHRDTYDVEYRVRTRSGDFRWIQSRGQALWNSAGEPYRMVGWIMDVTDRKRDEDALRVSREELRRL
3PNX	160	GMENKKMNLLLFSGDYDKALASLIIANAAREMEIEVTIFCAFWGLLLLRDPEKASQEDKSLYEQAFSSLTPREA EELPLSKMNLGGIGKKMLLEMMKEEKAPKLSDLLSGARKKEVKFYACQLSVEIMGFKKEELFPEVQIMDVKEYL KNALESDLQLFI
3MSE	180	GISPNVLNNMKSYMKHSNIRNIIINIMAHELSVINNHIKYINELFYKLDTNHNGSLSHREIYTVLASVGIKKWD INRILQALDINDRGNITYTEFMAGCYRWKNIESTFLKAAFNKIDKDEDGYISKSDIVSLVHDKVLDNNDIDNFF LSVHSIKKGIPREHIINKISFQEFKDYMLSTF
3MR7	189	SNAERRLCAILAADMAGYSRLMERNETDVLNRQKLYRRELIDPAIAQAGGQIVKTTGDGMLARFDTAQAALRCA LEIQQAMQQREEDTPRKERIQYRIGINIGDIVLEDGDIFGDAVNVAARLEAISEPGAICVSDIVHQITQDRVSE PFTDLGLQKVKNITRPIRVWQWVPDADRDQSHDPQPSHVQH
3MQZ	215	SNAMSVQTIERLQDYLLPEWVSIFDIADFSGRMLRIRGDIRPALLRLASRLAELLNESPGPRPWYPHVASHMRRR VNPPPETWLALGPEKRGYKSYAHSGVFIGGRGLSVRFILKDEAIEERKNLGRWMSRSGPAFEQWKKKVGDLRDFG PVHDDPMADPPKVEWDPRVFGERLGSLKSASLDIGFRVTFDTSLAGIVKTIRTFDLLYAEAEKGS
3NO3	238	GKDNTKVIAHRGYWKTEGSAQNSIRSLERASEIGAYGSEFDVHLTADNVLVVYHDNDIQGKHIQSCTYDELKDLQ LSNGEKLPTLEQYLKRAKKLKNIRLIFELKSHDTPERNRDAARLSVQMVKRMKLAKRTDYISFNMDACKEFIRLC PKSEVSYLNGELSPMELKELGFTGLDYHYKVLQSHPDWVKDCKVLGMTSNVWTVDDPKLMEEMIDMGVDFITTDL PEETQKILHSRAQ
3NO7	248	MGSDKIHHHHHHHENLYFQGMTFSKELREASRPIIDDIYNDGFIQDLLAGKLSNQAVRQYLRADASYLKEFTNIYA MLIPKMSSMEDVKFLVEQIEFMLEGEVEAHEVLADFINEPYEEIVKEKVWPPSGDHYIKHMYFNAFARENAAFTI AAMAPCPYVYAVIGKRAMEDPKLNKESVTSKWFQFYSTEMDELVDVFDQLMDRLTKHCSETEKKEIKENFLQSTI HERHFFNMAYINEKWEYGGNNNE
3ON7	280	GMKLETIDYRAADSAKRFVESLRETGFGVLSNHPIDKELVERIYTEWQAFFNSEAKNEFMFNRETHDGFFPASIS ETAKGHTVKDIKEYYHVYPWGRIPDSLRANILAYYEKANTLASELLEWIETYSPDEIKAKFSIPLPEMIANSHKT LLRILHYPPMTGDEEMGAIRAAAHEDINLITVLPTANEPGLQVKAKDGSWLDVPSDFGNIIINIGDMLQEASDGY FPSTSHRVINPEGTDKTKSRISLPLFLHPHPSVVLSERYTADSYLMERLRELGVL

Structure Prediction (CASP) competition[2]. These date were also used in [Rashid1].

To compare and evaluate the performance of k-ACO algorithm with the state-of-the-art approaches, we use the measure Relative Improvement (RI). Let denote E_A and E_B is the average energy values achieved by k-ACO algorithm and the state-of-the-art approaches.

$$RI = \frac{E_A - E_B}{E_B}$$

TABLE V: Comparison between GA and ACO when the running time of ACO has been increased

Protein details				GA plus			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

[2]http://predictioncenter.org

k-ACO was compared with 2 other algorithms TLS [33] and GA [19]. For each protein, each of 3 algorithms were run 50 times. The table below shows the best and the average result of 50 runs for each protein.

From the table II, it is straightforward to see that k-ACO finds the better result when compared to TLS. However, the results of k-ACO and GA are nearly the same, the difference between them always below 3%. K-ACO performed better than GA in 10 protein sequences while GA found better results than k-ACO in 7 protein sequences.

To further compare with GA, we increased the number of loops to 60000 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.

C. MJ energy model

In this section, data in table IV were used for MJ energy model. These data were also used in [20].

We run k-ACO on the dataset above and compare the result with other algorithms, namely Hybrid (Ullah,2010), Local search (Shatabda,2013), GA(Rashid,2016). This is the best and average result taken from 50 runs for each protein sequence.

From the column RI of Table VI, we see that for all proteins sequences, our algorithm improves the average energy ranging

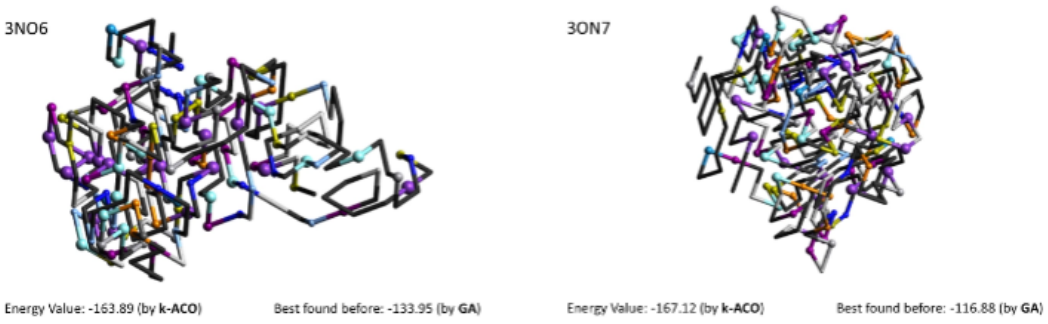

Fig. 3: New best Structure found by k-ACO for two largest datasets

TABLE VI: Comparing k-ACO algorithm against other proposed algorithms. The bold values are the best one in their row

Protein details			Hybrid		Local search		GA		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

from 2.3% to 65.86%.

TABLE VII: Running time of ACO and GA

Protein details			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	

V. Conclusion

In this paper, we presented the k-ACOPSP algorithm to predict the protein structure on FCC lattice, using two different energy models: HP model and MJ model. 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 ant colony optimization algorithms for other similar problems.

References

[1] A. Smith, "Protein misfolding," *Nature Reviews Drug Discovery*, vol. 426, no. 6968, pp. 78–102, 2003.

[2] C. M. Dobson, "Protein folding and misfolding," *Nature 426*, pp. 884–890, 2003.

[3] A. Breda and N. F. Valadares, "Protein structure, modelling and applications," *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*, 2007.

[4] P. Veerapandian, *Structure-based drug design*, 1997.
[5] B. Alberts, A. Johnson, M. Lewis, Julianand Raff, K. Roberts, and P. Walter, "The shape and structure of proteins," *Molecular Biology of the Cell, 4th edition*, 2002.
[6] C. Anfinsen, "The principles that govern the folding of protein chains," *Science 191(4069)*, pp. 223–230, 1973.
[7] "So much more to know," *The Science Editorial*, vol. 309, no. 5731, 2005.
[8] L. Bragg, *The Development of X-Ray Analysis, 1st Edition*, 1975.
[9] E. T. Baldwin, I. T. Weber, and R. S. Charles, "Crystal structure of interleuk in 8: symbiosis of nmr and crystallography," *Proc Natl Acad Sci USA*, pp. 502–506, 1991.
[10] C. A. Floudas, "Computational methods in protein structure prediction," *Biotechnology and Bioengineering*, vol. 97, pp. 207–213, 2007.
[11] C. M. Dobson, "Computational biology: protein predictions," pp. 176–177, 2007.
[12] H. Berman, "The protein data bank," *Nucleic Acids Res*, pp. 235–242, 2000.
[13] A. Bechini, "On the characterization and software implementation of general protein lattice models," *PLoS ONE*, 2013.
[14] I. Dotu, M. Cebrian, P. V. Hentenryck, and P. Clote, "On lattice protein structure prediction revisited," *IEEE/ACM Trans Comput Biol Bioinform*, 2011.
[15] M. Mann and R. Backofen, "Exact methods for lattice protein models, bio-algorithms and med-systems," vol. 10, pp. 213–225, 2014.
[16] D. Covell and R. Jernigan, "Conformations of folded proteins in restricted spaces," *Biochemistry*, pp. 3287–94, 1990.
[17] T. C. Hales, "A proof of the kepler conjecture," *The Annals of Mathematics*, vol. 162, no. 3, pp. 1065–1185, 2005.
[18] 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.
[19] 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.
[20] 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," *Computational Biology and Chemistry*, pp. 162–177, 2016.
[21] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, vol. 22, pp. 3986–3997, 1989.
[22] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules 18(3)*, pp. 534–552, 1985.
[23] S. Miyazawa and R. Jernigan, "Residueresidue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," *J Mol Biol*, pp. 623–644, 1996.
[24] 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*, pp. 1183–1198, 1993.
[25] M. Paterson and T. Przytycka, "On the complexity of string folding," *Discrete Applied Mathematics*, vol. 71, pp. 217–230, 1996.
[26] 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, pp. 101–117, 2007.
[27] M. T. Hoque, M. Chetty, and A. Sattar, "Protein folding prediction in 3d fcc hp lattice model using genetic algorithm," *IEEE Congress on Evolutionary Computation*, pp. 4138–4145, 2007.
[28] 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," *A novel ab-initio genetic-based approach for protein folding prediction*, pp. 393–400, 2007.
[29] "A genetic algorithm for 3d protein folding simulations. in: The 5th international conference on genetic algorithms."
[30] L. Kapsokalivas, X. Gan, A. Albrecht, and K. Steinhfel, "Population-based local search for protein folding simulation in the mj energy model and cubic lattices," *Comput Biol Chem*, pp. 283–294, 2009.
[31] N. Mansour, "Particle swarm optimization approach for protein structure prediction in the 3d hp model," *Interdiscip Sci 4*, pp. 190–200, 2013.
[32] A. Shmygelska and H. H. Hoos, "An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem," *BMC Bioinformatics*, 2005.
[33] P. Clote, M. Cebrian, I. Dotu, and P. V. Hentenryck, "Protein structure prediction on the face centered cubic lattice by local search," *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 241–246, 2008.
[34] A. D. Ullah and K. Steinhfel, "A hybrid approach to protein folding problem integrating constraint programming with local search," *Selected articles from the Eighth Asia-Pacific Bioinformatics Conference*, vol. 11, 2010.
[35] S. Shatabda, M. A. H. Newton, and A. Sattar, "Mixed heuristic local search for protein structure prediction," *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.
[36] M. Dorigo and C. A. Maniezzo, Vittorio and, "Positive feedback as a search strategy," Tech. Rep., 1991.
[37] D. D. Dong, H. X. Huan, and D. Q. Huy, "On the pheromone update rules of ant colony optimization approaches for the job shopscheduling problem," pp. 153–160, 2008.

Dang Thanh Hai: The problem of 3D protein structure prediction is an important and challenge task in Bioinformtics. This paper presents an efficient ant colony optimization based algorithm for predicting the protein structure on a three-dimensional face-centered cubic lattice coordinate. The algorithm use the hydrophobic-polar (HP) model and Miyazawa-Jernigan (MJ) model to calculate the free energy as the score of the objective function. The approach descirbed in the paper is highly potential for a publication in the follow-up.