Protein Structure Prediction Based on Profile HMM and DMQPSO

Haixia Long, Shulei Wu, and Chun Shi

***Abstract*—Protein structure prediction is a challenging field strongly associated with protein function and evolution determination, which is crucial for biologists and the pharmaceutical industry. Despite significant process made in recent years, protein structure prediction maintains its status as one of the prime unsolved problems in computational biology. In this study, we have developed a method for protein structure prediction based on profile Hidden Markov Model (HMM) and Quantum Particle Swarm Optimization (QPSO) with diversity-maintained algorithm (DMQPSO). The profile HMM can reduce the number of states using secondary structure information about proteins for each fold, which is called a 7-state HMM. The DMQPSO is an efficient optimization algorithm which is used to train profile HMM. Experiment results show that the proposed method is reasonable and the accuracy of protein secondary structure prediction is increased.**

***Index Terms*—Protein structure prediction, protein secondary structure, fold recognition, profile HMM, DMQPSO.**

I. Introduction

The research of protein secondary structure plays a major role in determining sub-cellular locations and improving the sensitivity of fold recognition method [1]. Many models and methods have been applied to predict secondary structure. For example, Jpred [2], PSI-PRED [3], and PHD-PSI [4] are methods based on neural networks. Context- based secondary structure potential approach(CSSP) [5], and some other statistical classification methods, such as K-nearest neighbor method [6], support vector machines (SVM) [7], and hidden Markov model [8] have been well investigated. Lampros proposed a reduced state-space HMM, which simultaneously finds amino acid sequences and secondary structures for proteins [9]. However, the main disadvantage of HMMs is the employment of large model architectures, which require large data sets and high computational resource for training. It is necessary to reduce the parameters of HMMs, while HMMs maintain performance of fold recognition. Also, HMMs must consider sequential information about secondary structures of proteins, to improve prediction performance.

Therefore, we propose a novel fold recognition method for protein tertiary structure prediction based on a hidden Markov model. The method consists of seven hidden states and uses protein sequences and secondary structure information to recognize the fold of proteins. Our contribution can be summarized as follows: first, we present the 7-state HMM to protein fold recognition. The seven states of the model are divided into three components {H, E, and C}: -helix, -strand, and coil. H and E states of three components represents: "beginning", "middle" and "ending" for the secondary structure; H beginning(HB), H middle (H), H ending (HE), E beginning(EB), E middle (E) , E ending (EE). Second, we avoid utilizing the computationally expensive Baum-Welch algorithm [10], but we use the QPSO algorithm [11] to optimize the parameters of HMM. Finally, we evaluate the proposed HMM for a low homology dataset.

A new variant of PSO [12], called quantum-behaved particle swarm optimization (QPSO), has been proposed in order to improve the global search ability of the original PSO. The iterative equation of QPSO is far different from that of PSO in that it is needs no velocity vectors for particles, has fewer parameters to adjust and can be implemented more easily. It has been proved that this iterative equation leads QPSO to be global convergent [13]. The QPSO algorithm has been aroused the interests of many researchers from different communities. It has been shown to successfully solve a wide range of continuous optimization problems. Among these applications, it has been used to tackle constraint optimization problems [14], multi-objective optimization problems [15], neural network training [16], economic dispatch problems [17], electromagnetic design [18], semiconductor design [19], bioinformatics [20]. In this paper, we make analyses for a single particle's behavior in QPSO, deriving the sufficient and necessary condition for probabilistic roundedness of the particle that can guarantee the particle swarm to converge. Then based on the analyses, we propose an improved QPSO, called diversity-maintained QPSO (DMQPSO), in which the diversity is maintained at a certain level to enhance the global search ability of QPSO. Finally, the DMQPSO algorithm is used to train the HMMs for protein structure prediction and tested on two protein data base.

II. Methods

A Topology of the 7-State HMM

HMMs are widely used in bioinformatics. An HMM describes a probability distribution over a potentially infinite number of sequences. The HMM is a doubly embedded stochastic process with an underlying stochastic process that is not observable (it is hidden), and can only be observed via

Manuscript received August 9, 2014; revised November 20, 2014. This work was supported by the National Natural Science Fund (No. 61163042, No. 61362016), the Hainan Province Natural Science Fund (No. 614235), the Higher School Scientific Research Project of Hainan Province (Hjkj2013-22).

The authors are with School of Information Science Technology, Hainan Normal University, Haikou 571158, HaiNan, China (e-mail: haixia_long@163.com, 595615374 @qq.com, 605515770 @qq.com).

another set of stochastic processes that produces the sequence of observations.

Fig.1 describes a simple HMM model. The HMM consists of a set of N states $S = \{S_1, S_2, \cdots, S_N\}$, and q_T is a state of T moment, where $q_T \in S$. O_T is an observation signal at T moment. States are connected to each other by transition probability matrix A. Emission probability matrix B emits an observable a symbol from an output alphabet with a probability. The joint probability of the symbol sequence and the state sequence is a product of all transition and emission probabilities.

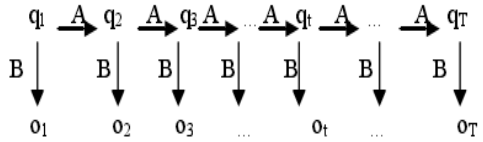


Fig. 1. A simple HMM model.

In the following, we introduce 7-state Hidden Markov model shown in Fig. 2 and describe the protein fold recognition method using the model.

The Hidden Markov model is a statistical model, which is very well suited for many tasks in molecular biology [9]. We utilize the HMM to recognize the folds of unknown protein structures. To improve the performance of the HMM, we apply information about secondary structure obtained from a database of known protein structures to the HMM. Also, we propose a novel HMM with seven hidden states, called the 7-state HMM. The model uses secondary structure information about proteins and has reduced parameters for the HMM. The seven states of the model are divided into three components {H, E, and C}: -helix, -strand, and coil. H and E states of three components represents: "beginning", "middle" and "ending" for the secondary structure; H beginning(HB), H middle (H), H ending (HE), E beginning(EB), E middle (E) , E ending(EE). The secondary structure states are obtained from DSSP [21]. The DSSP 8-state secondary structure representation (H, G, E, B, I, S, T, -) was grouped according to the 3-state scheme as follows: H, G, and I are regarded as helix (H), E and B as strand (E) and all others as coils.

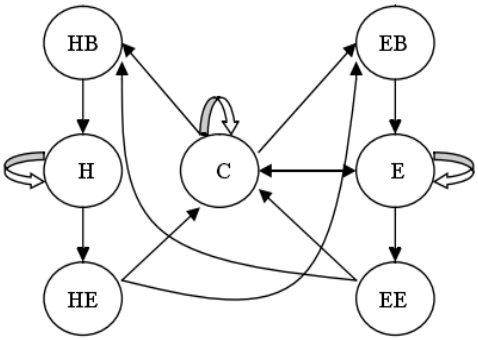


Fig. 2. The 7-states HMM.

In the 7-state HMM, every state emits a signal, one of the 20 amino acids, according to a set of emission probabilities, then stochastically transmits the signal to some other states with a probability according to the previous state. Since the 7-state HMM has seven states and emits 20 amino acids for each state, the number of emission probabilities is 7 × 20, the number of transition probabilities is 7×7, and the number of beginning probabilities is 1×7. The total number of the model parameters is 196.

The goal in HMM learning is to determine model parameters—the transition probability and the emission probability—from an ensemble of training samples. There is no known method for obtaining the optimal or most likely set of parameters from the data, but we can nearly always determine a good solution in a straight forward method. The forward–backward algorithm is an instance of a generalized expectation maximization algorithm. The general approach will iteratively update the weights in order to explain observed training sequences better. We utilize newly optimization algorithm to train the HMM, that is DMQPSO algorithm.

B DMQPSO Algorithm

In the PSO with M individuals, each individual is treated as volume-less particle in the D-dimensional space, with the current position vector and velocity vector of particle i at the nth iteration represented as $X_{i,n} = \left(X_{i,n}^1, X_{i,n}^2, \cdots, X_{i,n}^D\right)$ and $V_{i,n} = \left(V_{i,n}^1, V_{i,n}^2, \cdots, V_{i,n}^D\right)$. The particle moves according to the following equations:

$$V_{i,n+1}^j = w \cdot V_{i,n}^j + c_1 r_{i,n}^j \left(X_{i,n}^j - P_{i,n}^j\right) + c_2 R_{i,n}^j \left(X_{i,n}^j - G_n^j\right) \quad (1)$$

$$X_{i,n+1}^j = X_{i,n}^j + V_{i,n+1}^j \quad (2)$$

for $i = 1, 2, \cdots, M$; $j = 1, 2, \cdots, D$, where c_1 and c_2 are called acceleration coefficients. The parameter is known as the inertia weight which can be adjusted to balance the explorative search and the exploitive search of the particle. Vector $P_{i,n} = \left(P_{i,n}^1, P_{i,n}^2, \cdots, P_{i,n}^D\right)$ is the best previous position (the position giving the best objective function value or fitness value) of particle and called personal best $(pbest)$ position, and vector $G_n = \left(G_n^1, G_n^2, \cdots, G_n^D\right)$ is the position of the best particle among all the particles in the population and called global best $(gbest)$ position. Without loss of generality, we consider the following maximization problem:

$$Maximize\ f(x), \quad s.t.\ X \in S \subseteq R^D \quad (3)$$

where $f(x)$ is an objective function continuous almost everywhere and is the feasible space. Accordingly, $P_{i,n}$ can be updated by

$$P_{i,n} = \begin{cases} X_{i,n} & if\ f(X_{i,n}) \succ f(P_{i,n-1}) \\ P_{i,n-1} & if\ f(X_{i,n}) \leq f(P_{i,n-1}) \end{cases} \quad (4)$$

and G_n can be found by $G_n = P_{g,n}$, where $g = \arg\max_{1 \leq i \leq M} \left[f\left(P_{i,n}\right)\right]$. The parameters $r_{i,n}^j$ and

$R_{i,n}^j$ are sequences of two different sequences of random numbers distributed uniformly within(0, 1), which is denoted by $r_{i,n}^j, R_{i,n}^j \sim U(0,1)$. Generally, the value of $V_{i,n}^j$ is restricted in the interval $[-V_{\max}, V_{\max}]$.

Trajectory analysis in [12] showed that convergence of the PSO algorithm may be achieved if each particle converges to its local attractor, $p_{i,n} = \left(P_{i,n}^1, P_{i,n}^2, \cdots, P_{i,n}^D\right)$ defined at the coordinates

$$p_{i,n}^j = \left. c_1 r_{i,n}^j P_{i,n}^j + c_2 R_{i,n}^j G_n^j \right/ c_1 r_{i,n}^j + c_2 R_{i,n}^j \quad 1 \le j \le D \tag{5}$$

$$\text{Or} \quad p_{i,n}^j = \varphi_{i,n}^j \cdot P_{i,n}^j + \left(1 - \varphi_{i,n}^j\right) \cdot G_n^j \tag{6}$$

where $\varphi_{i,n}^j = c_1 r_{i,n}^j \, c_1 r_{i,n}^j / \left(c_1 r_{i,n}^j + c_2 R_{i,n}^j\right)$ with regard to the random numbers $r_{i,n}^j$ and $R_{i,n}^j$ in Eq.(2) and (4). In PSO, the acceleration coefficients c_1 and c_2 are generally set to be equal, i.e. $c_1 = c_2$, and thus $\varphi_{i,n}^j$ is a sequence of uniformly distributed random numbers within (0,1). As a result, Eq. (6) can be restated as

$$p_{i,n}^j = \varphi_{i,n}^j \cdot P_{i,n}^j + \left(1 - \varphi_{i,n}^j\right) \cdot G_n^j \quad \phi_{i,n}^j \sim U(0,1) \tag{7}$$

In QPSO, each single particle is treated as a spin-less one moving in quantum space. Thus state of the particle is characterized by wave function, where $|\psi|^2$ is the probability density function of its position. Inspired by convergence analysis of the particle in PSO, we assume that, at the nth iteration, particle flies in the D-dimensional space with a potential well centered at $p_{i,n}^j$ on the j^{th} dimension $\left(1 \le j \le D\right)$. Let $Y_{i,n+1}^j = \left| X_{i,n+1}^j - p_{i,n}^j \right|$, we can obtain the normalized wave function at iteration $n+1$

$$\Psi\left(Y_{i,n+1}^j\right) = \frac{1}{\sqrt{L_{i,n}^j}} \exp\left(-Y_{i,n+1}^j / L_{i,n}^j\right) \tag{8}$$

which satisfies the bound condition that $\Psi\left(Y_{i,n+1}^j\right) \to 0$ as $Y_{i,n+1}^j \to \infty$ is the characteristic length of the wave function. By the statistical interpretation of wave function, the probability density function is given by

$$Q\left(Y_{i,n+1}^j\right) = \left|\psi\left(Y_{i,n}^j\right)\right|^2 = \frac{1}{L_{i,n}^j} \exp\left(-\frac{2Y_{i,n+1}^j}{L_{i,n}^j}\right) \tag{9}$$

and thus the probability distribution function is

$$F\left(Y_{i,n+1}^j\right) = 1 - \exp\left(\left. -2Y_{i,n+1}^j \right/ L_{i,n}^j\right) \tag{10}$$

Using Monte Carlo method, we can measure the j^{th} component of position of particle at the $(n+1)th$ iteration by

$$X_{i,n+1}^j = p_{i,n}^j \pm \frac{L_{i,n}^j}{2} \ln\left(\left. 1 \right/ \mu_{i,n+1}^j\right) \quad \mu_{i,n+1}^j \sim U(0,1) \tag{11}$$

where $\mu_{i,n+1}^j$ is a sequence of random numbers uniformly distributed within $(0,1)$. The value of $L_{i,n}^j$ is determined by:

$$L_{i,n}^j = 2\alpha \cdot \left| X_{i,n}^j - C_n^j \right| \tag{12}$$

where $C_n = \left(C_n^1, C_n^2, \cdots, C_n^D\right)$ is called mean best (*mbest*) position defined by the average of the pbest position of all particles, i.e. $C_n^j = \left(\left. 1 \right/ M\right) \sum_{i=1}^{M} P_{i,n}^j \left(1 \le j \le D\right)$. Thus the position of the particle updates according to the following equation:

$$X_{i,n+1}^j = p_{i,n}^j \pm \alpha \cdot \left| X_{i,n}^j - C_n^j \right| \cdot \ln\left(\left. 1 \right/ \mu_{i,n+1}^j\right) \tag{13}$$

The parameter α in Eq. (12) and (13) is called contraction-expansion (CE) coefficient, which can be adjusted to balance the local search and the global search of the algorithm during the optimization process. The current position of the particle in QPSO is thus updated according to Eq. (7) and (13).

The QPSO algorithm starts with the initialization of the particle's current positions and their pbest positions (setting $P_{i,0} = X_{i,0}$), followed by the iteration of updating the particle swarm. At the each iteration, the *mbest* position of the particle swarm is computed and the current position of each particle is updated according to Eq. (7) and (13). Before each particle updates its current position, its fitness value is evaluated and then its *pbest* position and the current *gbest* position are updated. In Eq. (13), the probability of using either operation "+" or operation "-" is equal to 0.5. The search procedure continues until the termination condition is met.

We outline the procedure of the QPSO algorithm as follows:

Procedure of the QPSO:

Step 1: Initialize the population;

Step 2: Execute the following steps;

Step 3: Compute mean best position C;

Step 4: Properly select the value of α ;

Step 5: For each particle in the population, execute form Step 6 to Step 8;

Step 6: Evaluate the objective function value;

Step 7: Update *pbest* position and the *gbest* position;

Step 8: Update each component the particle's position according to Eq. (7) and (13);

Step 9: While the termination condition is not met, return to Step 2;

Step 10: Output the results.

QPSO is a promising optimization problem solver that outperforms PSO in many real application areas. First of all, the introduced exponential distribution of positions makes

QPSO global convergent. The QPSO algorithm in the initial stage of search, as the particle swarm initialization, its diversity is relatively high. In the subsequent search process, due to the gradual convergence of the particle, the diversity of the population continues to decline. As the result, the ability of local search ability is continuously enhanced, and the global convergence ability is continuously weakened. In early and middle search, reducing the diversity of particle swarm optimization for contraction efficiency improvement is necessary, however, to late stage of search, because the particles are gathered in a relatively small range, particles swarm diversity is very low, the global search ability becomes very weak, the ability for a large range of search has been very small, this algorithm will occur the phenomenon of premature.

To overcome this shortcoming, we introduce diversity-maintained into QPSO.

The population diversity of the DMQPSO is denoted as *diversity*(*pbest*) and is measured by average Euclidean distance from the particle's position to the mean position, namely

$$diversity(pbest) = \frac{1}{M \cdot |A|} \cdot \sum_{i=1}^{M} \sqrt{\sum_{j=1}^{D} \left(x_{i,j} - \bar{x}_j \right)^2} \quad (10)$$

where *M* is the population of the particle, $|A|$ is the length of longest the diagonal in the search pace, and *D* is the dimension of the problem. Hence, we may guide the search of the particles with the diversity measures when the algorithm is running.

In the DMQPSO algorithm, only low bound d_{low} is set for *diversity*(*pbest*) to prevent the diversity from constantly decreasing. The procedure of the algorithm is as follows. After initialization, the algorithm is running in convergence mode. In process of convergence, the convergence mode is realized by Contraction-Expansion (CE) Coefficient. On the course of evolution, if the diversity measure *diversity*(*pbest*) of the swarm drops to below the low bound d_{low}, the particle swarm turns to be in explosion mode in which the particles are controlled to explode to increase the diversity until it is larger than d_{low}.

C Protein Structure Prediction Based on Profile HMM and DMQPSO

The procedure of protein structure prediction is shown in Fig. 3. A set of data consists of the training data and test data. When using the DMQPSO algorithm to train the parameters of the HMM, namely, the transition and emission probabilities. The number of emission probabilities is 7×20, the number of transition probabilities is 7×7, and the number of beginning probabilities is 1×7. The total number of the model parameters is 196. During the each iteration of QPSO, a copy of the population is created. All particles in this copy are normalized such that the constraints on the transition and emission probabilities are satisfied. Each particle in the copy population is evaluated by the following equation:

$$Q = \frac{N_{correct}}{N_{total}} \times 100\% \quad (14)$$

where $N_{correct}$ is the number of protein sequence correctly predicted and N_{total} is the total number of protein sequences.

After training of the HMM with DMQPSO, the output *gbest* position of the particles represents the optimized parameters of the HMM, i.e. transition and emission probabilities. The trained HMM can be considered as a profile for the set of test data. Finally, the resulting prediction is evaluated according to the Eq. (14).

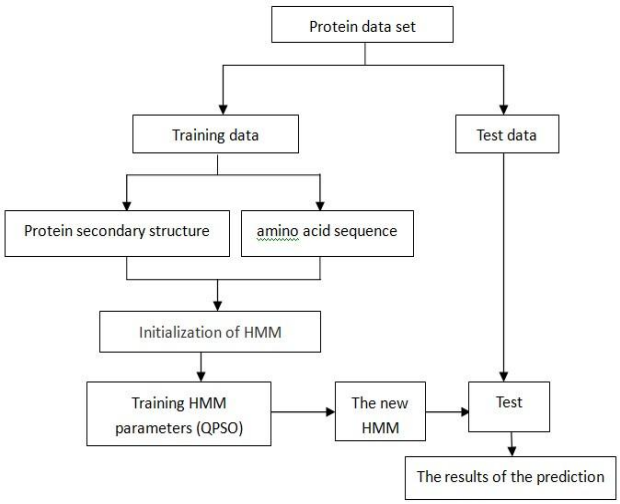


Fig. 3. The procedure of protein structure prediction.

III. Simulation and Results

A Data Set and Parameter Setting

In order to validate the proposed improvements, all proteins have been selected from the Protein Data Bank (PDB) [22] which is presented in Table Ⅰ and known class labels and fold labels in SCOP [23] which is presented in Table II. Then the group of sequences is separated in training and test sets. The total numbers of folds are 20 and 8 respectively.

The proposed QPSO was used to train the HMMs on the datasets. The population with the size of 20 was initialized randomly with uniform distribution in the search scope [0, 1]. The parameters of DMQPSO were determined according to published recommendations [24].

B Results and Discussion

We tested the performance of DMQPSO with 7-state HMM for predicting protein structure by compared it with 9-state HMM, 3-state HMM and SAM [25].

Table III presents the accuracy of each fold of PDB data set and the total accuracy for the 7-state HMM with DMQPSO compared with 9-state HMM, 3-state HMM and SAM. In the case of fold a3, the total number of members is 10 and the number of protein sequences correctly predicted is 7. Therefore, using Eq. (14), the prediction accuracy for the fold a3 is 70%. The prediction accuracies for all folds and the overall accuracy are computed in an identical manner. The fold prediction performance of the 7-state HMM with DMQPSO is assessed by comparing it with the 9-state HMM, 3-state HMM and SAM model trained with identical data sets. The overall accuracy for 7-state HMM with DMQPSO, 9-state HMM, 3-state HMM and SAM is 32.2%, 27.2%,

22.2%, 28.1%, respectively. The comparison shows that 7-state HMM with DMQPSO outperforms the other models for most types of folds and the overall accuracy is better than the other models.

TABLE I: PDB DATA SET

Fold	Index	Number of sequences in the training set	Number of sequences in the test set
Globin-like	a1	23	11
Cytochrome c	a3	20	10
DNA-binding 3-helical bundle	a24	32	15
EF-hand	a39	32	15
SAM domain-like	a118	35	16
All alpha proteins		142	67
Immunoglobin-like beta sandwich	b1	142	71
Galactose-binding domain-like	b18	23	10
ConA-like lectins/glucanases	b29	25	13
PDZ domain-like	b36	25	13
Double-stranded beta-helix	b82	32	16
All beta proteins		247	123
(TIM)-barrel	c1	151	75
P-loop containing nucleotide	c37	103	51
Ribonuclease H-like motif	c55	33	17
PLP-dependent transferases	c67	31	16
Periplasmic binding protein-like Ⅱ	c94	25	13
Alpha and beta proteins(a/b)		343	172
b-grasp	d15	55	28
Cystatin-like	d17	21	10
Ferredoxin-like	d58	117	58
Protein kinase-like (PK-like)	d144	25	12
C-type lectin-like	d169	20	10
Alpha andbetaproteins (a+b)		238	118
Overall		970	480

TABLE II: SCOP DATA SET

Fold	Index	Number of sequences in the training set	Number of sequences in the test set
Putative DNA-binding domain	a6	7	4
CH domain-like	a40	7	3
All alpha proteins		14	7
Prealbumin-like	b3	9	4
PUA domain-like	b122	9	5
All beta proteins		18	9
Barrel-sandwich hybrid	c30	8	4
Amino acid dehydrogenase-like, N-terminal domain	c58	9	4
Alpha and beta proteins(a/b)		17	8
Lysozyme-like	d2	7	4
IL8-like	d9	7	4
Alpha andbetaproteins (a+b)		14	8
Overall		63	32

Table IV presents the accuracy of each fold and the total accuracy for the SCOP data set. The overall accuracy for 7-state HMM with DMQPSO, 9-state HMM, 3-state HMM and SAM is 81.2%, 78.1%, 71.8%, 68.7%, respectively. The comparison shows that 7-state HMM with DMQPSO is substantially better than the accuracy achieved with other models.

TABLE III: COMPARISON OF THE PROPOSED MODEL WITH 9-STATE HMM, 3-STATE HMM AND SAM FOR THE 20 FOLDS OF THE FIRST DATASET

Fold index	7-state HMM with DMQPSO		9-state HMM		3-state HMM		SAM	
a 1	8 / 1 1	72.7%	5 / 1 1	45.5%	5 / 1 1	45.5%	9 / 1 1	81.8%
a 3	8 / 1 0	80%	5 / 1 0	50%	3 / 1 0	30%	6 / 1 0	60%
a 24	6 / 1 5	40%	2 / 1 5	13.3%	1 / 1 5	6.7%	2 / 1 5	13.3%
a 39	10/15	66.7%	8 / 1 5	53.3%	6 / 1 5	40%	11/15	73.3%
a118	10/16	62.5%	9 / 1 6	56.3%	6 / 1 6	37.5%	0 / 1 6	0 %
b 1	38/71	53.5%	31/71	43.9%	29/71	40.9%	23/71	33.3%
b 18	3 / 1 0	30%	2 / 1 0	20%	2 / 1 0	20%	3 / 1 0	30%
b 29	5 / 1 3	38.5%	3 / 1 3	23%	2 / 1 3	15.3%	3 / 1 3	23%
b 36	11/13	84.6%	11/13	84.6%	10/13	76.9%	11/13	84.6%
b 82	5 / 1 6	31.3%	2 / 1 6	12.5%	1 / 1 6	6.2%	1 / 1 6	6.2%
c 1	12/75	16%	6 / 7 5	8 %	5 / 7 5	6.6%	7 / 7 5	9.3%
c 37	10/51	19.6%	3 / 5 1	5.8%	2 / 5 1	3.9%	25/51	49%
c 55	9 / 1 7	52.9%	5 / 1 7	29.4%	4 / 1 7	23.5%	3 / 1 7	17.6
c 67	12/16	75%	10/16	62.5%	8 / 1 6	50%	11/16	68.7%
c 94	9 / 1 3	69.2%	9 / 1 3	69.2%	7 / 1 3	53.8%	5 / 1 3	38.4%
d 15	8 / 2 8	28.6%	5 / 2 8	17.8%	4 / 2 8	14.2%	1 / 2 8	3.5%
d 17	3 / 1 0	30%	2 / 1 0	20%	1 / 1 0	10%	0 / 1 0	0 %
d 58	4 / 5 8	6.9%	2 / 5 8	3.4%	2 / 5 8	3.4%	3 / 5 8	5.1%
d144	9 / 1 2	75%	6 / 1 2	50%	5 / 1 2	41.6%	9 / 1 2	75%
d169	6 / 1 0	60%	5 / 1 0	50%	4 / 1 0	40%	2 / 1 0	20%
overall	186/480	38.8%	131/480	27.2%	107/480	22.2%	135/480	28.1%

TABLE IV: COMPARISON OF THE PROPOSED MODEL WITH 9-STATE HMM, 3-STATE HMM AND SAM FOR THE 10 FOLDS OF THE SECOND DATASET

Fold index	7-state HMM with DMQPSO		9-state HMM		3-state HMM		SAM	
a6	3/4	75%	2/4	50%	1/4	25%	2/4	50%
a40	3/3	100%	3/3	100%	3/3	100%	2/3	66.6%
b3	4/4	100%	4/4	100%	4/4	100%	2/4	50%
b122	5/5	100%	4/5	80%	4/5	80%	5/5	100%
c30	3/4	75%	3/4	75%	2/4	50%	4/4	100%
c58	4/4	100%	3/4	75%	3/4	75%	2/4	50%
d2	3/4	75%	2/4	50%	2/4	50%	1/4	25%
d9	4/4	100%	4/4	100%	4/4	100%	4/4	100%
overall	29/32	90.6%	25/32	78.1%	23/32	71.8%	22/32	68.7%

IV. DISCUSSION

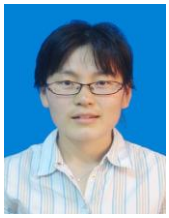
In this paper, we used DMQPSO algorithm to optimize 7-state HMM parameters, and then the optimized model is used to predict protein structure. Each state of the model corresponds to the secondary structure of helix (H), strand (E) and coil. To test the effect of the model, in the experiment we used the model with the PDB data set and SCOP date set and compared with the other model. The results of the 7-state HMM with DMQPSO outperforms than other models.

ACKNOWLEDGMENT

This work was financially supported by the National Natural Science Fund (No. 61163042, No. 61362016), the Hainan Province Natural Science Fund (No. 614235), the Higher School Scientific Research Project of Hainan Province (Hjkj2013-22).

REFERENCES

[1] S. Montgomerie, S. Sundararaj, W. J. Gallin *et al*., "Improving the accuracy of protein secondary structure prediction using structural alignment," *BMC Bioinform*, vol. 7, pp. 301-313, 2006.
[2] C. Cole, J. D. Barber, and G. J. Barton, "The Jpred3 secondary structure prediction server," *Nucleic Acids Res*., vol. 36, pp. 197–201, 2008
[3] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, pp. 404–405, 2000.
[4] D. Przybylski and B. Rost, "Alignments grow, secondary structure prediction improves," *Proteins: Struct. Funct. Bioinform*, vol. 46, pp. 197–205, 2002.
[5] Y. Li, H. Liu, I. Rata *et al*, "Building a knowledge-based statistical potential by capturing high-order inter-residue interactions and its applications in protein secondary structure assessment," *J. Chem. Inform. Model*, vol. 53, pp. 500–508, 2013.
[6] K. Joo, I. Kim, S. Y. Kim *et al*., "Prediction of the secondary structure of proteins using Predict, a nearest neighbor method on pattern space," *J. Kor. Phys. Soc*., vol. 45, pp. 1441–1449, 2004.
[7] X. B. Zhou, C. Chen, Z. C. Li *et al*., "Improved prediction of sub-cellular location for apoptosis proteins by the dual-layer support vector machine," *Amino Acids*, vol. 35, 383–388, 2008.
[8] S. A. Malekpour, S. Naghizadeh, and H. Pezeshk, "Protein secondary structure prediction using three neural networks and a segmental semi Markov model," *Math. Biosci*., vol. 217, pp. 145–150, 2009.
[9] C. Lampros, C. Papalukas, T. P. Exarchos *et al*., "Sequence-based protein prediction using a reduced state-space hidden Markov model," *Comput. Biol. Med.*, vol. 37, pp. 1211-1224, 2006
[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, pp. 257–285, 1989.
[11] J. Sun, B. Feng, and W. B. Xu, "Particle swarm optimization with particles having quantum behavior," in *Proc. Congress on Evolutionary Computation*, Portland, USA, pp. 326, 2004.
[12] M. Clerc and J. Kennedy, "The particle swarm-explosion, stability and convergence in a multidimensional complex space," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 58–73, 2002.
[13] W. Fang, J. Sun, Z. P. Xie *et al*., "Convergence analysis of quantum-behaved particle swarm optimization algorithm and study on its control parameter," *Acta Physica Sinica*, vol. 59, pp. 3686-3694, 2010.
[14] J. Sun, J. Liu, and W. B. Xu, "Using quantum-behaved particle swarm optimization algorithm to solve non-linear programming problems," *International Journal of Computer Mathematics*, vol. 84, pp. 261–272, 2007.
[15] S. N. Omkara, R. Khandelwala, T. V. S. Ananthb *et al.*, "Quantum behaved particle swarm optimization (QPSO) for multi-objective design optimization of composite structures," *Expert Systems with Applications*, vol. 36, pp. 11312–11322, 2009.
[16] S. Y. Li, R. G. Wang, W. W. Hu *et al*., "A new QPSO based BP neural network for face detection, advances in soft computing," *Fuzzy Information and Engineering*, vol. 40, pp. 355-363, 2007.
[17] J. Sun, W. Fang, D. Wang, *et al*., "Solving the economic dispatch problem with a modified quantum-behaved particle swarm optimization method," *Energy Conversion and Management*, vol. 50, pp. 2967–2975, 2009.
[18] L. S. Coelho and P. Alotto, "Global optimization of electromagnetic devices using an exponential quantum-behaved particle swarm optimizer," *IEEE Transactions on Magenetics*, vol. 44, pp. 1074–1077, 2008.
[19] S. L. Sabata, L. S. Coelho and A. Abrahamc, "MESFET DC model parameter extraction using quantum particle swarm optimization," *Microelectronics Reliability*, pp. 49, vol. 660–666, 2009.
[20] Y. J. Cai, J. Sun, J. Wang *et al*., "Optimizing the codon usage of synthetic gene with QPSO algorithm," *Journal of Theoretical Biology*, vol. 254, pp. 123–127, 2008.
[21] W. Kabsch and C. Sander, "Directory of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.
[22] H. M.Berman, J. Westbrook, Z. Feng *et al.*, "The protein data bank," *Nucleic Acids Res*., vol. 28, 235–242, 2000.
[23] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner *et al.*, "SCOP: A structural classification of proteins," *Nucleic Acids Res*, vol. 25, vol. 254-256, 1997.
[24] J. Sun, X. J. Wu, W. Fang *et al*., "Multiple sequence alignment using the Hidden Markov Model trained by an improved quantum-behaved particle swarm optimization," *Information Sciences*, vol. 182, pp. 93-114, 2012.
[25] Y. L. Sun, J. Y. Lee, K. S. Jung *et al.*, "A 9-state Hidden Markov model using protein secondary structure information for protein fold recognition," *Computers in Biology and Medicine*, vol. 39, pp. 527-534, 2009.



Haixia Long was born in Jiangsu, China, on February 1, 1980. She received the PhD in computer application technology from the University of Jiangnan, Wuxi, Jiangsu, China, in 2010. Her major field of study is bioimformatics.

Since 2010, she has been an associate professor in Information Science and Technology College, Hainan Normal University. She is the author or coauthor of more than 20 papers. Her research interests include artificial intelligence and bioinformatics.

Shulei Wu was born in Hainan, China, on May 29, 1974. She received the M.S. degrees in computer application technology from the University of ChongQing, ChongQing, China, in 2005.

Since 2011, she has been a professor in Information Science and Technology College, Hainan Normal University. She has served as a reviewer for International Conference on Computational Intelligence and Security. She is the author or coauthor of more than 20 papers. Her research interests include remotely sensed imaging, image processing and video retrieval.

Chun Shi received the M.Sc. degree and Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2008 and 2011, respectively. He is now an associate professor at the School Of Information Science and Technology, Hainan Normal University, P. R. China. His research interest covers wireless medium access control protocols, Ad Hoc networks and software design.