Fast Algorithm for Pseudoknotted RNA Structure Prediction

Hengwu Li
School of Computer Science and Technology, Shandong Economic University, Jinan 250014, China
Email: hengwuli@mail.sdu.edu.cn

Abstract— **Pseudoknotted RNA structure prediction is an important problem in bioinformatics. Existing polynomial time algorithms can handle only limited types of pseudoknots, or have too high time or space to predict long sequences. In this paper a heuristic algorithm is presented to maximize stems and predict arbitrary pseudoknots with $O(n^3)$ time and $O(n)$ space for a large scale of 5000 bases. Compared with maximum weighted matching algorithm, our algorithm reduce space complexity from $O(n^2)$ to $O(n)$; and the experimental results show that its sensitivity is improved from 80% to 87.8%, and specificity is increased from 53.7% to 75.9%. Compared with genetic algorithm with the accuracy of 83.3% and simulated annealing algorithm with the accuracy of 79.7%, our algorithm increases the predicted accuracy to 87.5%.**

Index Terms— **RNA structure, algorithm, pseudoknots**

I. Introduction

RNA structures prediction plays an important role in functional analysis of RNA molecules. Among the most prevalent RNA structures is a motif known as pseudoknot. Pseudoknots play a variety of diverse roles in biology. These roles include forming the catalytic core of various ribozymes, self-splicing introns, and telomerase. Additionally, pseudoknots play critical roles in altering gene expression by inducing ribosomal frameshifting in many viruses [1]. Plausible pseudoknotted structures have been proposed (Pleij et al.) in 1985 and confirmed (Kolk et al.) in 1998 for the 3' end of several plant viral RNAs, where pseudoknots are apparently used to mimic tRNA structure [2]. Recently, pseudoknots were confirmed in some RNAs of humans and many other species [3], [4].

Currently pseudoknot is not included in the majority of the study for RNA secondary structure prediction. The best algorithm predicts RNA secondary structure without pseudoknots with $O(n^3)$ time and $O(n^2)$ space and is implemented by MFOLD and ViennaRNA programs. Finding the best secondary structure including arbitrary pseudoknots has been proved to be NP-hard [5].

An approach to pseudoknot prediction is the maximum weighted matching (MWM) algorithm, considering only the base paired action and no stacking action. It folds an optimal pseudoknotted structure in $O(n^3)$ time with low accuracy and seems best suited to folding sequences for which a previous multiple alignment exists [6].

Another approach adopts dynamic programming to predict the tractable subclass of pseudokonts based on complex thermodynamic model in $O(n^4) - O(n^6)$ time [7]–[9].

Most methods for RNA folding which are capable of folding pseudoknots adopt heuristic search procedures and sacrifice optimality. Examples of these approaches include genetic algorithms and simulated annealing algorithms [10], [11].

Adjacent base pairs form stack, stacking and base pairing actions in RNA molecules are the most primary and stable actions. The adjacent continuous stack constitute stem, and crossed stems construct pseudoknots, so search the optimal structures based on combination with stem zones has become new method to RNA structure prediction, such as stem zone stacking algorithm presented by Benedeti and stem zone random stacking algorithm presented by Li WJ [12], [13], predict nested secondary structures. Recently Ruan give a heuristic algorithm based on stem zone to predict secondary structure containing pseudoknots with $O(n^4)$ time and $O(n^2)$ space [14].

Now the existing polynomial time algorithms can handle only limited types of pseudoknots, or have too high time and space to predict long RNA molecular. In this paper for pseudoknotted RNA secondary structure prediction, considering only stacking energy and neglecting other secondary role, a heuristic algorithm is presented to predict pseudoknotted RNA structure to maximize stems. It takes $O(n^3)$ time and $O(n)$ space to predict RNA sequences with 5000 bases, and have good accuracy with the sensitivity of 87.5% and specificity of 78.9% for the sequences in Pseudobase database[15].

In section 2 we give the energy model for pseudoknotted RNA secondary structure prediction. In section 3 a heuristic algorithm is presented. In section 4 we briefly conclude the paper.

II. RNA Structure Prediction

Let $s = s_1, s_2, ..., s_n$ be an RNA sequence, base $s_i \in \{A, U, C, G\}, 1 \leq i \leq n$. The subsequence $s_{i,j} = s_i, s_{i+1}, ..., s_j$ is a segment of s, $1 \leq i \leq j \leq n$.

If $si\&sj \in \{A\&U, C\&G, U\&G\}$, then s_i and s_j may constitute base pair (i, j). Each base can at most take part in one base pair. RNA secondary structure S is a set of

Manuscript received August 12, 2009; revised November 11, 2009; accepted December 21, 2009.

This work was supported by NSFC grants 603077, Shandong Province Natural Science Foundation of China (No. Y2008G37, Y2008G29, Y2007G59) and the Research Foundation of Shandong Economic University (No. 08009).

doi:10.4304/jcp.5.7.1100-1104

base pairs for s. If (i,j) and $(i+1,j-1) \in S$, base pairs (i,j) and $(i+1,j-1)$ constitute stack $(i,i+1:j-1,j)$, and $m(\geq 2)$ consecutive stacks form the stem $(i,i+m:j-m,j)$ with the length of $m+1$. The energy of stem $(i,i+m:j-m,j)$ is denoted as $E(i,i+m:j-m,j)$.

If base pairs (i,j) and (k,l) are parallel ($i<j<k<l$ or $k<l<i<j$) or nested ($i<k<l<j$ or $k<i<j<l$), then base pairs (i,j) and (k,l) are compatible, otherwise base pairs (i,j) and (k,l) constitute pseudoknots ($i<k<j<l$ or $k<i<l<j$) as Fig.1.

The major driving force of structure formation for RNA molecules is Watson-Crick base pair and wobble G, U base pair, and in particular stacking of adjacent base pairs [16]. Pseudoknotted RNA secondary structure is a set of base pair. Base pair and internal unpaired bases construct loops. Stack doesn't contain unpaired bases, and any other kinds of loops contain one or more unpaired bases. Since unpaired bases are destabilizing, stack is the only type of loops that stabilize the secondary structure [16]. Therefore for pseudoknotted RNA structure prediction, we give the general energy model considering only stacking energy and neglecting other secondary role.

Definition 1: SEM (stacking energy model of pseudoknotted RNA structure prediction)

For RNA sequence s, $s \in \{A,U,C,G\}^*$, a secondary structure S is a set of base pairs such that if $(i,j) \in S$ then

1) $(i',j') \in S$, if $(i,j)\bigcap(i',j') \neq \emptyset$, then $(i,j) = (i',j')$.

2)$(i,j) \in S$, $(i,j) \in (A,U),(C,G),(U,G)$ and $j-i \geq 4$.

3) if $(i+1,j-1) \in S$, then (i,j) and $(i+1,j-1)$ form stack with the energy of $E(i,i+1:j-1,j)$.

4) if $(i+1,j-1),(i',j'),(i'+1,j'-1) \in S, s_i = s_{i'}, s_j = s_{j'}$and $s_{i+1} = s_{i'+1}, s_{j-1} = s_{j'-1}$, then $E(i,i+1:j-1,j) = E(i',i'+1:j'-1,j')$. That is, the size of stacking force is determined by base pair itself and adjacent bases pair.

5) if $(i+1,j-1) \in S$, then the energy of S is $E(S) = \Sigma_{1\leq i<j\leq n} E(i,i+1:j-1,j)$.

So the problem of pseudoknotted RNA structure prediction is to find a secondary structure S containing pseudoknots with minimal energy for given RNA sequence s under SEM model.

III. Heuristic Algorithm

The maximum weighted matching algorithm considers only the base paired action and no stacking action, so calculated results contain many redundant base pairs, which don't exist in real structure.

Adjacent base pairs form stack, and the adjacent continuous stack constitute stem. A pseudoknot is the cross of two stems. Stems reduce folding free energy of pseudoknots structure, and are the main action formation to stabilize RNA structure. The more the number of base pairs is, the more stable the stem is. So we study stem problem to find the key to RNA structure prediction.

A. Algorithm

Now existing polynomial time algorithms can handle only limited types of pseudoknots, or have too high time and space to predict long RNA molecular. Based on relatively stable structure characteristics of stem, we design a heuristic algorithm GS for pseudoknotted RNA structure prediction to maximize stems as follows.

GS algorithm:

Given sequence s, let E be the energy of the maximal stem.

step 1:

Search the maximal stem.

$min = 0$

For $i = 1$ to $n-8$

For $j = 9$ to n

If $s_{i.i+1}$ and $s_{j-1,j}$ form stack, calculate the energy of maximal stem

If $E < min$, $min = E$, store the start position, the end position and the length of maximal stem.

End for

End for

step 2:

If $(min < 0)$, output the maximal stem with minimal energy. Replace the bases in maximal stem with empty string, then go to step1.

First we calculate the energy of all impossible stems in the sequence, and search the maximum stem. Then marked the bases in the maximum stems and make them can't be paired in the back calculation. The same handle is implemented until no stem.

If $(i,j) \in S$, then several stems may both start with (i,j). But we need only select the maximal stem from them in GS algorithm, and the number of the selected stems is at most $O(n^2)$.

In GS algorithm, the bases in maximal stem are replaced with empty, not be delete and be paired in the back calculation. The sequence and position of the bases in maximal stem are kept, so that the redundancy stack is eliminated, which don't exist in real structure, and the accuracy of calculation is improved.

If we denote stem list with matrix, the space complexity is $O(n^2)$. We use variables to store the positions and length of two ends to stem, and only store the maximal stem and no stem list in every loop, so the space complexity is reduced to $O(n)$.

B. Complexity

Only the data of maximal stem is stored in GS algorithm to reduce the time and space complexity.

The dual loop were used in the search of stem, its time complexity is $O(kn^2)$(k is the length of stem). Let the length of searched stem k_1,k_2,k_m, then $k_1+k_2++k_m \leq n/2$. So the time complexity of search for maximal stem is $O(k_1n^2+k_2n^2++k_mn^2) \leq O((n/2)n^2) = O(n^3/2)$, and the time complexity of GS algorithm is $O(n^3)$.

Only one-dimensional array and auxiliary variables are used in the search of stem, so the space complexity of GS algorithm is $O(n)$.

C. Test results

There is no nested restriction to the stems in GS algorithm, so the calculated structures can contain psudoknots.

We test the sequences in PseudoBase. Computer test results show that the accuracy rate of GS algorithm is more than 95 percent for about 20 percent sequences, and more than 78 percent for about 60 percent sequences. The algorithm can calculate the sequence with 5000 bases.

Four experimental examples are given as follows. They are all computed in $1s$. In the results, the unpaired bases are denoted as :, and base pairs are denoted as () and [].

E1:

brome mosaic virus
:(((:[[[[[])):::]]]]]:
Output is as Fig.1.
:(((:[[[[[])):::]]]]]:
The accuracy is 100%.

E2:

physalis mottle virus
::::::::::::::(((::[[[[[[ ))):::]]]]]]::::
Output is as Fig.2.
:(((:::::)))(((::[[[[[[ ))):::]]]]]]::::
The accuracy is 100%.

E3:

odontoglossum ringspot virus
:(((((((((::::[[[[(((((:::::(((((((((:::::)))))))))))))))))))))
:::::::]]]]:
Output is as Fig.3.
:(((((((((::::[[[[(((((:::::(((((((((:::)))))))))))))))))))))
:::::::]]]]:
The accuracy is 100%.

E4:

Homo sapiens mRNA
:::::::((((((((((::[[[[[::[[[[[::::[[[[[[[[::[[[::[[[[:
:::))))))):::::::]]]]:]]]]]]]]]]:::]]]]]:]]]]]::::]]]:]]]]]
Output is as Fig.4.
:::::::((((((((((((((:::[[[[[:::::[[[[[[[[[[[[::)))))))
::)))))))]]]]:::(((((::::]]]]]]]:::)))))):]]]]]::::::
The accuracy is 50%.

Figure 1. The result of brome mosaic virus

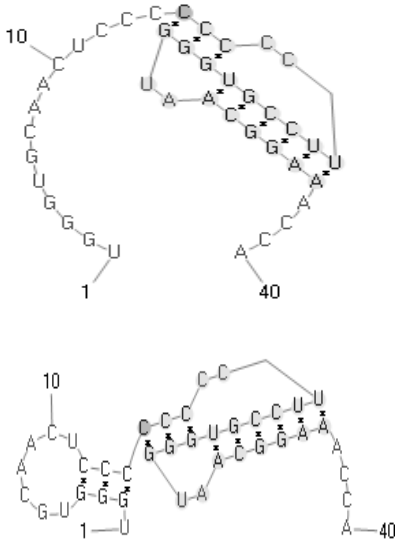

Figure 2. The result of physalis mottle virus
The above one is the result of comparative sequence analysis, another is that of GS algorithm.

IV. Comparison of Performance

According to the valuation criterion of gene prediction algorithm, presented by Burset and Guigo in 1996, we compare the predicted result of GS algorithm with that of comparative sequence analysis, using sensitivity and specificity as a target. The equation of sensitivity and specificity is as follows:

Sensitivity (Sn):

$$SN = \frac{TP}{TP + FN} \tag{1}$$

Specificity (Sp):

$$SP = \frac{TP}{TP + FP} \tag{2}$$

TP (true positive): the correctly predicted number of base pair for true base pairs.

FN (false negative): the incorrectly predicted number of base pair for true base pairs.

FP (false positive): the predicted number of base pair for true single bases.

The sensitivity and specificity of four experiments with GS algorithm are showed in Table.I. The average sensitivity of GS algorithm is 87.5% and the average specificity of GS algorithm is 78.9%.

MWM algorithm is the optimal combination one, PKNOTS algorithm is the minimal energy one, and ILM is the heuristic one based on stem, so we compare GS algorithm with above three algorithms. ILM algorithm use six sequences to compute and compare the performance, we use the same six sequences to compare their sensitivity and specificity as Table.II.

It is showed that the performance of GS algorithm is better than MWM algorithm, and many redundant base pairs are deleted, through replacing delete with mark.

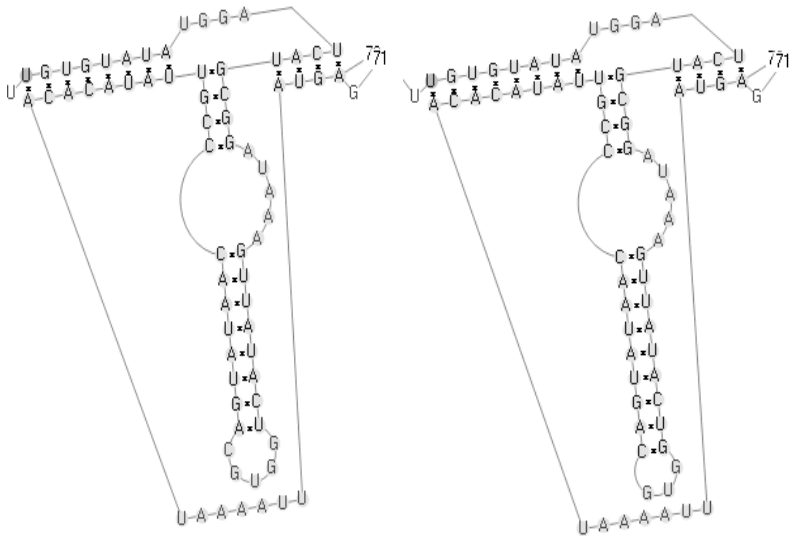

Figure 3. The result of odontoglossum ringspot virus
The left one is the result of comparative sequence analysis, another is that of GS algorithm.

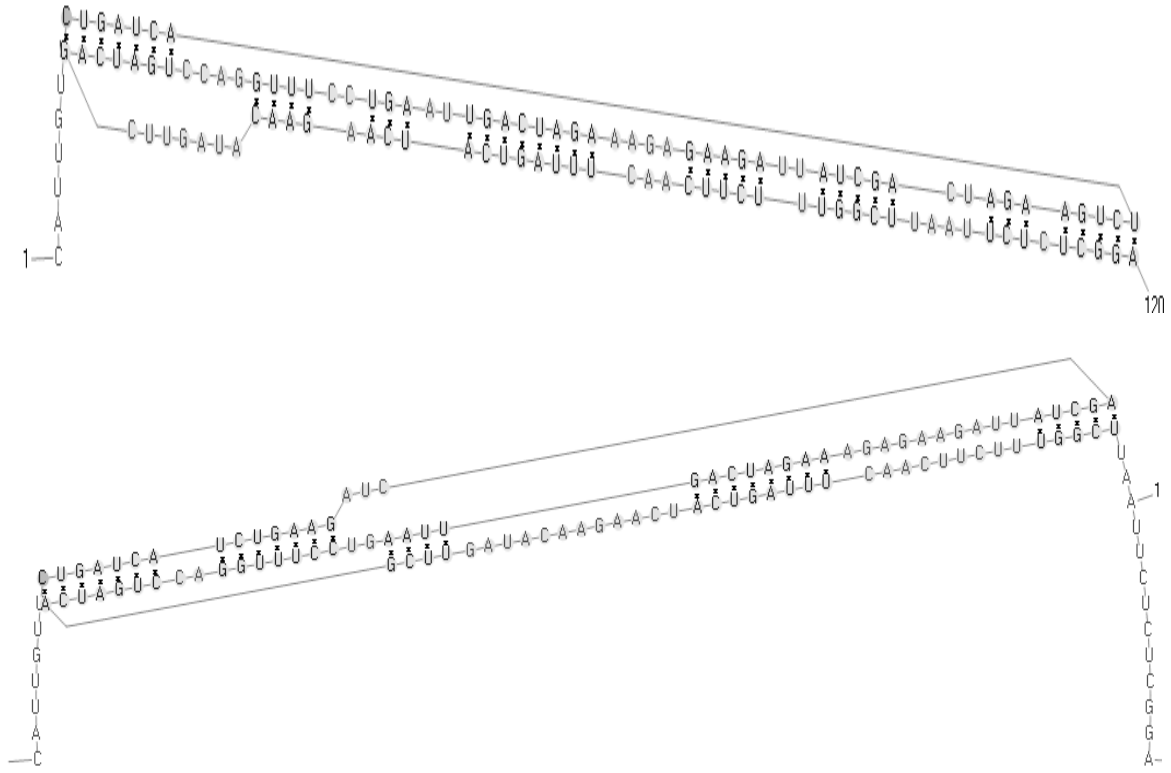

Figure 4. The result of Homo sapiens mRNA
The above one is the result of comparative sequence analysis, another is that of GS algorithm.

For single sequence, PKNOTS algorithm gets the optimal structure, ILM algortihm computes the optimal structure in each iterated loop, GS algorithm computes the maximal stem in each iterated loop. So the accuracy of PKNOTS and ILM algorihm is better than GS algorithm. But compared with MWM, PKNOTS and ILM algorithm, GS algorithm remarkably reduces time and space, through replacing base pairs with stem.

We compare GS algorithm with genetic and simulated annealing algorithms as Table.III. It is showed that the

TABLE I.
EXPERIMENTAL DATA

Experiment	1th	2nd	3rd	4th	Average
The number of bases	21	40	69	120	63
The number of base pairs	8	9	24	40	20
The number of predicted base pairs	8	12	25	45	23
The number of correct base pairs	8	9	24	20	15
Sensitivity	100%	100%	100%	50%	87.5%
Specificity	100%	75%	96%	44.4%	78.9%

TABLE II.
COMPARISON OF GS ALGORITHM WITH MWM, PKNOTS AND ILM ALGORITHM

Sequence	**MWM**		**PKNOTS**		**ILM**		**GS**	
Target	SN%	SP%	SN%	SP%	SN%	SP%	SN%	SP%
HIV-1-RT	100	84.6	100	100	100	100	100	100
TYMV	100	63.2	100	96.0	100	82.8	100	100
TMV-3'up	68.0	41.5	52.0	59.1	80.0	80.0	80.0	60.6
TMV-3'down	73.5	49	97.0	97.0	76.5	68.4	58.8	46.5
HDV	67.8	45.2	85.7	75.0	100	82.4	96.4	81.8
Anti-HDV	70.8	38.6	95.8	69.7	100	66.7	91.7	66.7
Average	80	53.7	88.4	82.8	92.8	80.0	87.8	75.9

predicted accuracy of GS algorithm is better than genetic and simulated annealing algorithm.

TABLE III.
COMPARISON OF GS ALGORITHM WITH GENETIC AND SIMULATED ANNEALING ALGORITHM

Algorithm	Accuracy
GS algorithm	87.5%
Genetic algorithm	83.3%
Simulated annealing algorithm	79.7%
Genetic and simulated annealing algorithms	87.0%

V. CONCLUSION

In this paper SEM model is built based on base pair stacking force and neglecting other secondary role, and a heuristic algorithm with $O(n^3)$ time and $O(n)$ space is presented to predict pseudoknotted RNA structure under the model.

Computer test results show that the fast algorithm can deal with a large scale of 5000 bases and have good accuracy with the sensitivity of 87.5% and specificity of 78.9% for the sequences in Pseudobase database.

Compared with maximum weighted matching algorithm, our algorithm reduce space complexity from $O(n^2)$ to $O(n)$; and the experimental results show that its sensitivity is improved form 80% to 87.8%, and specificity is increased from 53.7% to 75.9%. Compared with genetic algorithm with the accuracy of 83.3% and simulated annealing algorithm with the accuracy of 79.7%, our algorithm increases the predicted accuracy to 87.5%.

ACKNOWLEDGMENT

We thank the reviewers for their detailed comments.

REFERENCES

[1] D.W. Staple, and S.E. Butcher, "Pseudoknots: RNA structures with diverse functions," *PLoS. Biol.*, vol. 3, pp. 213, 2005.
[2] M. H. Kolk, M. van der Graff, S. S. Wijmenga, C. W. A. Pleij, H. A.Heus, and C. W. Hilbers, "NMR structure of a classical pseudoknots: interplay of single- and double-stranded RNA," *Science*, vol. 280, pp. 434-438, 1998.
[3] D.H, Mathews, and D.H.Turner, "Prediction of RNA secondary structure by free energy minimization," *Current Opinion in Structural Biology*, vol. 16, pp. 270-278, 2006.
[4] I.Barette, G.Poisson, and P. Gendron, "Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching," *Nucleic Acids Research*, vol. 29, pp. 753-758, 2001.
[5] R.B. Lyngsy, and C.N. Pedersen, "RNA pseudoknot prediction in energy based models," *Journal of Computational Biology*, vol. 7, pp. 409-428, 2001.
[6] J.E. Tabaska, R.B. Cary, H.N. Gabowad, and G.D. Stormo, "An RNA folding method capable of identifying pseudoknots and base triples," *Bioinformatics*, vol. 14, pp. 691-699, 1998.
[7] S.R. Rivas and Eddy, "A dynamic programming algorithm for RNA structure prediction includdimg pseudoknots," *Journal of Molecular Biology*, vol. 285, pp. 2053-2068,1999.
[8] J. Reeder, and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinformatics*, vol. 5, pp. 104, 2004.
[9] Li Hengwu, Zhu Daming, Liu Zhendong, and Li Hong, "Prediction for RNA planar pseudoknots," *Progress in Nature Science*, vol. 17, pp. 717-724, 2006.
[10] A.P. Gultyaev, F.H. van Batenburg, and C.W.A Pleij, "The computer simulation of RNA folding pathways using a genetic algorithm," *J. Mol. Biol.*, vol. 250, pp. 37-51, 1995.
[11] M. Schmitz, and G. Steger. "Description of RNA folding by Simulated Annealing," *J.Mol. Biol.*, vol. 255, pp. 254-266, 1996.
[12] G. Benedeti, and P.D. Santis, "New method to find a set of energetically optimal RNA secondary structure," *Nucl.Acids.Res.*, vol. 17, pp. 5149-5161, 1989.
[13] W.J. Li, and J.J. Wu, "Prediction of RNA secondary structure based on helical regions distribution," *Bioinformatics*, vol. 14, pp. 700-706, 1998.
[14] J. Ruan, G.D. Stormo, and W. Zhang, "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots," *Bioinformatics*, vol.20, pp. 58-66, 2004.
[15] PseudoBase homepage: *http://wwwbio.LeidenUniv.nl /~Batenburg/PKB.htm*
[16] S. Ieong, M.Y. Kao, T. W. .Lam, W.K. Sung, and S.M. Yiu, "Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs," *Journal of Computational Biology*, vol. 6, pp. 981-995, 2003.

Hengwu Li was born in Jiaonan city, China, in 1969, received his PhD degree in Software and Theory of Computer from Shandong University, China, in 2008.

He is currently an Associate Professor in School of Computer Science and Technology of ShanDong Economic University, China. His current research interests include the design of algorithm, biological computer, and intelligent computer.