An Efficient Algorithm Based on Hopfield Neural Network for RNA Secondary Structure Prediction

Yanqiu Che†, Zheng Tang†, and Shaozhi Liu††

†Faculty of Engineering, Toyama University, Toyama, 930-8555 Japan
††Hokuriku Computer Service Co. Ltd., Toyama, 939-2708 Japan

Summary
During the past years, many improvements have been made in computational methods for the prediction of RNA secondary structure to provide insight into many functions RNAs perform in biology processes. In this paper, we propose an efficient algorithm that can increase energy temporarily and then energy declines again by introducing stochastic dynamics into the Hopfield neural network. The proposed algorithm can help the Hopfield network escape from local minima and find the optimal or near-optimal solutions. The proposed algorithm has been applied to RNA secondary structure prediction problem. Simulation results verify that it has the ability to search the more stable RNA secondary structure for an RNA sequence compared to other neural network methods.

Key words:
RNA secondary structure prediction; Hopfield neural network; stochastic dynamics

1. Introduction

RNAs are molecules that are important for many processes in the cell. A molecule of RNA consists of a long chain of subunits, called nucleotides. Each nucleotide contains one of four possible bases: Adenine(A), Guanine(G), Cytosine(C), and Uracil(U). Under normal physiological conditions, a nucleotide chain can fold back upon itself by forming pairs of bases. The base pairing of RNA is generally called the secondary structure which determines how the RNA will interact and react with other components. Yet base pairing dose not occur arbitrarily. A-U and C-G form stable pairs and are known as the Watson-Crick base pairs [1]. Other base pairs are less stable and often ignored.

Work on the determination of RNA structure has been carried out for decades by a number of research groups. There are two direct methods including using X-ray crystallography [2][3]and using NMR spectroscopy [4][5] to determinate RNA structure. Appealing computational methods for the prediction of RNA structure have been developed to provide insight into functions that RNA has. Searching for configurations of maximum base pairing or of minimum free energy is the general approach. Many algorithms have been proposed for predicting RNA secondary structure with respect to minimum free energy. Early algorithm was made by Zuker and Stiegler[6]. The Zuker's algorithm (implemented in the programs called MFOLD[7]) is an efficient dynamic programming algorithm for identifying the globally minimal energy structure for an RNA sequence, as defined by such a thermodynamic model[8]. Zuker's energy calculations have been further improved [9][10]and are probably the most used RNA secondary structure prediction method today. Another algorithm concerning minimum free energy makes use of artificial neural networks. Artificial neural networks are models of highly parallel and adaptive computation, based very loosely on current theories of brain structure and activity, and have been applied with some success to optimization problems such as TSP. In 1992, Takefuji presented an algorithm based on the Hopfield neural network for RNA secondary structure prediction [11]. But the major weakness of this algorithm is still due to its failure in finding the global minimum solution.

In this paper, we propose an efficient algorithm that can increase energy temporarily and then energy declines again by introducing stochastic dynamics into the Hopfield neural network. The proposed algorithm can help the Hopfield network escape from local minima and get an optimal or a near-optimal solution. This algorithm has been applied to RNA secondary structure prediction problem and simulation results verify that the proposed algorithm has the ability to search the more stable RNA secondary structure for an RNA sequence compared to other neural network methods.

2. Problem Formulation

The first stage of our algorithm is to select all the possible helices for a given RNA sequence. A helix refers to an anti-parallel complementary strand whose length must be greater than or equal to 3 base pairs, as shown in Fig, 1.

Manuscript received May 11, 2007.
Manuscript revised May 25, 2007.

The Watson-Crick base pairs (A-U and C-G) are permitted in the helix and the minimum length of a hairpin loop is 3 bases.

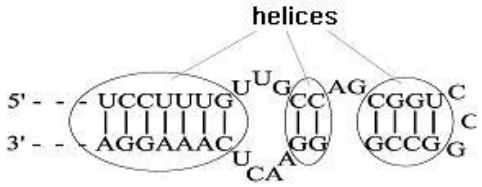


Fig. 1 A simple graph shows helices of RNA secondary structure

The stability of RNA secondary structure is evaluated by free energy. Much progress has been made on the problem of assigning free energy values to substructures. The most useful free energy data have been extrapolated from experiments on particular kinds of RNA carried out by Tinoco and Uhlenbeck [12][13]. For a helix, the free energy is calculated according to Table 1(units is Kcal/mol).

Table 1: Free energy calculation

5'-3'/3'-5'	A-U	U-A	G-C	C-G
A-U	-1.2	-1.8	-2.1	-2.1
U-A	-1.8	-1.2	-2.1	-2.1
G-C	-2.1	-2.1	-4.8	-4.8
C-G	-2.1	-2.1	-3.0	-4.8

For a given RNA sequence $\boldsymbol{S}$, $\boldsymbol{R}$= $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, $\boldsymbol{R}_3$, ...$\boldsymbol{R}_n$ is a set of helix, and $E_1, E_2, E_3, \ldots E_n$ are free energy values of these helices calculated according to Table 1. The most stable secondary structure must have the lowest free energy and have no inconsistencies. Base pairs with the cross or overlap states are called inconsistencies.

According to the above conditions, this optimization problem can be formulated by an objective function whose minimum value corresponds to the most stable RNA secondary structure. In a reasonable formulation, there are two components to the objective function: one is used to select stack domain candidates where the sum of free energy is the lowest; the other is used to guarantee there is no inconsistency in RNA secondary structure. Thus, this optimization problem can be mathematical formulated as follows:

$$E = \sum_{i=1}^{n} E_i V_i + \sum_{i=1}^{n} \left(|E_i| \sum_{j=1}^{n} c_{ij} V_i V_j \right) \quad (1)$$

$$V_i = \begin{cases} 1 & if\ R_i\ is\ selected \\ 0 & otherwise \end{cases}, i \in n \quad (2)$$

where c_{ij} is a factor that indicates there is inconsistency or not. If both $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$ are selected and there is inconsistency between them, then c_{ij} = 1; If both $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$ are selected and there is no inconsistency between them, then c_{ij} = 0.

3. Description of the Proposed Algorithm

Hopfield neural networks (HNN) have provided a parallel and powerful method of solving difficult optimization problems [14][15][16]. But, due to its inherent local minimum problem, the global minimum or good solution is usually difficult to be found [17]. By incorporating stochastic dynamics into the Hopfield neural network, we propose an efficient algorithm that allows energy to be increased temporarily, which helps the network escape from local minima.

Fig.2 is a conceptual graph of the energy landscape with a local minimum and a global minimum which shows the dynamics of the proposed improved network simply. The X-coordinate denotes the state of the network and the Y- coordinate denotes the value of energy function. For example, if the network is initialized onto point *A* (Fig.2(a)). Because of the mechanism of the HNN, the state of network moves towards negative gradient direction and reaches the local minimum. If we change the dynamics of the HNN by introducing stochastic term at point *B* to increase the energy temporarily, point *B* will become a new point *B*'(Fig.2(b)). From point *B*', the network returns to move towards negative gradient direction and reaches the global minimum point *C* (Fig.2(b)).

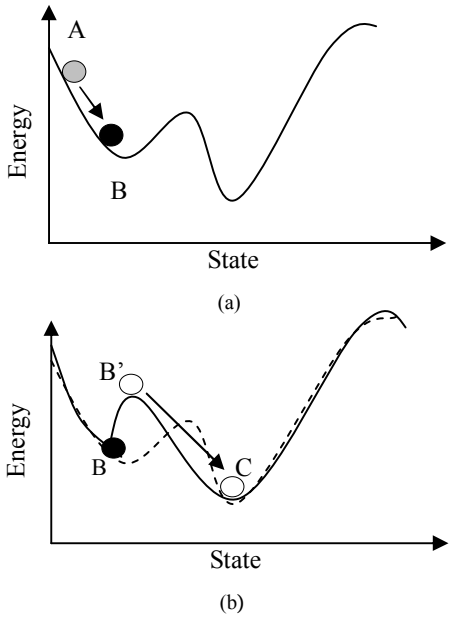


Fig. 2 The relation between energy and state transition in the learning process of the HNN with two stable states.

The energy function of the HNN at time t is given by:

$$E(t) = (-\frac{1}{2})\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}V_i(t)V_j(t) + \sum_{i=1}^{n}\theta_i V_i(t) \quad (3)$$

The inputs of the neurons are computed by the updating rule:

$$U_i(t) = \sum_{j=1}^{n} w_{ij}V_j(t) - \theta_i \quad (4)$$

The input-output function is:

$$V_i(t+1) = \begin{cases} 1 & U_i(t) \geq 0 \\ 0 & U_i(t) < 0 \end{cases} \quad (5)$$

where U_i is the input, V_i is the output, w_{ij} is the symmetric interconnection strength form neuron j to neuron i (w_{ii}=0), and θ_i is the offset bias.

In the proposed algorithm, the input and output of neuron i is modified as follows:

$$u_i(t) = \lambda(t)U_i(t) \quad (6)$$

$$V_i(t+1) = \begin{cases} 1 & u_i(t) \geq 0 \\ 0 & u_i(t) < 0 \end{cases} \quad (7)$$

$$\lambda(t) = random(h(t),1) \quad (8)$$

$$h(t) = 1 - 2e^{-t/\beta} \quad (9)$$

Fig.3 shows how the value of h changes with t for β=60. Obviously, h(t) increases form -1 to 1 while the updating proceeds.

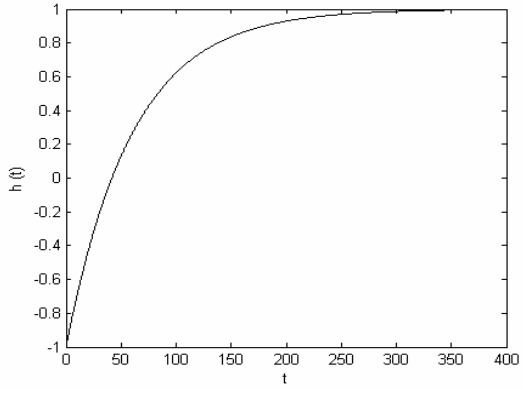


Fig. 3 Graph of $h(x) = 1 - 2e^{-t/\beta}$ with β=60.

The following analysis can explain the dynamics of our proposed algorithm which includes stochastic dynamics. Suppose that at time t, the state of neuron k is changed. The change in the state of neuron k is:

$$\Delta V_k(t) = V_k(t+1) - V_k(t) \quad (10)$$

The energy caused by the change in the state of neuron k is:

$$\begin{aligned} E(t+1) = & \left[-\frac{1}{2}\sum_{i=1,i\neq k}^{n}\sum_{j=1,j\neq k}^{n} w_{ij}V_i(t)V_j(t) + \sum_{i=1,i\neq k}^{n}\theta_i V_i(t) \right] \\ & + \left[-\frac{1}{2}\sum_{j=1,j\neq k}^{n} w_{kj}V_k(t+1)V_j(t) - \frac{1}{2}\sum_{i=1,i\neq k}^{n} w_{ik}V_i(t)V_k(t+1) \right] \\ & + \theta_k V_k(t+1) \end{aligned} \quad (11)$$

The change of energy caused by the change in the state of neuron k is:

$$\begin{aligned} \Delta E = & -\frac{1}{2}\sum_{j=1,j\neq k}^{n} w_{kj}(V_k(t+1)V_j(t) - V_k(t)V_j(t)) \\ & -\frac{1}{2}\sum_{i=1,i\neq k}^{n} w_{ik}(V_i(t)V_k(t+1) - V_i(t)V_k(t)) + \theta_k(V_k(t+1) - V_k(t)) \end{aligned} \quad (12)$$

Then replace i with j. Because of w_{kj}=w_{jk}, Eq. (12) can be reduced to:

$$\begin{aligned} \Delta E &= -\sum_{j=1,j\neq k}^{n} w_{kj}(V_k(t+1)V_j(t) - V_k(t)V_j(t)) + \theta_k(V_k(t+1) - V_k(t)) \\ &= -\Delta V_k(t)\left(\sum_{j=1,j\neq k}^{n} w_{kj}V_j(t) - \theta_k \right) \end{aligned} \quad (13)$$

Using Eq. (4), Eq. (13) can be rewritten as follows:

$$\Delta E = -\Delta V_k(t)U_k(t) \quad (14)$$

Initially, when h(t)<0, it is possible that $\lambda(t)$<0 according to Eq.(8). Consider two cases. Case a: If $U_k(t)\geq 0$, then $u_i(t)\leq 0$. According to Eq. (7) the value of $V_k(t+1)$ keeps 0 or changes from 1 to 0, thus $\Delta V_k(t)\leq 0$. Case b: If $U_k(t)<0$, then $u_i(t)>0$. According to Eq. (7) the value of $V_k(t+1)$ keeps 1 or changes from 0 to 1, thus $\Delta V_k(t)>0$. Consider Case a and Case b, from Eq. (14) we can see that when $\lambda(t)<0$, $\Delta E>0$ is possible. This increase of energy can help the network escape from local minima when the network suffers from local minima. The possibility of an increase of energy becomes smaller as h(t) increases until

$h(t)>0$. When $h(t)>0$, $\lambda(t)>0$. The dynamics of the network tends toward the original updating mode and finally the network has the steepest descent which guarantees that the network always converges to a stable state. Thus the proposed algorithm provides a mechanism for escaping from local minima and converging to a better stable state by introducing stochastic dynamics.

The proposed algorithm always permits descent of the energy function, but ascent of the energy is allowed initially and becomes less likely as time goes. However, it is not guaranteed that the proposed algorithm will always find the global minimum. Instead, usually it has higher search capability than the original Hopfield neural network.

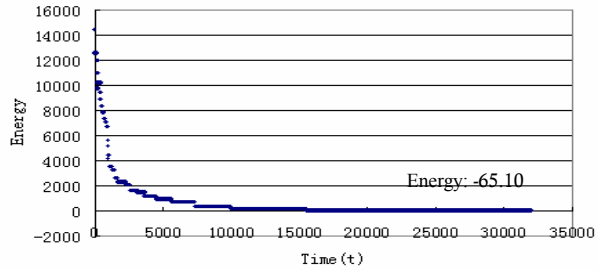
4. Test Results

The algorithm proposed in this paper has been applied to find the optimal or near-optimal structure of an RNA sequence based on free energy rules. The original Hopfield neural network and Tekefiji's algorithm were also executed for comparison. Each algorithm has been carried out on five different RNA sequences: 61, 77, 120, 401 and 511 bases. We performed 100 runs of each algorithm on all RNA sequences with different initial values. The parameters, β=30, were used in the simulations of the proposed algorithm. The performance of these algorithms was summarized in Table 2. The results that we recorded for each RNA sequence were the lowest energy among 100 runs. The columns "Length" represent the shortest length of possible helix is 2, 3, and 4 base pairs respectively. From Table 2, we can see that the proposed algorithm is comparable with the algorithm of Hopfield neural network and Takefuji's algorithm for small size problems (Example 1-2). For the larger size problems (instance 3-5), Table 2 shows that the proposed algorithm can find better solutions than other algorithms.

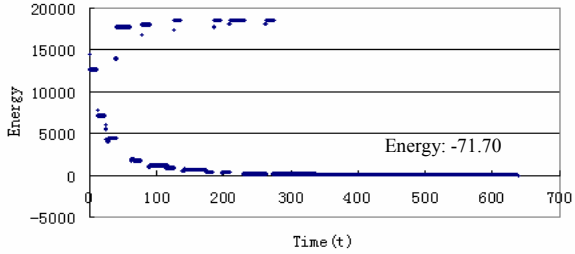
Table 2: Test results on RNA sequence

Example	length	Takefuji	HNN	Proposed
Example 1	2	-30.30	-30.90	-39.90
	3	-39.90	-39.90	-39.90
	4	-39.90	-39.90	-39.90
Example 2	2	-46.20	-37.80	-50.70
	3	-43.50	-40.80	-49.80
	4	-43.50	-43.50	-43.50
Example 3	2	-76.50	-69.60	-89.40
	3	-76.20	-74.70	-80.60
	4	-65.10	-71.70	-71.10
Example 4	3	no result	-107.40	-214.80
	4	-173.10	-103.50	-186.90
Example 5	3	no result	-116.40	-241.80
	4	no result	-113.40	-226.80

In order to show how the proposed algorithm works, we chose RNA sequence with 120 bases as an example. Fig.4(a) shows the evolution of energy function in Takefuji's algorithm. One can observe that the energy descended on every step and had a tendency to converge to the local minima. The evolution of energy function in the proposed algorithm is demonstrated in Fig.4(b). The energy in Fig.4(b) fluctuated with large changes before t =276 due to stochastic dynamics and then gradually descended. The fluctuations in the proposed algorithm helped the Hopfield network escape from the local minima and found a better solution.



(a)



(b)

Fig. 4 The evolution of energy function in Takefuji's algorithm and the proposed algorithm with 120 bases

In order to demonstrate the reliability of our algorithm, the secondary structure of $tRNA^{phe}$ and pre-$tRNA^{tyr}$ were predicted by our proposed algorithm, compared to those predicted by Zuker's algorithm (http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi). $tRNA^{phe}$ contains 76 bases and has a well-known cloverleaf structure[18](5' and 3' represent the start and the end of an RNA sequence):

5'---GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA---3'

pre-$tRNA^{tyr}$ secondary structure consists of five helices [19]. It contains 92 bases:

5'---CUCUCGGUAGCCAAGUUGGUUUAAGGCGCAAGACUGUAAUUUAUCACUACGAAAUCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGACCA---3'

The secondary structure of $tRNA^{phe}$ and pre-$tRNA^{tyr}$ predicted by our proposed algorithm are shown in Fig.5(a)

and Fig.6(a). Fig.5(b) and Fig.6(b) shows the minimum free energy secondary structure predicted by Zuker's method with free energy. The bases pairs in the rectangles are the true positive base pairs in the known secondary structure. The complex substructural components - such as hairpin loops and bulge loops have not been included in our algorithm. If we think more about it, we can achieve more accurate and reasonable prediction results (the figures of our predicted structures were obtained by the efn server).

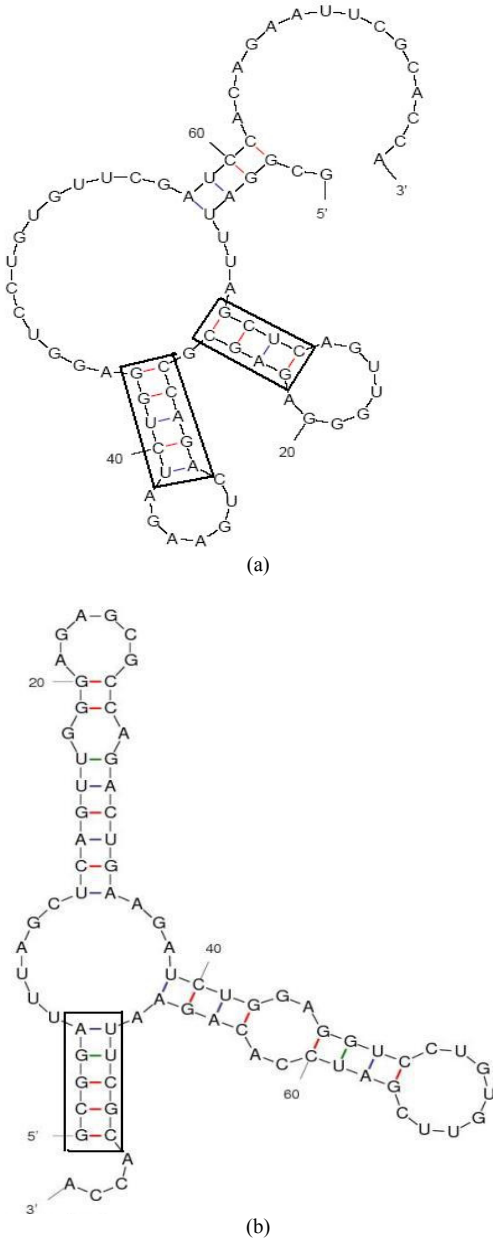


Fig. 5 tRNA^{phe} secondary structure predicted by the proposed algorithm and Zuker's method.

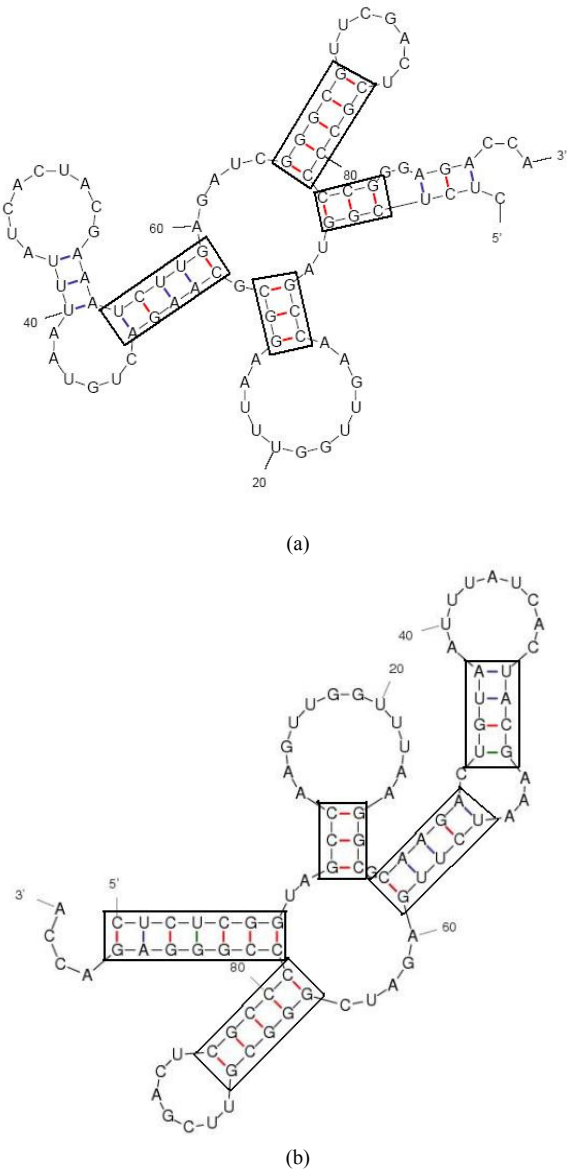


Fig. 6 Secondary structure of pre-tRNA^{tyr} predicted by the proposed algorithm and Zuker's method

5. Conclusion

This paper presented a new algorithm that introduces stochastic dynamics into the Hopfield network for RNA secondary structure prediction problem. The proposed algorithm can increase energy temporarily because of stochastic dynamics and then energy declines again, which helps the Hopfield network escape from local minima and find optimal or near-optimal solutions. This algorithm has been applied to RNA secondary structure problem and has been compared with other methods. The simulation results gave the evidence that the proposed algorithm had the

ability to search the more stable RNA secondary structure for an RNA sequence compared to other neural network methods.

References

[1] J. D. Watson, F. H. C. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," Nature, vol.171, pp.737-738, 1953.

[2] J. H. Cate, J. A. Doudna, "Solving large RNA structure by X-ray crystallography," Methods Enzymol, vol.317, pp. 169-180, 2000.

[3] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Cater, C. Vonrhein, T. Hartsch, V. Ramakrishnan, "Structure of the 30S ribosomal subunit," Nature, vol. 407, pp.327-339, 2000.

[4] G. Varani, I. Jr. Tinoco, "RNA structure and NMR spectroscopy," Q Rev Biohpys, vol.24(4), pp.479-532, 1991.

[5] P. J. Lukavsky, J. D. Puglisi, "RNAPack: an integrated NMR approach to RNA structure determination," Methods, vol.25(3), pp.316-332, 2001.

[6] M. Zuker, P. Stiegler, "Optimal computer folding of lager RNA sequences using thermodynamic and auxiliary information," Nucleic Acids Res., vol.9, pp.133-148, 1981.

[7] M. Zuker, "Compute prediction of RNA structure.," Meth. Enzymol, vol.180, pp.262-288, 1989.

[8] M. Zuker, D. H. Mathews, D. H. Turner, "Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide," In J. Barszewski, B. Clark (Eds.). RNA Biochem & Biotech, Kluwer Academic Publishers, Dordrecht, NL, pp11-43, 1999.

[9] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, M. Zuker, "Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding," Proc. Natl. Acad. Sci., vol.91, pp. 9218-9222, 1994.

[10] D. H. Mathews, J. Sabina, M. Zuker, D. H. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure," J. Mol. Biol., vol. 288, pp.911-940, 1999.

[11] Y. Takefuji, D. Ben-Alon, A. Zaritsky, "Neural computing in discovering RNA interactions," BioSystems, vol. 27(2), pp.85-96, 1992.

[12] I. Tinoco, O. C.Uhlenbeck, M. D. Levine, "Estimation of secondary structure in ribonucleic acids," Nature, vol.230, pp.362-367, 1971.

[13] I. Tinoco, P. N. Borer, B. Dengler, M. D. Levine,O. C. Uhlenbeck, D. M. Crothers, J. Gralla,"Improved estimation of secondary structure in ribonucleic acids," Nature New Biology, vol.246, pp.40-41, 1973.

[14] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational capabilities," Proceedings of the Nat. Acad. and Sci., vol.79, pp.2554-2558, 1982.

[15] J. J. Hopfield, "Neurons with graded response have collective computational capabilities like those of two-state neurons," Proceedings of the Nat. Acad. and Sci., vol. 81, pp. 3088-3092, 1984.

[16] J. J. Hopfield, D. W. Tank, "Neural computation of decisions in optimization problems," Biological Cybernetics, vol.52, pp.141-152, 1985.

[17] G. V. Wilson, G. S. Pawley, "On the stability of the Traveling Salesman Problems," Biological Cybernetics, vol.58, pp. 63-70,1988.

[18] M. Sprinzl, D.H. Gauss, "Compilation of sequences of tRNA genes," Nucleic Acids Res. vol.12(Suppl), 1984.

[19] H. M. Martinez, "An RNA folding rule," Nucleic Acids Res., vol.12, pp.323-334, 1984.

Yanqiu Che received a B.S. degree and an M.S. degree from Dalian University of Technology, Liaoning, China in 2002 and 2005 respectively. Now she is working toward the Ph.D degree at Toyama University, Japan. Her main research interests are neural networks, optimization problems and RNA structure prediction.

Zheng Tang received a B.S. degree from Zhejiang University, Zhejiang, China in 1982 and an M.S. degree and a D.E. degree from Tsinhua University, Beijing, China in 1984 and 1988, respectively. From 1990 to 1999, he was an Associate Professor in Miyazaki University, Miyazaki, Japan. From 2000, he is a professor in Toyama University, Toyama, Japan. His current research interests include intellectual information technology, neural networks, and optimizations.

Shaozhi Liu received a B.S. degree from Dalian University of Technology, Liaoning, China in 1999 and an M.S. degree from Toyama University, Toyama, Japan in 2005. Now he is working in Hokuriku Computer Service Co. Ltd., Japan.