International Journal of Soft Computing 1 (1) : 61-66, 2006

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

[1]YanQiu Che, [2]Qiping Cao and [1]Zheng Tang
[1]Department of Intellectual Information Systems Engineering 3190 Gofuku, Toyama 930-8555, Japan
[2]Tateyama Institute of System, Toyama 930-0004, Japan

Abstract: We present an improved method based on the Hopfield neural network for RNA secondary structure prediction in this study. The proposed method adjusts two parameters of the energy function in gradient ascent direction when the Hopfield neural network traps in a local minimum. The correction of the two parameters can increase the energy temporarily and help the network escape from the local minimum. The proposed algorithm was analyzed theoretically and evaluated experimentally through predicting RNA secondary structure. The simulation results on four RNA sequence show that the proposed algorithm performs better than others and has the ability to search the more stable RNA secondary structure for a RNA sequence.

Key words: RNA secondary structure, hopfield neural network, local minimum, gradient ascent learning

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). Many RNAs fold into structure that are important for regulatory, catalytic, or structure roles in the cell. RNA's structure is dominated by base pairing interactions, most of which are Watson-Crick pairs between complementary bases[1]. The base paired structure of RNA is called its secondary structure. Examples of such structures are found in tRNA[2], Rrna[3, 4], tmRNA[5] and SRP RNA[6].

Work on the determination of RNA secondary structure has been carried out for decades by a number of research groups. Appealing computational methods for the prediction of RNA structure from knowledge of primary structure have been developed to provide insight into functions that RNA serves. Searching for configurations of maximum base pairing or of minimum free energy is the best current 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[7]. The Zuker's algorithm (implemented in the programs called MFOLD[8]) is an efficient dynamic programming algorithm for identifying the globally minimal energy structure for a RNA sequence, as defined by such a thermodynamic model[9]. It computes an optimal structure for a sequence of N nucleotides in time proportional to N^3. Zuker's energy calculations have been further improved[10, 11] and are probably the most used RNA secondary structure prediction method today. But the Zuker's algorithm is incapable of predicting pseudoknots, and this is one well-known limitation of it.

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 a Hopfield neural network for RNA secondary prediction[12]. The results showed that this neural network algorithm performed better than the previous algorithms in the aspect of calculating time and accuracy. But the major weakness of this algorithm is still due to its failure in finding the global minimum solution.

In this study, we propose an improved method based on the Hopfield neural network for RNA secondary structure prediction. The proposed method adjusts two parameters of the energy function in gradient ascent direction when the Hopfield neural network traps in a local minimum. The proposed algorithm consists of two stages. In the first stage, a

Corresponding Author: YanQiu Che, Department of Intellectual Information Systems Engineering 3190 Gofuku, Toyama 930-8555, Japan

feasible solution is calculated by the Hopfield neural network. In the second, the parameters are changed and the network continues with the updated parameters. The proposed algorithm was applied to RNA secondary structure prediction. The simulation results on four RNA sequence showed that the proposed algorithm performed better than others and had the ability to search the more stable RNA secondary structure for a RNA sequence.

PROBLEM FORMULATION

For a given RNA sequence S, at the first all possible stack domain candidates are selected and listed. A set of adjacent base pairs is called stack domain, as showed in Fig.1. It is known that the longer the stack domain is, the more stable the RNA secondary structure is.

The stability of RNA secondary structure is evaluated by free energy. The most useful free energy data have been extrapolated from experiments on particular kinds of RNA carried out by Tinoco and Uhlenbeck[13][14]. For stack domain, the free energy is calculated according to Table 1(units is Kcal moL^{-1}).

For RNA sequence S, R = $R_1, R_2, R_3, \ldots, R_n$ is a set of stack domain candidates and $e_1, e_2, e_3, \ldots e_n$ are free energy values of these stack domain candidates 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, as shown in Fig.2.

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 following:

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

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

c_{ij} is a factor that indicates there is inconsistency or not. If both R_i and R_j are selected and there is inconsistency between them,then $c_{ij} = 1$; If both R_i and R_j are selected and there is no inconsistency between them, then $c_{ij} = 0$.

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

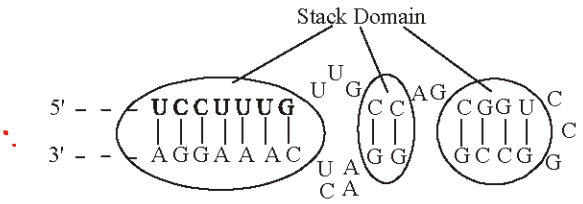

Fig. 1: Stack domain

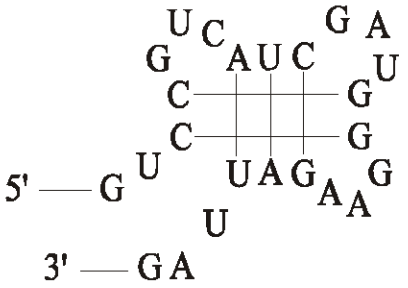

Fig.2: Inconsistency of stack domain candidates

DESCRIPTION OF THE PROPOSED METHOD

Takefuji's method for RNA secondary structure prediction: Takefuji proposed a Hopfield network algorithm for the prediction of RNA secondary structure[12]. In his algorithm, the number of stack domain candidates determines the number of neurons. The Hopfield network consists of *N* neurons, and each neuron has an input U_i and an output $V_i \in \{0,1\}$. The energy function of the Hopfield network is expressed in Eq. 1. And the inputs of the neurons are computed by the following Equation:

$$U_i = -A\left(e_i \sum_{k=1}^{N} c_{ik} V_k \right) + Be_i \qquad (3)$$

The output of the *i*th neuron is given by following:

$$V_i = \begin{cases} 1 & U_i \le 0 \\ 0 & U_i > 0 \end{cases} \qquad (4)$$

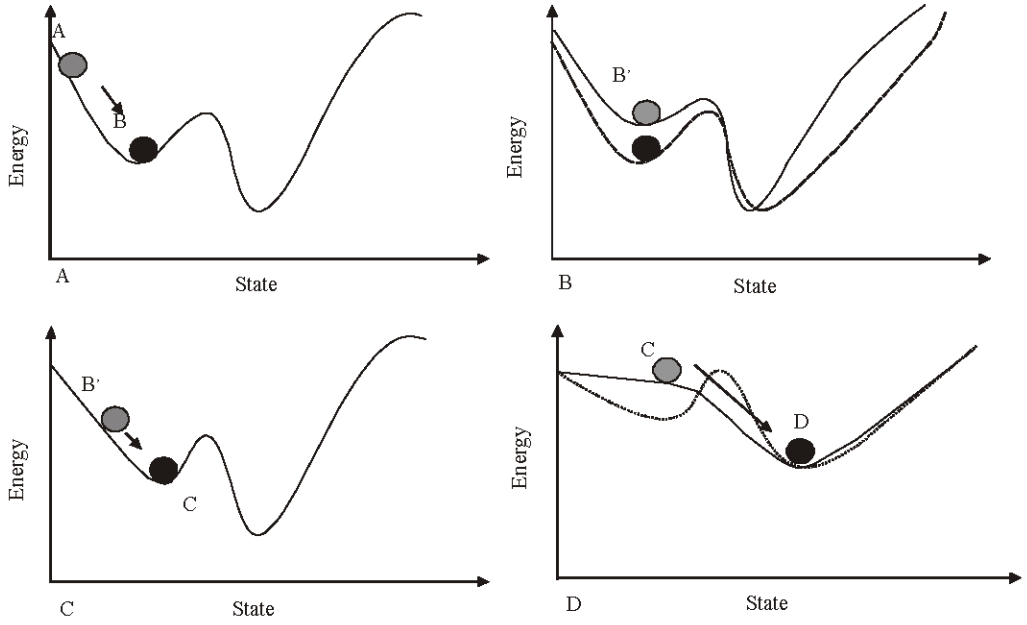

Fig.3: The relation between energy and state transition in the proposed algorithm with two stable states

where A and B are parameters, and 0<A<<B is set at the first. The value of parameter A increases little by little for every step to reduce the inconsistencies until there is no inconsistency in RNA secondary structure. The simulations results verify that the free energy is smaller than the previous algorithms and the more stable RNA secondary structure is obtained. However, the major weakness of this algorithm is still due to its failure in finding the global minimum solution.

Description of the proposed method: In order to overcome the disadvantage of the Takefuji's method and prediction the more stable RNA secondary structure, we propose an improved algorithm that permits energy to be increased temporarily, which helps the Hopfield neural network escape from local minimum. In the proposed algorithm, the energy function can be divided into two parts:

$$E = w_1 E_1 + w_2 E_2 \tag{5}$$

$$E_1 = \sum_{i=1}^{n} e_i V_i \tag{6}$$

$$E_2 = \sum_{i=1}^{n} \left(|e_i| \sum_{j=1}^{n} c_{ij} V_i V_j \right) \tag{7}$$

The motion Equation is described as follows:

$$U_i = -w_2 \left(e_i \sum_{k=1}^{N} c_{ik} V_k \right) + w_1 e_i \tag{8}$$

$$V_i = \begin{cases} 1 & U_i \le 0 \\ 0 & U_i > 0 \end{cases} \tag{9}$$

The parameters w_1 and w_2 are not constants in the proposed algorithm. When the network traps into a local minimum, the w_1 and w_2 will be updated by using positive gradient direction.

$$w_1 = w_1 + p \frac{\partial E}{\partial w_1} = w_1 + pE_1 \tag{10}$$

$$w_2 = w_2 + q \frac{\partial E}{\partial w_2} = w_2 + qE_2 \tag{11}$$

In order to explain the proposed algorithm, we use a conceptual graph of the energy landscape (Fig. 3) with a local minimum and a global minimum. 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.3 a). Because of the mechanism of the HNN, the state of network moves towards negative gradient direction and reaches the local minimum *B* (Fig. 3 b).This is called the HNN updating phase. Then we change parameters w_1 and w_2 in gradient ascent direction so that the energy is increased temporarily, and point *B* becomes a new point *B*' (Fig. 3 b). This is called the gradient ascent phase. From point *B*', the network returns to the HNN updating phase with new

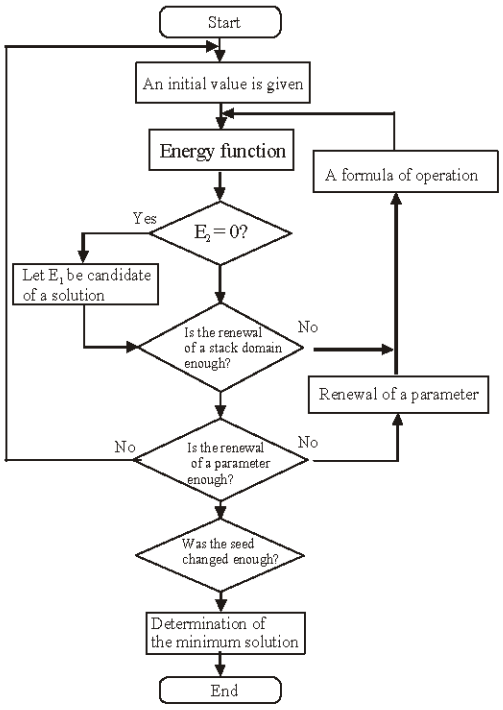

Fig. 4: The flow chart of the proposed algorithm

parameters w_1 and w_2 and reaches a new stable state C (Fig. 3c). If the HNN updating phase and the gradient ascent phase are carried out in turn, the network can escape from the local minimum and get the global minimum (Fig. 3d).

This technique can not always find the global minimum. Instead, it can get a lot of local minima and choose the smallest one as the solution. This technique has higher search capability than those which only converge at on local minimum.

Algorithm procedure: The proposed algorithm consists of two stages. In the first stage, a feasible solution is calculated by the Hopfield neural network. In the second, the parameters are changed and the network continues with the updated parameters. As a result, an optimal or near optimal solution can be obtained. The Fig. 4 describes the proposed method.

SIMULATION RESULT

In order to verify the effectiveness of the proposed algorithm for RNA secondary prediction, simulations has been carried out on four RNA sequences. All the simulations were implemented in C++ on PC (CPU:1.7GHz). Other algorithms were also executed for comparison. For Takefuji's algorithm, A and B were set 1 and 30, respectively and A was increased by 0.05

Table 2: Comparison of the value of 61 bases

Algorithm	Length = 2	Length = 3	Length = 4
Hopfield	-39.90(-20.18)	-39.90(-28.09)	-39.90(-34.60)
Takefuji	-30.30(-29.50)	-39.90(-38.48)	-39.90(-39.90)
proposed	-41.10(-39.95)	-39.90(-38.71)	-39.90(-39.90)

Table 3: Comparison of the value of 77 bases

Algorithm	Length = 2	Length = 3	Length = 4
Hopfield	-38.70(-18.17)	-40.80(-24.57)	-43.50(-29.92)
Takefuji	-46.20(-42.60)	-43.50(-38.56)	-43.50(-43.48)
proposed	-51.36(-46.75)	-49.80(-49.76)	-43.50(-43.48)

Table 4: Comparison of the value of 120 bases

Algorithm	Length = 2	Length = 3	Length = 4
Hopfield	-69.60(-24.84)	-74.70(-41.82)	-71.70(-52.85)
Takefuji	-76.50(-74.87)	-76.20(-73.96)	-65.10(-62.63)
proposed	-99.00(-93.10)	-88.50(-84.23)	-71.70(-71.54)

Table 5: Comparison of the value of 401 bases

Algorithm	Length = 3	Length = 4
Hopfield	-107.40(-33.57)	-103.50(-56.37)
Takefuji	-----	-173.10(-164.06)
proposed	-204.00(-195.54)	-178.20(-172.47)

when A needed to be updated. For the proposed algorithm, the values of parameters are initialized as follows:

$$w_1 = 1.0;\ w_2 = 1.0;\ p = 0.0001;\ q = 0.0001$$

For each of algorithms, the simulation program ran 100 times. The results that we recorded for each RNA sequence are the lowest energy and the average energy among 100 runs. The comparisons were arranged in Table 2~ 5. The columns Length = 2, Length = 3, Length = 4 represent the length of possible stack domain that starts from 2, 3, 4, respectively.

Example 1: 61 bases of RNA sequence
Consider the folding of a RNA sequence that has 61 bases:
5'---ACAGGAGUAAUUCCCGCCGAAACAGGGUUUU
CACCCUUCCUUUC
UUCGGGUGUCCUUCCUC ---3'

Example 2: 77 bases of RNA sequence
Consider the folding of a RNA sequence that has 77 bases:
5'---AGGCUUGUAGCUCAGGUGGUUAGAGCGCA
CCCCUGAUAAGGGUGA
GGUCGGUGGUUCAAGUCCACUCAGGCCUACCA---3'

Example 3: 120 bases of RNA sequence
Consider the folding of a RNA sequence that has 120 bases:
5'---UGCCUGGCGGCCUUAGCGCGGUGGUCC
CACCUGACCCCAUGCCGA
ACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAG
UGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCC
AGGCAU---3'

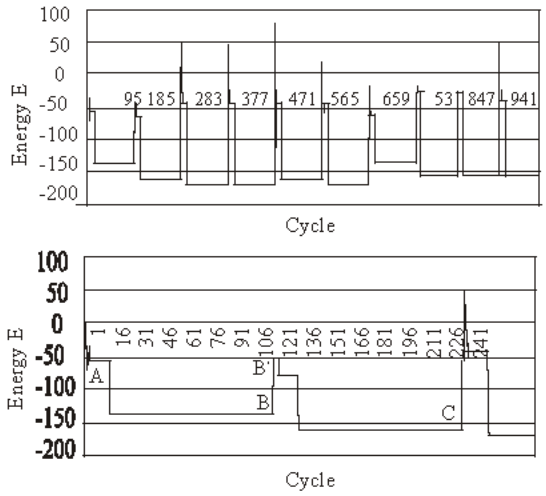

Fig. 5: The variation process of energy as parameters change

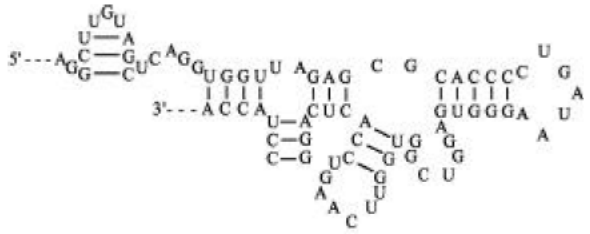

Fig. 6: The secondary structure of example 2:77 bases

Example 4: 401 bases of RNA sequence

Consider the folding of a RNA sequence that has 401 bases:

5'---AUGUGCAAUUUUUGGAUAAUCGCGUGA
GGAGAAUUGUUUCUCAU
GAGGAAAGUCCAUGCUAGCACAGGCUGUGAUGCC
UGUAGUGUUUGUGCUAGGCGAAACCAUAAGCCU
AGGGACGAGAAAUCGUUACGGCAGUUGAAAUGG
CUAAGUCCUUGGAUAGGCCAGAGUAGGCUUGAA
AGUGCCACAGUGACGGAGUCUUUCUGGAAACAGA
GAGAGUGGAACGCGGUAAACCCCUCAAGCUAGCA
ACCCAAAUUUUGGUCGGGGCAUGGAGUACGCGGA
AACGAACGUAGUAUUCUGACUGCUAUCAGCUAG
AGCUGUUAGUGGUAGACAGAUGAUUAUCGAAGG
AAGUGGUCCUAGUCACUUCUGGAACAAAACAUG
GCUUAUAGAAAAUUGCAUAUAGG---3'

From the comparisons from Table 2 to Table 5, we can see that the proposed algorithm outperformed the other algorithms with respect to the lowest energy. That is to say that the proposed algorithm has the ability to escape from the local minima and find the more stable RNA secondary structure. Moreover, we can find that for large problems, for example, 120 and 401 bases, the proposed algorithm was more effective than Takefuji's algorithm. We can draw a conclusion that the longer RNA sequence is, the more effective the proposed algorithm is.

In order to show how the proposed algorithm works, we chose RNA sequence with 401 bases as an example. Fig. 5 shows the evolution of energy function in the proposed algorithm. Fig. 5b is an enlarged figure of Fig. 5a. From Fig. 5, we can see the network started from the state A where energy was -30.30 and then trapped into local minimum B:-135.60. In order to escape from this local minimum, the parameters were updated by positive gradient direction and the state of the network was changed from B to B':-39.60. And then form the new state B' the network continued to calculate until it served another local minimum C:-160.20. The parameters were changed again to escape from C. The two stages repeated by turns and at last the optimal or near optimal solution could be reached.

The RNA secondary structure will not be understood if only the value of energy is list. Fig. 6 shows the prediction result of example 2: 77 bases.

CONCLUSIONS

This study presents an improved method based on the Hopfield neural network for RNA secondary structure prediction. The proposed method adjusts two parameters of the energy function in gradient ascent direction when the Hopfield neural network traps in a local minimum. The correction of the two parameters can increase the energy temporarily and help the network escape from the local minimum. The proposed algorithm was analyzed theoretically and evaluated experimentally through predicting RNA secondary structure. The simulation results on four RNA sequence showed that the proposed algorithm performed better than others and had the ability to search the more stable RNA secondary structure for a RNA sequence.

REFERENCES

1. Watson, J.D. and F.H.C. Crick, 1953. A Structure for Deoxyribose Nucleic Acid. Molecular Structure of Nucleic Acids. Nature, 171: 737-738.
2. Sprnzl, M., C. Horn, M. Brown, A. Ioudovitch and S. Steinberg, 1998. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res., 26: 148-153.

3. Wuyts, J., P. De Rijk, Y. Van de Peer, T. Winkelmans and De R. Wachter, 2001. The European large sununit ribosomal RNA database. Nucleic Acids Res., 29: 175-177.
4. Wuyts, J., Y. Van de Peer, T. Winkelmans and De R. Wachter, 2002. The European database on small subunit ribosomal RNA. Nucleic Acids Res., 30: 183-185.
5. Zwieb, C., J. Gorodkin, B. Knudsen, J. Burks and J. Wower, 2003. tmRDB(tmRNA database). Nucleic Acids Res., 32: 446-447.
6. Rosenblad, M.A., J. Gorodkin, B. Knudsen, C. Zwieb and T. Samuelsson, 2003. Nucleic Acids Res., 31: 363-364.
7. Zuker, M. and P. Stiegler, 1981. Optimal computer folding of lager RNA sequences using thermodynamic and auxiliary information. Nucleic Acids Res., 9: 133-148.
8. Zuker, M., 1989. Compute prediction of RNA structure. Meth. Enzymol., 180: 262-288.
9. Zuker, M., D.H. Mathews and D.H. Turner, 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. RNA Biochemistry and Biotechnology, pp: 11-43.
10. Walter, A.E., D.H. Turner, J. Kim, M.H. Lyttle, P. Müller, D.H. Mathews and M. Zuker, 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc. Natl. Acad. Sci., 91: 9218-9222.
11. Mathews, D., H.J. Sabina, M. Zuker and D.H. Turner, 1999. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J. Mol. Biol., 288: 911-940.
12. Takefuji, Y., D. Ben-Alon and A. Zaritsky, 1992. Neural computing in discovering RNA interactions. BioSystems, 27: 85-96.
13. Tinoco, I., O.C. Uhlenbeck and M.D. Levine, 1971. Estimation of Secondary Structure in Ribonucleic Acids. Nature, 230: 362-367.
14. Tinoco, I., P.N. Borer, B. Dengler, M.D. Levine, O.C. Uhlenbeck, D.M. Crothers and J. Gralla, 1973. Improved Estimation of Secondary Structure in Ribonucleic Acids. Nature New Biol., 246: 40-41.