Biol. Cybern. 63, 337–340 (1990)

Biological Cybernetics

A Parallel Algorithm for Estimating the Secondary Structure in Ribonucleic Acids

Y. Takefuji [1], C. W. Lin [2], and K. C. Lee [1]

[1] Department of Electrical Engineering and Applied Physics and [2] Department of Biomedical Engineering, Center for Automation and Intelligent Systems Research, Case Western Reserve University, Cleveland, OH 44106, USA

Received October 25, 1989/Accepted in revised form April 30, 1990

Abstract. A parallel algorithm for estimating the secondary structure of an RNA molecule is presented in this paper. The mathematical problem to compute an optimal folding based on free-energy minimization is mapped onto a graph planarization problem. In the planarization problem we want to maximize the number of edges in a plane with no two edges crossing each other. To solve a sequence of n bases, $n(n-1)/2$ processing elements are used in our algorithm.

Introduction

Fresco has used the first RNA secondary structure model for predicting the secondary structure in Ribonucleic Acids (Fresco and Albert 1960). Two types of RNA folding algorithms have been reported: the "combinatorial" method introduced by Pipas (Pipas and McMahon 1975) and the "recursive" or dynamic programming method introduced by Nussinov (Nussinov et al. 1978). Both algorithms including the latest method proposed by Zuker (Zuker 1989) are all based on the sequential computation. Unfortunately few parallel algorithms based on molecular thermodynamics models have been reported. Recently Qian and Sejnowski (Qian and Sejnowski 1988), and Holley and Karplus (Holley and Karplus 1989) have reported a backpropagation algorithm using a three-layer feed-forward neural network for protein secondary structure prediction. Their method is based on the correlation between secondary structure and amino acid sequences. They have the following drawbacks over the conventional RNA folding algorithms based on molecular thermodynamics models.

1. They need a teacher to let the network to learn the correlation between secondary structure and amino acid sequences. The molecular thermodynamics models do not need the teacher.

2. The correlation models cannot provide accurate prediction if a completely new datum is given where the previously learned correlation is useless.

3. Their feed-forward neural network requires a prohibitively long learning process to deal with a long sequence of bases for the RNA secondary structure prediction.

4. No theorem is given to determine the neural network architecture including how many layers and how many hidden neurons should be used.

A suboptimal parallel algorithm for estimating the secondary structure of the RNA is introduced in this paper. Our algorithm is based on the molecular thermodynamice models. It does not require a teacher, nor a learning process. The algorithm using $n(n-1)/2$ processors can yield the suboptimum solution where n is the number of bases. A sequence of fifty-five bases from R17 viral RNA (Tinoco et al. 1971) was used to verify our algorithm.

Parallel algorithms based on artificial neural network model have been successfully applied to solving NP-complete optimization problems such as graph planarization (Takefuji and Lee 1989), four-coloring (Takefuji and Lee 1988), tiling (Takefuji and Lee 1990), sorting (Takefuji and Lee 1990). Our algorithm uses a massive number of simple processing elements. The processing element is called the neuron, because it performs the function of a simplified biological neuron. In our algorithm the binary neurons play a key role where the output of the ith neuron V_i follows:

$$V_i = 1 \quad \text{if} \quad U_i > 0$$

0 otherwise, where U_i is the input of the ith neuron.

The stability number for a given RNA secondary structure is the sum of the contributions of the loops, bulges, and helices. The structure with the highest number is the most stable, called optimal folding. The mathematical problem to compute an optimal folding based on free-energy minimization is mapped onto a graph planarization problem. In the planarization problem we want to maximize the number of edges in a plane with no two edges crossing each other. An *A-U* or *G-C* base pairs are only considered as possible edges to be embedded in a plane while the bases are the

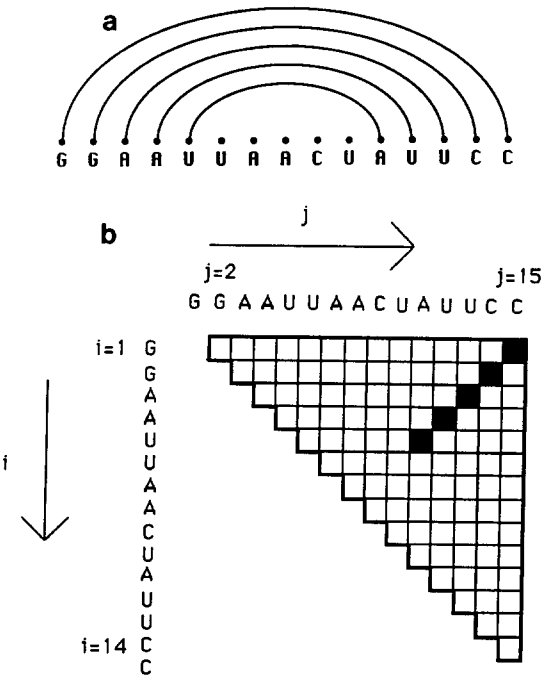

Fig. 1. a A single-row representation. **b** A neural network array for a sequence of fifteen bases

vertices. In other words, for a fragment stretching from ribonucleotides i to j, it is denoted by the subscript ijth neuron where the output and the input is depicted by V_{ij} and U_{ij} respectively for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$.

Consider a sequence of fifteen bases as shown in Fig. 1a. In our algorithm a single-row representation is used were five edges are embedded. A $15 \times 14/2$ neural network array in Fig. 1b is used to predict the secondary structure of this problem. The following five functions are considered: $b(i,j)$, $g(i,j,k)$, $f(i,j)$, $p(i,j,t)$, and $h(x)$. The function $b(i,j)$ denotes the possible pairing: $b(i,j) = 1$ if i and j bases are one of four base pairs (G-C, C-G, A-U, or U-A base pairs), 0 otherwise. The cross bonding is sterically impossible so that the graph must be planar where two edges must not cross each other. A violation function $g(i,j,k)$ for graph planarization was described in our paper (Takefuji and Lee 1989): $g(i,j,k) = 1$ if $i < j < k$, 0 otherwise. The function $f(k,l)$ indicates the strength of a base pair bond between k and 1 bases: $f(k,l) = 2$ if k and 1 bases are a G-C pair, 1 if they are an A-U pair, 0 otherwise. This strength of a base pair bond is based on the result of Tinoco (Tinoco et al. 1971). The hairpin loop constraint is also considered in our algorithm where more than two bases are required to make a hairpin loop. The hairpin constraint function $p(i,j,t)$ is given by: $p(i,j,t) = 1$ if (j-hairpin) $< i$, 0 otherwise where hairpin is given by hairpin $= 4$ if $t = 0$, 55-t if 55-$t > 4$, 4 otherwise, $h(x)$ is the hill-climbing function, 1 if $x = 0$, 0 otherwise.

To predict suboptimal foldings of a sequence of n bases, $n(n-1)/2$ neurons are required. The motion equation of the ijth neuron is given by:

$$\begin{aligned}\frac{dU_{ij}}{dt} = &-A_1\left(\sum_{k\neq i} V_{ik} - 1\right) b(i,j) \\ &-A_2\left(\sum_{k\neq j} V_{jk} - 1\right) b(i,j) \\ &-B_1 \sum_{k<i<l<j} V_{kl} g(k,i,l) g(i,l,j) f(k,l) \\ &-B_2 \sum_{i<k<j<l} V_{kl} g(i,k,j) g(k,j,l) f(k,l) \\ &-Cp(i,j,t) + Dh\left(\sum_{k\neq i} V_{ik}\right) \qquad (1)\end{aligned}$$

where V_{xy} is V_{xy} if $x < y$, V_{xy} otherwise. The first term in (1) forces the ith base to have one and only one bond. Note that if the ith base has strong violations caused by other bases, it cannot have any bond. The second term also forces the jth base to have one and only one bond. The third and fourth terms are always inhibitory which always satisfy planarization conditions. The fifth term is the inhibitory hairpin constraint which prohibits less than three bases to make a hairpin loop. The last term is a hill-climbing force which allows the state of the system to escape from the local minimum.

Parallel Algorithm

The following procedure describes the proposed parallel algorithm based on the first order Euler method. It yields the suboptimal secondary structure of the RNA from the given sequence of bases. 0. Set $t = 0$, $A_1 = A_2 = B_1 = B_2 = D = 1$, and $C = 3$.

1. The small negative number is assigned to the initial values of $U_{ij}(t)$ for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$.

2. Evaluate values of V_{ij} based on the binary function for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$.
$V_{ij}(t) = 1$ if $U_{ij} > 0$
0 otherwise

3. Use the motion equation in (1) to compute $\Delta U_{ij}(t)$.

$$\begin{aligned}\Delta U_{ij}(t) = &-A_1\left(\sum_{k\neq i} V_{ik}(t) - 1\right) b(i,j) \\ &-A_2\left(\sum_{k\neq j} V_{jk}(t) - 1\right) b(i,j) \\ &-B_1 \sum_{k<i<l<j} V_{kl}(t) g(k,i,l) g(i,l,j) f(k,l) \\ &-B_2 \sum_{i<k<j<1} V_{kl}(t) g(i,k,j) g(k,j,l) f(k,l) \\ &-Cp(i,j,t) \\ &+Dh\left(\sum_{k\neq i} V_{ik}(t)\right)\end{aligned}$$

4. Compute $U_{ij}(t+1)$ based on the first order Euler method: $U_{ij}(t+1) = U_{ij}(t) + \Delta U_{ij}(t)$ where $i = 1$, $n-1$ and $j = i+1, \ldots, n$.

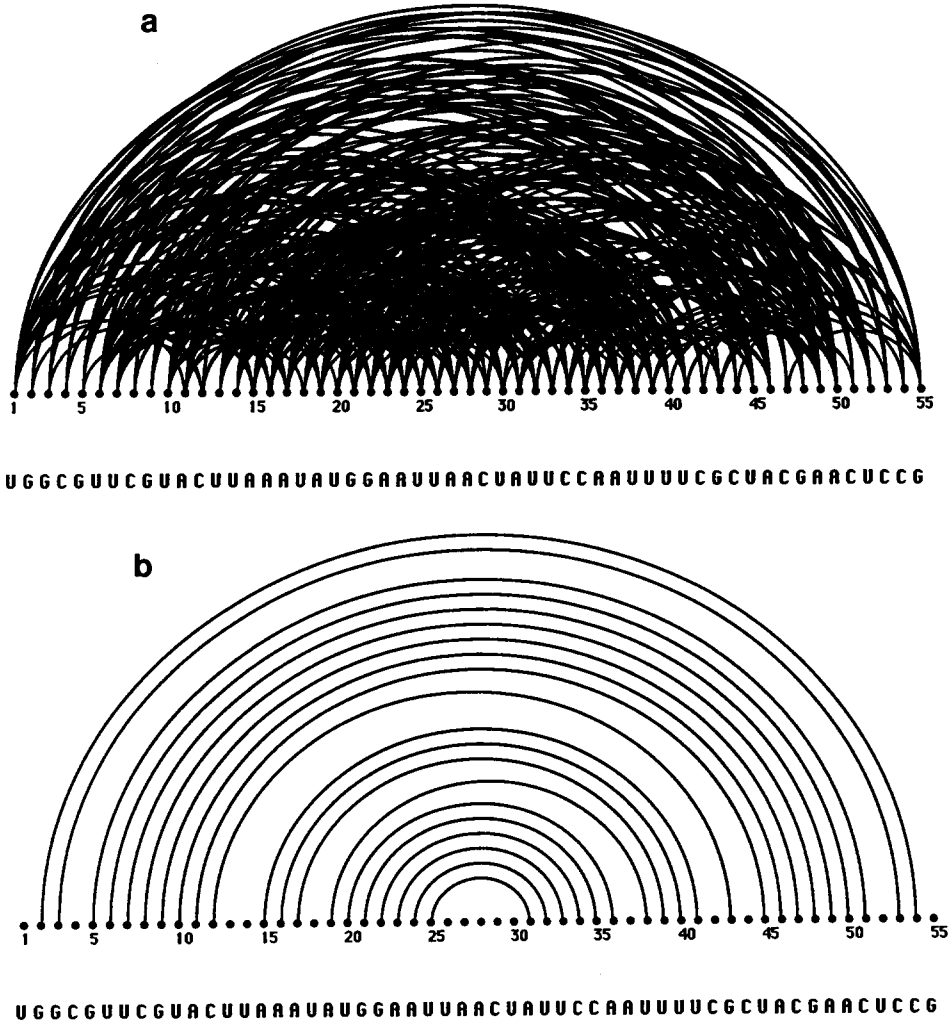

Fig. 2. a The convergence of the neural network to a solution. This shows the state of 1485 neurons after the first iteration. **b** The convergence of the neural network to a solution. This shows the state of 1485 neurons after the sixty-first iteration

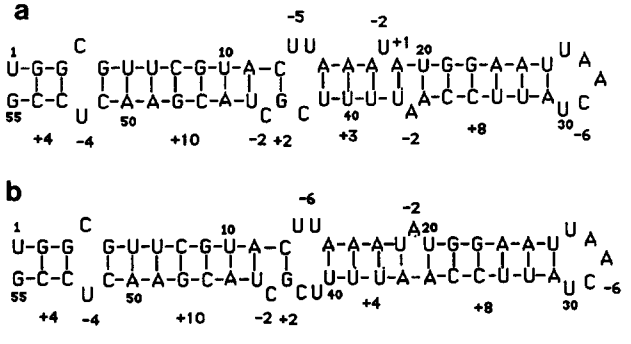

Fig. 3. a The predicted secondary structure. **b** The optimum structure

5. Increment t by 1. If $t = T$ then terminate this procedure else go to step 2.

Consider a sequence of fifty-five bases from *R*17 viral RNA (Tinoco et al. 1971). 1485 ($=55 \times 54/2$) neurons are used to solve this problem. Figure 2a and b shows the state of the system after the first iteration and the sixty-first iteration respectively. Figure 3a redraws the result of the secondary structure of the sequence shown in Fig. 2b. The total stability of the structure is $+7$ which is equivalent to the free energy $\Delta G = -8.4$ kcal. When pairs (A^{15} to U^{41}), (A^{16} to U^{40}), and (A^{17} to U^{39}), are shifted to (A^{15} to U^{40}), (A^{16} to U^{39}), and (A^{17} to U^{38}) respectively, the total stability of the structure becomes $+8$ which is the optimum structure (Tinoco et al. 1971) in Fig. 3b.

Conclusion

We showed the parallel algorithm for estimating the secondary structure in Ribonucleic Acids. The proposed algorthm requires $n(n-1)/2$ processing elements for a chain length of n bases to compute suboptimal secondary structure of RNA molecules. As of August 1987, the Genbank database had approximately 15,000 entries with nearly 15 million nucleotides. The average length of a sequence is approximately 1000 nucleotides and the longest is 172,282. In order to minimize the number of processing elements for the long sequence, we have developed the constraint of the possible base pairs (i and j): $n - \alpha < i + j < n + \beta$ where α and β are positive integer. For example, in a sequence of 359 bases from the potato sprindle tuber viroid (PSTV), the following condition $350 < i + j < 370$ is given for the possible base pairs (i and j) constraint. Consequently it

requires only 1017 processing elements for the sequence of 359 bases instead of 128,881 processing elements. Our system generated the suboptimal secondary structure in a sequence of 359 bases from the PSTV.

Acknowledgements. This work is partly supported by the National Science Foundation Grant MIP-8902819.

References

Fresco JR, Alberts BM, Doty P (1960) Some molecular details of the secondary structure of rinobucleic acid. Nature 188:98

Holley H, Karplus M (1989) Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 86

Nussinov R, Pieczenik G, Griggs JR, Klietman DJ (1978) Algorithms for loop matchings. SIAM J Appl Math 35:68

Pipas JM, McMahon JE (1975) Method for predicting RNA secondary structure. Proc Natl Acad Sci USA 72:2017

Qian N, Sejnowski T (1988) Predicting the secondary structure of globular protein using neural network models. J Mol Biol 202:865

Takefuji Y, Lee KC (1988) CAISR Tech. Rep. TR 88–139. Case Western Reserve University

Takefuji Y, Lee KC (1989) A near-optimum parallel planarization algorithm. Science 245:1221

Takefuji Y, Lee KC (1990) A parallel algorithm for tiling problems. IEEE Trans Neural Networks 1:1

Takefuji Y, Lee KC (1990) A superior sorting parallel algorithm based on neural networks. IEEE Trans Circ Syst 37:8

Tinoco I, Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in Ribonucleic Acids. Nature 230

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48

Y. Takefuji
Department of Electrical Engineering
and Applied Physics
Case Western Reserve University
Glennan Building
Cleveland, OH 44106
USA