

# New Method for Predicting RNA Secondary Structure

Hirotoishi Taira<sup>†</sup>, Tomonori Izumitani<sup>†</sup>, Takeshi Suzuki<sup>‡</sup> and Eisaku Maeda<sup>†</sup>

<sup>†</sup>NTT Communication Science Laboratories  
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan  
{taira, izumi, maeda}@cslab.kecl.ntt.co.jp

<sup>‡</sup>Department of Electrical Engineering, Nagaoka University of Technology  
1603-1 Kamitomioka-machi, Nagaoka Niigata 940-2188, Japan  
takesu@pelican.nagaokaut.ac.jp

**Keywords:** RNA secondary structure prediction, Nussinov algorithm, SCFG, F-measure

## Abstract

It has become clear recently that there are many RNAs that are not translated into proteins, instead they work as functional molecules. These RNAs are called “non-coding RNAs.” Predicting the secondary structure of these RNAs is important for understanding their functions. We will therefore focus on Nussinov’s algorithm and the SCFG version of Nussinov’s algorithm as useful techniques for predicting RNA secondary structures. We introduce a new scoring table and loop length restriction to improve these algorithms. and the improved algorithms provided better levels of performance than the originals.

## 1 Introduction

Recently, many DNA gene sequences have been discovered, as genome projects including the human genome project, study various types of organism. According to the central dogma, each DNA gene sequence is copied to a sequence of mRNAs by “transfer”, and subsequently, the sequence is “translated” into a protein. By contrast, there are RNAs that are separated from the dogma, and are not translated into proteins. Instead, these RNAs themselves work

as functional molecules. This RNA is called “non-coding RNA” and there are various kinds. They include tRNA, which is used to transfer the gene information contained in mRNA into the amino acid sequences constituting a protein and rRNA, one of the subunits of a ribosome. Non-coding RNAs occupy almost all parts of RNA and have important roles within living bodies. If we identify the structure correctly, the function of the RNA can be predicted more easily and will be more useful in the development of new medicines. This makes it very important to get to know these solid structures. However, in order to identify the exact three-dimensional structure of molecules, the molecules must be crystallized and this process is difficult and time-consuming. Therefore, it is useful if the three-dimensional structure can be predicted from a one-dimensional sequence of RNA by calculation. Since it is difficult to predict the three-dimensional structure of RNA correctly with the present technology, we tried instead to predict two-dimensional structure of RNA.

There are two main conventional algorithms for the prediction of secondary structures. They are Nussinov’s algorithm (Nussinov et al., 1978) and Zuker’s algorithm (Zuker, 1989). In this paper, we focus on Nussinov’s algorithm. This algorithm utilizes dynamic programming to search for remote base pairs. We add a new scoring mechanism and control the loop’s min-

imum length to obtain high levels of performance when predicting secondary structures.

In the next section, we describe the Nussinov algorithm. In the third section, we present the Nussinov algorithm using stochastic context-free grammars (SCFGs) (Eddy and Durbin, 1994; Grate, 1995; Lefebvre, 1995; Lefebvre, 1996). In the fourth section, we present our new scoring technique and algorithm. In the fifth section, we describe experiments using this algorithm and discuss the results. In the last section, we draw conclusions based on these results.

## 2 Nussinov's Algorithm

Nussinov's algorithm is a simple algorithm that finds the RNA secondary structure with the highest number of base pairs using dynamic programming (DP) (Nussinov et al., 1978; Durbin et al., 1998). This algorithm can be described as follows:

Initialize:

$$\begin{aligned}\gamma(i, i-1) &= 0 \quad (i = 1, 2, \dots, L) \\ \gamma(i, i) &= 0 \quad (i = 2, 3, \dots, L)\end{aligned}$$

Repeat:  $for(l = 2; l \leq L; l++)\{$   
 $for(i = 1; i \leq L - l + 1; i++)\{$

$$\gamma(i, j) = \max \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + \delta(i, j) \\ \max_{i < k < j} [\gamma(i, k) \\ + \gamma(k+1, j)] \end{cases}$$

Here,  $i$  and  $j$  indicate the one-dimensional location of an RNA sequence.  $L$  indicates the length of the sequence.  $\delta(i, j)$  indicates whether there is a base pair between the  $i$ -th base and the  $j$ -th base (1) or not (0).  $\gamma(i, j)$  indicates the number of base pairs between  $i$  and  $j$ . Hence, the value of  $\gamma(1, L)$  is the number of base pairs of an optimal structure.

The optimum secondary structures can then be predicted by performing a traceback.

## 3 Nussinov's Algorithms using SCFG

Next, we present the SCFG version of Nussinov's algorithm. In this algorithm, the procedure of the regular Nussinov's algorithm is replaced by an SCFG generation rule (Durbin et al., 1998). By using SCFG, we can give the algorithm the probabilistic framework.

$$S \rightarrow aS|cS|gS|uS \quad (1)$$

$$S \rightarrow Sa|Sc|Sg|Su \quad (2)$$

$$S \rightarrow aSu|uSa|cSg|gSc \quad (3)$$

$$S \rightarrow SS \quad (4)$$

$$S \rightarrow \epsilon \quad (5)$$

Rules (1) and (2) correspond to the generation of a non-base pair. (3) corresponds to the generation of a regular base pair, (4) to the generation of a branch, and (5) to the end of generation process, respectively. The process using this grammar works as follows:

Initialize:

$$\begin{aligned}& \text{set initial values} \\ & \text{to } p(x_i S), p(Sx_j), p(x_i Sx_j), p(SS) \\ & \gamma(i, i-1) = -\infty \quad (i = 2, 3, \dots, L) \\ & \gamma(i, i) = \max\{\log p(x_i S), \log p(Sx_i)\} \\ & \quad (i = 1, 2, \dots, L)\end{aligned}$$

Repeat:

$$\begin{aligned}& for(l = 2; l \leq L; l++)\{ \\ & for(i = 1; i \leq L - l + 1; i++)\{ \\ & j = i + l - 1 \\ & \gamma(i, j) = \\ & \quad \max \begin{cases} \gamma(i+1, j) + \log p(x_i S) \\ \gamma(i, j-1) + \log p(Sx_j) \\ \gamma(i+1, j-1) + \log p(x_i Sx_j) \\ \max_{i < k < j} \{ \gamma(i, k) \\ + \gamma(k+1, j) + \log p(SS) \} \end{cases} \\ & \quad \} \\ & \quad \}\end{aligned}$$

Here,  $x_i$  and  $x_j$  are the  $i$ -th and the  $j$ -th bases in the sequence, respectively. We can then obtain the predicted result of an optimal secondary structure by performing a traceback.

## 4 Our Methods

The Nussinov Algorithm and Nussinov Algorithm using SCFG are not without problems. One problem is that the algorithm only considers the maximum number of base pairs when searching for an optimum structure. Hence, even if the predicted loops are short, the algorithm tends to make a base pair. In the real world, since short loops are often thermodynamically unstable structures, we will obtain many incorrect structures. To compensate for these short loops, we ensured that the loop length was six or more.

Another problem with the Nussinov algorithm is that it only takes regular base pairs into consideration. Figures 1 and 2 show models of regular base pairs. The hydrogen bonds between Adenine and Uracil, and between Guanine and Cytosine, are drawn as dotted lines. By contrast, the pair of Guanine and Uracil pair has two hydrogen bonds and there is rebounding between an oxygen and an oxygen as shown in Fig. 3.

The scoring table used by the Nussinov algorithm is shown in Fig 4. In this table, the score given to a regular base pair is 1. Considering the above discussion, we propose a new scoring table in Fig. 5. In proportion to the number of hydrogen bonds, we give 2 and 3 for the regular base pairs, respectively, and 1 for the G-C non-regular base pair considering the oxygen-oxygen rebounding, and -1 to the other combinations. Based on these restrictions and the scoring table, we predicted the secondary structure of RNA with the Nussinov algorithm and the Nussinov algorithm using SCFG.

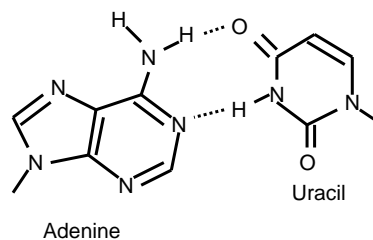


Figure 1: A-U regular base pair.

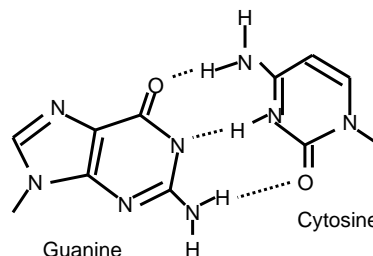


Figure 2: G-C regular base pair.

### 4.1 Experimental Result

#### 4.1.1 Experimental Setting

We used 20 RNA sequences taken from various web sites. Their lengths were 21-38 bases. These sequences are shown in Fig. 6.

The actual secondary structure of these RNA sequences is shown in the “RE” columns. In this figure, it is shown that “>” forms a base pair with “<”, and “-” shows this position is in the loop domain. “>” expresses one side of the new base pair, and “<” shows the base that has not yet made the pair in “>” which appeared most recently.

For the experiment, we used the original Nussinov (NA) and improved Nussinov (NB) algorithm, as well as the Nussinov algorithm using SCFG (SA), and the improved Nussinov algorithm using SCFG (SB).

#### 4.1.2 Evaluation Methods

The algorithm was evaluated by accuracy, F-measure, and evaluation by hand.

Accuracy is the prediction rate showing whether the position is a base pair or a loop.

Table 1: Accuracy vs F-measure vs Evaluation by hand

Seq No.	NA	NB	lp	gu	SA	SB	lp	gu
1	0.346 / 0.166 / 1	0.692 / 0.555 / 2	O		0.769 / 0.625 / 2	1.000 / 1.000 / 5	O	O
2	0.862 / 0.666 / 2	0.896 / 0.727 / 3	O	O	0.896 / 0.727 / 3	1.000 / 1.000 / 5		O
3	0.517 / 0.375 / 1	0.724 / 0.636 / 2	O		0.655 / 0.555 / 1	0.620 / 0.500 / 1		
4	0.761 / 0.777 / 1	0.714 / 0.666 / 3			0.904 / 0.888 / 4	0.714 / 0.666 / 3		
5	0.473 / 0.000 / 1	0.894 / 0.833 / 4	O		1.000 / 1.000 / 5	1.000 / 1.000 / 5		
6	0.685 / 0.375 / 1	0.885 / 0.777 / 3	O	O	0.628 / 0.333 / 1	0.885 / 0.777 / 4	O	O
7	0.428 / 0.272 / 1	0.714 / 0.642 / 1	O		0.657 / 0.538 / 1	0.657 / 0.657 / 1		
8	0.454 / 0.428 / 1	0.696 / 0.666 / 1	O		0.878 / 0.875 / 1	0.818 / 0.785 / 3		
9	0.538 / 0.000 / 1	0.794 / 0.428 / 2	O	O	0.538 / 0.100 / 1	0.794 / 0.333 / 1		O
10	0.380 / 0.142 / 1	0.619 / 0.500 / 2	O	O	0.428 / 0.444 / 1	0.619 / 0.500 / 2	O	O
11	0.538 / 0.200 / 1	1.000 / 1.000 / 5	O	O	0.692 / 0.428 / 1	1.000 / 1.000 / 5	O	O
12	0.461 / 0.250 / 1	0.923 / 0.900 / 3	O	O	0.538 / 0.333 / 1	0.923 / 0.888 / 2		O
13	0.615 / 0.333 / 1	1.000 / 1.000 / 5	O	O	0.769 / 0.625 / 2	0.923 / 0.875 / 2	O	O
14	0.692 / 0.400 / 1	1.000 / 1.000 / 5	O	O	0.692 / 0.333 / 1	1.000 / 1.000 / 5	O	O
15	0.653 / 0.428 / 1	1.000 / 1.000 / 5	O	O	0.692 / 0.500 / 1	0.692 / 0.500 / 2	O	O
16	0.375 / 0.250 / 2	0.575 / 0.588 / 1			0.525 / 0.562 / 2	0.450 / 0.400 / 2		
17	0.629 / 0.600 / 1	0.629 / 0.500 / 1			0.518 / 0.333 / 1	0.518 / 0.333 / 1		
18	0.523 / 0.400 / 1	0.904 / 0.875 / 4	O		1.000 / 1.000 / 5	1.000 / 1.000 / 5		
19	0.208 / 0.111 / 1	0.291 / 0.384 / 1	O		0.750 / 0.750 / 2	0.833 / 0.833 / 3		
20	0.416 / 0.500 / 1	0.500 / 0.545 / 1			0.500 / 0.600 / 1	0.333 / 0.363 / 1		
avg.	0.527 / 0.333	0.772 / 0.711			0.701 / 0.577	0.787 / 0.714		

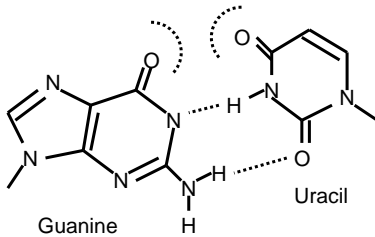


Figure 3: G-U non-regular base pair.

The F-measure is defined as follows.

For every prediction of base, we can calculate  
 $a$  = (the number of bases the predictor evaluates as loop for actual loop ),

$b$  = (the number of bases the predictor evaluates as loop for actual base pair),

$c$  = (the number of bases the predictor evaluates as base pair for actual loop).

Then, we can calculate the precision ( $P$ ) and recall ( $R$ ) as

$$P = \frac{a}{a+b}, \quad R = \frac{a}{a+c}.$$

By combining the precision and recall, the F-measure is defined as follows:

$$F = \frac{1 + \beta^2}{\frac{1}{P} + \beta^2 \frac{1}{R}}.$$

	A	C	G	U
A	0	0	0	1
C	0	0	1	0
G	0	1	0	0
U	1	0	0	0

Figure 4: Original scoring table.

The F-measure varies between 0 and 1. As the F-measure becomes larger, the prediction accuracy increases.  $\beta$  is a weight parameter and we set  $\beta = 1$ .

The evaluation by hand assigns five levels to the result structures.

- 5) Perfect match
- 4) Having one mistake related to the loop or bulge
- 3) Having two mistakes related to the loop or bulge
- 2) The number of hairpin loops is the same as

	A	C	G	U
A	-1	-1	-1	2
C	-1	-1	3	-1
G	-1	3	-1	1
U	2	-1	1	-1

Figure 5: Our scoring table.

the actual structure

1) No match

### 4.1.3 Experimental Results

We compared results obtained with Nussinov (NA), our own New Nussinov (NB), SCFG (SA), our new SCFG (SB).

Concrete predicted structures are shown in Fig. 6. The result is shown in Table 1. On average, the F-measure rises from 0.333 to 0.711 due to the improvements from NA to NB. Moreover, the F-measure rises from 0.577 to 0.714 due to the improvements from SA to SB. Accuracy evaluation shows similar results. In Table 1, Os in “lp” and “gu” columns indicate sequences that obtain higher F-measure with loop restriction and considering G-U pairs, respectively. The F-measure of 11 in 18 sequences increased with loop restriction, and the F-measure 9 sequence increased with considering G-U pair, due to the improvements from NA to NB. Moreover, the F-measure of 6 in 11 sequences increased with loop restriction and the F-measure of 9 sequences increased with considering G-U pairs, due to the improvements from SA to SB.

The result of hand evaluation is shown in Fig. 7. It turns out that the NB and SB results improve greatly here too. Especially, the number of 5) Perfect Match sequences increase

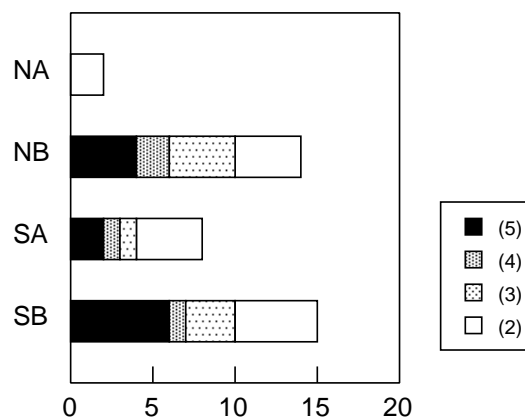


Figure 7: Total Evaluation by hand.

abruptly. This indicates that our method outperforms the original Nussinov algorithm and Nussinov algorithm using SCFG.

## 5 Conclusion

We presented a new Nussinov algorithm and a scoring table for the prediction of RNA secondary structure. Our experimental results indicate that this scoring approach and method work well. In the future, we would like to predict various RNA sequences.

## References

- R. Durbin, S. R. Eddy, A. Krogh, and G. J. Mitchison (Eds.). 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
- S. R. Eddy and R. Durbin. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088.
- L. Grate. 1995. Automatic RNA secondary structure determination with stochastic context-free grammars. In *Proc. of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 136–144. AAAI Press.
- F. Lefebvre. 1995. An optimized parsing algorithm well suited to RNA folding. In *Proc. of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 222–230. AAAI Press.
- F. Lefebvre. 1996. A grammar-based unification of several alignment and folding algorithms. In *Proc.*

*of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 143–154. AAAI Press.

R. Nussinov, G. Pieczenk, J. R. Griggs, and D. J. Kleitman. 1978. Algorithms for loop matching. *SIAM Journal of Applied Mathematics*, 35:68–82.

M. Zuker. 1989. Computer prediction of RNA structure. *Methods in Enzymology*, 180:262–288.

