

A New Method for Predicting RNA Secondary Structure

Hirotoishi Taira, Tomonori Izumitani, Eisaku Maeda
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0237, Japan
{taira,izumi,maeda}@cslab.kecl.ntt.co.jp

Takeshi Suzuki
Nagaoka University of Technology
1603-1 Kamitomioka-machi, Nagaoka
Niigata 940-2188, Japan
takesu@pelican.nagaokaut.ac.jp

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 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

There are two main conventional algorithms for the prediction of RNA’s secondary structures. They are Nussinov’s algorithm [5] and Zuker’s algorithm [6]. In this paper, we focus on Nussinov’s algorithm. This algorithm utilizes dynamic programming to search for remote base pairs.

The Nussinov Algorithm and Nussinov Algorithm using SCFG [1, 2, 3, 4] have some problems. One problem is that the algorithm only considers the maximum number of base pairs when searching for an optimal structure. Hence, even if the predicted loops are short, the algorithm tends to make base pairs. In the real world, since short loops are often thermodynamically unstable structures, we will obtain many incorrect structures.

Another problem with the Nussinov algorithm is that it only takes regular base pairs into consideration. The regular base pair, Adenin-Uracil, Guanine-Cytosine, has 2 or 3 hydrogen bonds. By contrast, the non-regular base pair, Guanin-Uracil, has two hydrogen bonds and there is rebounding between an oxygen and an oxygen. The scoring table used by the Nussinov algorithm, however, only counts regular base pairs and equate A-U with G-C.

2. Methods

To compensate for above two problems and obtain high levels of performance when predicting secondary structures, we add a new scoring table and control the loop’s minimum length.

In the loop length problem, we ensured that loop length was six or more. In the scoring table, in proportion to the number of hydrogen bonds, we give 2 and 3 for the regular base pairs, respectively, and 1 for the G-U 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.

3. Results and Discussion

3.1. Experimental Setting

We predicted the structure of 20 non-coding RNA sequences taken from various web sites. Their lengths were 21-38 bases. For the experiment, we used the original Nussinov and improved Nussinov algorithm, as well as the Nussinov algorithm using SCFG, and the improved Nussinov algorithm using SCFG.

3.2. Evaluation Methods

Their algorithms were evaluated by accuracy, F-measure, and shape evaluation. Accuracy is the prediction rate showing whether the position is a base pair or a part of a loop. The F-measure is defined as a harmonic average of the precision and recall. The evaluation by the shapes assigns five levels to the result structures. The evaluation levels are: Perfect match: 5, Having one mistake related to the loop or bulge: 4, Having two mistakes related to the loop or bulge: 3, The number of hairpin loops is the same as the actual structure: 2, No match: 1.

Table 1. Accuracy, F-measure, Shape Evaluation for NA, NB, SA and SB

Seq No.	NA		NB		SA		SB	
	Acc. / F / Shape	Acc. / F / Shape	lp	gu	Acc. / F / Shape	Acc. / F / Shape	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 / 1.1	0.772 / 0.711 / 2.7			0.701 / 0.577 / 1.85	0.787 / 0.714 / 2.9		

3.3. Experimental Results

We compared results obtained with the original Nussinov (NA), the improved Nussinov (NB), the original Nussinov using SCFG (SA), and the improved Nussinov using SCFG (SB).

Their results are 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, “O”s in “lp” and “gu” columns indicate sequences that obtain higher F-measure owing to loop restriction and considering G-U pairs, respectively. In the improvements from NA to NB, the F-measure of 11 in 18 sequences increased owing to loop restriction, and the F-measure 9 sequence increased owing to considering G-U pair. Moreover, in the improvements from SA to SB, 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.

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

4. Conclusion

We presented an improved Nussinov algorithm for the prediction of RNA secondary structure. Our experimental

results indicate that this scoring approach and method work well.

References

- [1] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.
- [2] L. Grate. 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, 1995.
- [3] F. Lefebvre. 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, 1995.
- [4] F. Lefebvre. 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, 1996.
- [5] R. Nussinov, G. Pieczenk, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM Journal of Applied Mathematics*, 35:68–82, 1978.
- [6] M. Zuker. Computer prediction of RNA structure. *Methods in Enzymology*, 180:262–288, 1989.