Pseudo Folding Simulation for RNA Secondary Structures Prediction

Minoru Asogawa, Akihiko Konagaya
Massively Parallel Systems NEC Laboratory, RWCP *
4-1-1, Miyazaki, Miyamaeku, Kawasaki, Kanagawa 216, Japan
asogawa@csl.cl.nec.co.jp, konagaya@csl.cl.nec.co.jp

1 Introduction

There are two major methods for predicting RNA secondary structure [Gribskov 91]. One is the simulated annealing, which is finding the lowest energy conformation for given single-stranded nucleic acid sequence, and the other is finding the optimal ensemble from a pool of stacking region candidates. For generating a pool of stacking region candidates, all base pairs less than some distance threshold are evaluated whether it produces Watson-Crick base pair. By utilizing the distance threshold and restricting the search space, required computation time reduces in the second order [Akiyama 92]. This method is widely used in former literatures, the distance threshold is fixed at certain value, something around 200.

It is difficult to choose a suitable threshold for obtaining good prediction with little computation. Since the distance threshold affects both prediction accuracy and required computation. From the computational view point, the required computation in this method is $O(n^2)$, where n is the distance threshold. Therefore, it is desirable to employ small distance threshold as possible. One of the defects using small distance threshold is that the base pairs, which are longer that the distance threshold, are never considered. Therefore, the output of the system could be poor. Thus, when the threshold is small, the system might produce poor result with little computation. When the threshold is large, the system might produce good result with huge computation.

Furthermore, only few methods consider pseudo knot, since this makes the problem much complex. Since the presented method simulates folding process, it can predict pseudo knots.

2 Pseudo Folding Simulation

The authors present a solution for choosing the distance threshold value for generating a pool of stacking region candidates. For this purpose, a distance map, which represents the distance of each two base, is utilized. In this method, one stacking region is predicted at a time. When one plausible stacking region is predicted, the distance map is updated to reflect its double-helical structure. Since taking account real tertiary distance may take huge computation, a new distance is defined for the approximation. The approximated distance corresponds to the minimum distance of the tertiary distance. The details is depicted in Fig. 1.

On the left in Fig. 1, a single-stranded nucleic acid sequence and its distance map are shown. When second, third and fourth necleic acids forms a Watson-Crik pair with tenth, ninth and eighth necleic acids, the single-stranded sequence changes on the right in Fig. 1. In this method, the distance between two nucleic acids of Watson-Crik pair is assumed as one, which is same to that of stranded sequences. Once a stacking region is predicted, nucleic acids contained in that region are never considered for further stacking region candidates. To make this pseudo folding simulation realistic, it is assumed that the double-helical region is rigid. Any nucleic acid pairs, just before and after the stacking region, are not chosen as a stacking region anymore, since the stiff stacking region is considered to prevent those nucleic acids to come close. For example, on the right of in Fig. 1, nucleic acid 1 and nucleic acid 5 are prohibited to make a pair. This is same to a pair of nucleic acid 7 and nucleic acid 7. Note that, the length of the stacking region affects this pair inhibition.

*RWCP: Real World Computing Partnership

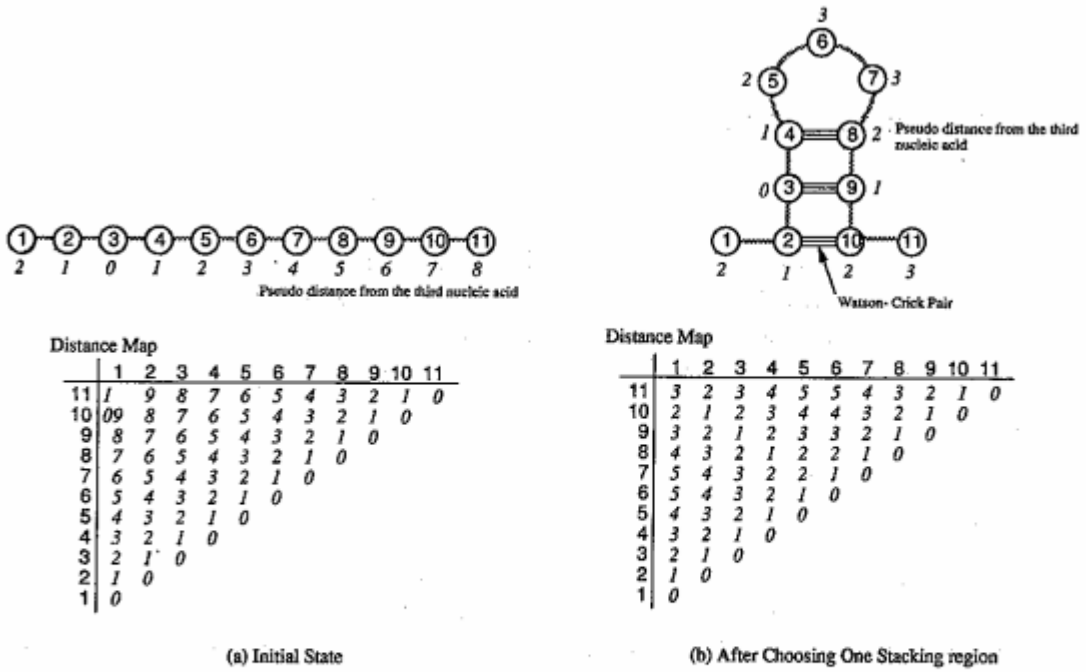

Figure 1: Distance Map Renovation

This method has been tested against several RNA sequences, which secondary structures are known. For about 80 percent of sequences, this method predicts their secondary structures correctly.

3 Conclusion

The authors present a novel method for generating a pool of stacking region candidates based on the small distance threshold. In this method, one stacking region is predicted at a time. A distance map, which represents the distance of each two base, is renovated, is updated to reflect its double-helical structure. This method can predict pseudo knots, by postulating that stacking region is stiff and prevents any nucleic acid pairs, just before and after the stacking region, to come close. With this method, required computation time reduces in $O(n^2)$. From preliminary experimental results with several RNA sequences, it is shown that this method is promising.

References

[Gribskov 91] Gribskov M. Devereux J., "Sequence Analysis Primer", Stockton Press, (1991).

[Akiyama 92] Akiyama Y., Furuya T., "A Fast RNA Secondary Structure Prediction System Based on Energy Minimization Property of a Hopfield Neural Network", *Proc. of 1992 Informatics Symp.*, (1992).