A Survey of RNA Structure Prediction Algorithms

Wenyan Jiang

***Abstract*— RNA plays an irreplaceable role in life activities. It not only participates in the expression of genes, but also plays an important role in various life activities. Therefore, it is the current research direction to predict RNA structure through RNA sequence information and to understand RNA function. Nowadays, predicting that RNA structures are maturing. The prediction of current RNA structure is mainly divided into the two level structure and tertiary structure. The prediction of two level structure mainly uses sequence similarity method and minimum free energy method. Predicting tertiary structure mainly uses molecular dynamics simulation. At present, the prediction of secondary structure has reached maturity, but the tertiary structure can only predict short sequences, so further research on tertiary structure is needed. This paper mainly introduces various methods.**

***Index Terms*— RNA secondary structure, RNA tertiary structure, molecular dynamics, sequence alignment**

I. INTRODUCTION

RNA is a genetic material that exists primarily in cells and viruses. A ribonucleotide molecule is composed of phosphoric acid, ribose and base. RNA can be divided into messenger RNA, transport RNA, and ribosomal RNA [1]. In the biological research, we have obtained a large amount of RNA one dimensional structure information. The one dimensional structure of RNA is mainly the linear structure formed by the nucleotide sequence according to the specific sequence of the nucleotide. The two level structure is a double helix formed by the folding of the nucleotide chain, which can be divided into a paired base pair and an unpaired ring. Based on two-dimensional structure, the tertiary structure is a spatial structure formed by the interaction between some structural units. For the study of the biological function of RNA, predicting the tertiary structure is the basis of the research, but the three-dimensional structure can not be directly obtained from the primary structure. So the tertiary structure of RNA should be obtained on the basis of the fine two-dimensional structure.

The two-dimensional structure of RNA is formed by base pairing, and the pairing type is divided into: A-U, G-C, and G-U [2]. Folding motion by a planar structure formed by components such as a single-stranded domain structure, a stem-loop structure, and a double-structure. The secondary structure of RNA is mainly composed of a multi-branched ring and a free single chain. The multi-branched strand is a hairpin loop that is not paired in the structure of the RNA molecule, and the free single strand is the single-stranded single strand in the RNA molecule. The common secondary structure mainly has a ladder structure, a ring structure, a convex hull structure and a tail structure. The trapezoid structure is a group of continuous base pairs. There are no other pairs of bases between the first group and the last group of base pairs, and they are matched in sequence from beginning to end. At the end of the ladder structure, a set of consecutive unmatched ribonucleotides are attached at both ends of the ladder, and the structure composed of these unmatched ribonucleotides is a ring structure. On the basis of the ladder structure, if there is more than one unmatched ribonucleotide on only one side of the ladder, the secondary structure composed of these consecutive unmatched ribonucleotides is a convex hull structure. At the end of the ladder structure is a set of consecutive unmatched ribonucleotides at the ends of the long strand of RNA, and the secondary structure consisting of these consecutive unmatched ribonucleotides is the tail structure.

The tertiary structure of RNA refers to the interaction between the secondary structural units and the spatial orientation of the secondary structural units. The spatial base stacking is the main force for stabilizing the RNA tertiary structure [3]. The local conformation of two or more helix plays an irreplaceable role in stabilizing the tertiary structure of RNA. In a structure such as a junction, one helix is stacked on another helix. The accumulation between helix determines the overall conformation of the RNA molecule.

II. RESEARCH METHODS

2.1 *Prediction method of RNA Secondary structure*

In the current biological research, the main methods for predicting the secondary structure of RNA are sequence alignment method and minimum free energy method.

A. Sequence comparison method

The sequence-specific approach is commonly used methods for predicting the RNA structure. The method is to judge whether or not the two sequences are similar through a large number of data sets, so that the function of the gene product generated by the corresponding sequence can be obtained. Sequence alignment methods can be divided into three methods: comparison before prediction, forecast before comparison, compare and predict at the same time. The first method is more suitable for multiple sequence alignment. The first method is more suitable for multi-sequence comparison. In general, multi-sequence comparison tools such as ClustalW [4], which is mainly carried out by Pfold [5] software. Combined with random context-independent grammar rules, training is carried out in known data sets, then find out the generation rule of the maximum sequence probability. This method is relatively high in prediction accuracy. Sankoff algorithm [6] is mainly used to compare and predict at the same time. Time complexity of the

Manuscript received November 19, 2019
Wenyan Jiang, School of computer science and technology, Tiangong University ,Tianjin,300387,China

algorithm is high, which is the combination of sequence alignment and Nussinov[7] algorithm, and its time complexity is O (N^{3m}).

The current mainstream algorithm includes the random context-independent grammar and the cooperative variation model. These two methods are often used in predictive structures.

Prediction based on random context-free grammar mainly marks bases in the form of characters and specifies the types of different substructures described by production. Using production rules to construct syntax tree represents a secondary structure of prediction. However, because of the different probability of the different generation, the probability is calculated by the algorithm to construct the most probable syntax tree. The disadvantage of this algorithm is that the computational complexity is too high.

The cooperative variation prediction algorithm is to find the potential secondary structure by optimizing the sequence.

B. Method based on minimum Free Energy

Based on the primary sequence of RNA without a priori knowledge, we often use the method of minimum free energy to predict RNA structure. Each branch structure has a homologous free energy, the free energy of the trapezoidal structure is negative, and the free energy of the annular structure is positive. The longer the ladder, the smaller the free energy, therefore, it can be concluded that the base free energy of the pairing can be reduced, and the free base free energy can be raised, so that the Nusinov algorithm uses the dynamic programming method. This is a folding method for finding the most matching pair, but its disadvantage is that it does not consider the influence of the free energy of the ring region, so its time complexity is reduced and becomes $O(n^3)$. The most classical method of dynamic programming is the Zuker algorithm [8] to compute the minimum free energy. By combining the free energy of the individual structures of the convex ring, the inner ring and the hairpin ring to obtain the free energy of the whole molecule, the time complexity of this method is $O(n^4)$. In terms of time complexity, the Zuker algorithm is higher than the Nussinov algorithm, and the former can repeatedly obtain the energy value of the ring structure. The same inner ring has the same free energy. After the calculation, the time complexity can be reduced to $O(n^3)$[9].

There is no large ring structure in the molecule, so if the inner ring size is at most k, the time complexity is $O(kn^2)$[10]. However, experiments have shown that the actual structure of RNA is inconsistent with the minimum free energy, so to achiene the purpose of solving the inconsistency between the two, Zuker[11] et al. proposed a suboptimal structure. Setting a reasonable threshold to forecast the true structure, the difference from the minimum structure of free energy is considered to be the true structure of RNA within the threshold. So the choice of threshold should be moderate, too large may make the research complicated, too small will affect the real Structural results.

2.2 Prediction method of RNA Tertiary structure

The prediction of RNA tertiary [12]structure is relatively immature, but RNA can only function normally if it forms a three-dimensional structure. The initial experimental method is not only costly but also technically challenging, such as X-ray crystal diffraction, NMR, etc. Molecular dynamics [13] simulations predict RNA structure is also an effective method. RNA sequence was selected as the research object. In the prediction of secondary structure, the three-dimensional initial state model was constructed by base pairing structure according to the lowest structure and molecular sequence of free energy. In view of this model, the empirical energy was scored. The function adjusts the three-dimensional structure, and finally obtains the three-dimensional structure of the RNA. Based on the two-dimensional structure, we often use methods such as MC-Fold [14]. MC-Fold predicts the secondary structure based on the sequence, and then the secondary structure predicts the tertiary structure as a topological constraint.

CONCLUSION

Structural prediction is the basis for understanding functionality. The prediction of RNA secondary structure is gradually applied to many fields. The combination of minimum free energy method and sequence alignment usually increases the prediction accuracy, but further research is needed, and the prediction of three-dimensional structure is still in the primary stage. The structural prediction of long RNA sequences and the improvement of the accuracy of energy scoring function will be solved in the future research.

REFERENCES

[1] Turner D H, Sugimoto N, Freier S M. RNA structure prediction[J]. Annual review of biophysics and biophysical chemistry, 1988, 17(1): 167-192.

[2] Abrahams J P, van den Berg M, Van Batenburg E, et al. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation[J]. Nucleic acids research, 1990, 18(10): 3035.

[3] Ya-Zhou S, Yuan-Yan W, Feng-Hua W, et al. RNA structure prediction: progress and perspective[J]. Chinese Physics B, 2014, 23(7): 078701.

[4] Li, Kuo-Bin. "ClustalW-MPI: ClustalW analysis using distributed and parallel computing." Bioinformatics 19.12 (2003): 1585-1586.

[5] Knudsen, Bjarne, and Jotun Hein. "Pfold: RNA secondary structure prediction using stochastic context-free grammars." Nucleic acids research 31.13 (2003): 3423-3428.

[6] Sankoff, David. "Simultaneous solution of the RNA folding, alignment and protosequence problems." SIAM journal on applied mathematics 45.5 (1985): 810-825.

[7] Nussinov, Ruth, et al. "Algorithms for loop matchings." SIAM Journal on Applied mathematics 35.1 (1978): 68-82.

[8] Jaeger, John A., Douglas H. Turner, and Michael Zuker. "Improved predictions of secondary structures for RNA." Proceedings of the National Academy of Sciences 86.20 (1989): 7706-7710.

[9] Waterman, Michael S., and Temple F. Smith. "Rapid dynamic programming algorithms for RNA secondary structure." Advances in Applied Mathematics 7.4 (1986): 455-464

[10] Lyngso R B , Zuker M , Pedersen C N . Fast evaluation of internal loops in RNA secondary structure prediction[J]. Bioinformatics, 1999, 15(6):440-445.

[11] Mathews, David H., et al. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." Journal of molecular biology 288.5 (1999): 911-940.
[12] Hajdin, Christine E., et al. "On the significance of an RNA tertiary structure prediction." Rna 16.7 (2010): 1340-1349.
[13] Haile, James M. Molecular dynamics simulation: elementary methods. Vol. 1. New York: Wiley, 1992.
[14] Parisien, Marc, and Francois Major. "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." Nature 452.7183 (2008): 51.