# Research Progress on Algorithms for RNA Secondary Prediction

**Guosong Jiang, Ke Chen**

*Abstract*— **Besides participating into genes' expression, modern cell biology scientists uncover that RNA plays diverse roles in whole life of cells, including catalyzation, regulation, etc. These types of RNA are usually called non-coding RNA, whose research have become a hot spot. Predicting RNA structures using information technology origins from the early works of Waterman in 1970s, he developed the theory of minimum free energy to predict unknown RNA structures. Lots of Corollaries proposed in this theory have become the theoretical foundation of many later prediction algorithms. After decades of development, the algorithms for RNA prediction become more and more mature and specific. RNA secondary structure prediction is the first step of RNA structure prediction. Nowdays, mainstream method generally includes minimum free energy methods and sequence alignment methods. State of the art methods usually focus on multi-branches prediction and combination of experimental data. This article reviews the methods in this area, concentration will give on the latest technology, including optimization of minimum free energy methods and multi-branches prediction.**

*Index Terms*—**RNA structure prediction, RNA secondary structure, free energy, sequence searching**

## I. INTRODUCTION

RNA's function is firstly proposed by Watson and Crick in their famous genetic central dogma [1]. In this dogma, RNA plays roles in helping gene's transcription, transportation and translation, and in the later few decades, people including biology scientists believe protein is the major and even the only tool natural lifes use to regulate its living [2]. Untill recent years, cell biology scientists decide to extend the old golden dogma, due to more and more organic molecule is found to play critical characters in cell's regulation, and RNA is not the exception. Non-coding RNA is the type of RNA which do not participate into protein translation but play roles in catalyzing and molecular slicing [3]. Recent study shows that the number of non-coding RNA is not even fewer than the number of proteins exist in cell, which indicates the developing status of RNA [4].

Traditional RNA structure resolution methods including nuclear magnetic resonance (NMR) and X-ray crystallography. NMR technology is one of nondestructive examination, which means it can get RNA structural information without hurting its structure. Therefore, using NMR to solve RNA structure usually have relatively higher accuracy than other experiment methods. At the same time, NMR technology has shortcomings of low detecting sensitivity, high equipment cost, difficulty in getting qualified RNA samples. X-ray crystallography is another technology often used by biologists to solve molecular structure, such as protein and RNA. But it is always a challenge especially for new laboratory assistant to carry out qualified RNA crystallization. Moreover, X-ray crystallography is also a expensive and time consuming way to get one aimed RNA structure. In general, traditional RNA structure resolution methods are not practical enough to solve a mass amount of RNA exist in nature, which cause a great gap between known RNA sequence and known RNA structure. Therefore, using information technology to predict RNA structure is the future of RNA structure solving [5].

RNA prediction algorithms are hierarchical [6]. RNA sequence is usually called the first structure and used as input of secondary structure prediction algorithms. The output of secondary structure prediction algorithms is RNA secondary structure, which consist of WC base-pairs. And then, the RNA secondary structure is used as input for tertiary structure prediction algorithms to generate RNA tertiary structure, which illustrates the spacial arrangement of all atoms of RNA and the interaction inside RNA molecules and with other molecules such as proteins. Therefore, RNA secondary structure prediction is the first step of the whole RNA prediction task, whose result will affect the quality of RNA tertiary structure prediction [7].

the mainstream methods of RNA secondary structure prediction include sequence alignment and minimum free energy methods [8]. Alignment is a classical method in evolutionary biology, it is also called comparative method. Sequence alignment bases on such a natural fact: When some species have evolutionary relationship, its genes show strong similarity, which causes its genes' products' similarity, such as protein and RNA sequence and structure [9]. It is generally believed that if two RNA sequence or part of two RNA sequence have strong similarity, they usually share the same RNA structural arrangement [10]. Given an aimed RNA sequence, sequence alignment algorithms search the backup RNA database for homologous sequences, and then uses its homologous sequences' known structure to generate the aimed RNA's secondary structure. algorithms of sequence alignment usually have O(m×n) and O(m×n) complexities for time and space respectively [11], where m and n represent the length of sequence for comparing. How to improve algorithms' efficiency is a challenge for comparative RNA scientists. What's more, sequence alignment requires homologous feature, which means this method do not support new types of RNA's structure prediction. This situation encourages scientists focus on another type of RNA prediction methods, the minimum free energy methods. This

Manuscript received Oct 09, 2018

**Guosong Jiang**, School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin, 300387, China

**Ke Chen**, *Corresponding Author, School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin, 300387, China

type of methods stems from Zuker and Stiegler's works on algorithm of searching for RNA secondary structure with minimum free energy [12]. The algorithm base on another theory Anfinsen proposed which illustrates that biomolecule is prone to status of minimum free energy. This theory builds the accuracy of minimum free energy methods [13]. Given an aimed RNA sequence, algorithm searches for all possible secondary structure, and then calculates their free energy one by one, at last, choosing the structure with minimum free energy as final result. The quality of final structure heavily depends on the free energy model. Tinoco and Salser proposed the free energy nearest-neighbor model which is wildly applied by scientists. This model is later standardized and extended by Turner's team [14]. The difficulty of the minimum free energy methods is exemplified by the fact that an RNA with n nucleotides can generate the order of $1.8^n$ possible secondary structures, optimization is necessary [15]. Other difficulties including multi-branches searching and accuracy improving. The following content will introduce the recent progress made to tackle these problems.

## II. RESEARCH PROGRESS

### *2.1 Optimization of the minimum free energy methods*

Minimum free energy methods are required to search every possible secondary structure to find the lowest free energy structure, if algorithm explicitly generate every possible secondary structure to calculate its free energy, it will encounter a problem that the number of possible structures for an RNA sequence increases exponentially with the sequence length. For sequences have length of 200 nucleotides, it will cost common personal computer with modern processor billions of years to find the lowest free energy structure in all possible secondary structure.

The basic and popular solution is the implementation of dynamic programming technology. This technology allows algorithm to avoid to generate every possible structure, but implicitly searches every short fragment to generate the minimum free energy structure at once. The algorithm for minimum free energy methods usually have two steps, fill and traceback, getting the minimum free energy and the structure of the minimum free energy respectively. The fatest algorithm using dynamic programming technology scale $O(n^3)$ [16].

GAs are another popular choices which based on the concepts of biological evolution. In every step of GAs, mutation, crossover and selection are introduced to make random changes to the solutions. Due to the nature of GAs is stochastic, GAs are usually run several times to obtain a consensus structure as final output structure [17].

### *2.2 Multi-branches' searching*

Multi-loops are loops where three or more helix encounter. Multi-loops are such important fragment of RNA secondary structure that affect the shape of RNA secondary structure and the arrangement of final tertiary structure in great magnitude. Original minimum free energy algorithms usually ignore multi-loops by giving them zero free energy. An ad-hoc, linear function considering the size of loops and the number of branches was firstly used to simulate Multi-loops energy. This type of linear model requires $O(n^3)$, time and $O(n^2)$, space, where n is the number of nucleotides in an RNA. Another model of multi-loops energy is called Jacobson–Stockmayer model, which would require exponential computation time. Ward and Datta optimize this model by improving the complexity of time and space to $O(n^4)$ and $O(n^3)$, which still lag behind the linear model. Ward and Datta also test the performance of the two models, which indicates the linear model is better [18].

### 2.3 Using specific motifs

Using RNA specific motifs to improve RNA structure prediction accuracy is another hot spot in recent years. RNA specific motifs, such as u-turns, k-turns and a-platform are widely occurring motifs in RNA, which are usually located in internal loops. Experiment shows they have strong relationship with many biological functions such as translation, regulation, etc. Due to these specific motifs usually associate with specific sequence pattern, it usually helps to improve the accuracy of RNA structure prediction. For example, Bayrak use sequence signatures and kink-turn motifs to successfully enhance their team's previous proposed algorithm [19].

## III. CONCLUSION

the ability of the classical minimum free energy methods that predict RNA structure directly from sequence data is still limited. Problems mainly come from two facts. On the one hand, the natural structure of an RNA is not static, which means the structure with the minimum free energy may not be the only natural structure or isn't the natural structure. On the other hand, solvent, ions, proteins and other biomolecules are important environmental factors affect the formation of RNA structure. Recently, a few algorithms trying to incorporate experimental information as constraints to improve the performance are proposed and exhibit bright future for the whole computational prediction methods.

## REFERENCES


[1] M.Parisien, F.Makor, The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data, Nature, 2008, 452:51–55 .

[2] B.A.Shapiro, Y.G.Yingling, W.Kasprzak, E.Bindewald, Bridging the Gap in RNA Structure Prediction, Current Opinion in Structural Biology, 2007,17:157–165..

[3] . N.Kim, S.Elmetwaly, S.Jung, T.Schlick, Graph-Based Sampling for Approximating. Global Helical Topologies of RNA, PNAS, 2014, 111 (11):4079–4084

[4] Z.Xu1, D.H.Mathews, Multilign: an Algorithm to Predict Secondary Structures Conserved in Multiple RNA Sequences, Bioinformatics, 2011, 27(5):626–632.

[5] Z.Miao, R.W. Adamiak, M.F.Blanchet, M.Boniecki, RNA-Puzzles Round II: Assessment of RNA Structure Prediction Programs Applied to Three Large RNA Structures, RNA,2015,21:1066–1084.

[6]. A Range of Complex Probabilistic Models for RNA Secondary Structure Prediction that Includes the Nearest-neighbor Model and More, RNA, 2012, 18:193–212

[7] N.B. Leontis, E.Westhof, Geometric Nomenclature and Classification of RNA Base Pairs, RNA, 2001, 7:499–512.

[8] The Building Blocks and Motifs of RNA Architecture, Curr Opin Struct Biol, 2006, 16(3):279–287.

[9] Revolutions in RNA Secondary Structure Prediction, J. Mol. Biol., 2006, 359:526–532

[10] M.Waterman, Secondary Structure of Single-Stranded Nucleic Acidst, Studies in Foundations and Combinatorics Advances in Mathematics Supplementary Studies, 1978, 1.

[11] The Kink-turn: A New RNA Secondary Structure Motif. The EMBO Journal.

[12] R.Nussinov, A.B.Jacobsont, Fast Algorithm for Predicting the Secondary Structure of Single-stranded RNA, Proc.Nati.Acad.Sci.

[13] M.Andronescu, A.Condon, H.H.Hoos, D.H.Mathews,K.P.Murphy, Computational Approaches for RNA Energy Parameter Estimation, RNA, 2010, 16:2304–2318.

[14] M.H.Bailor, A.M.Mustoe, C.L.Brooks, H.M.AlHashimi, 3D Maps of RNA Interhelical Junctions, nature protocols , 2011, 6(10):1536-1545.
[15] M.Hamada, H.Kiryu, K.Sato, T.Mituyama, Prediction of RNA Secondary Structure using Generalized Centroid Estimators, BIOINFORMATICS, 2009, 25(4): 465–473.
[16] Z.Xu, A.Almudevar, D.H.Mathews, Statistical Evaluation of Improvement in RNA Secondary Sructure Prediction, NAR, 2012, 40(4):e26.
[17] M.Ward, A.Datta, M.Wise, D.H.Mathews, Advanced Multi-loop Algorithms for RNA Secondary Structure Prediction Reveal that the Simplest Model is Best, NAR, 2017, 45(14): 8541–8550
[18] R.Das, D.Baker, Automated De novo Prediction of Native-like RNA Tertiary Structures, PNAS, 2007, 104(37): 14664–14669
[19] C.S.Bayrak, N.Kim, T.Schlick, Using Sequence Signatures and Kink-turn Motifs in Knowledge-based Statistical Potentials for RNA Structure Prediction, NAR, 2017, 45(9): 5414–5422.