Predicting RNA Secondary Structures: One-grammar-fits-all Solution

Menglu Li¹, Micheal Cheng², Yongtao Ye¹, Wk Hon², Hf Ting¹, Tw Lam¹, Cy Tang²,³, Thomas Wong⁴, and Sm Yiu¹*(Letter)

¹ Department of Computer Science, The University of Hong Kong, Hong Kong, China
² Department of Computer Science, National Tsinghua University, Hsinchu City, Taiwan
³ Department of Computer Science, Providence University, Taichung City, Taiwan
⁴ CSIRO Ecosystem Sciences, Canberra, Australia
smyiu@cs.hku.hk

Abstract. RNA secondary structures are known to be important in many biological processes. Many available programs have been developed for RNA secondary structure prediction. Based on our knowledge, however, there still exist secondary structures of known RNA sequences which cannot be covered by these algorithms. In this paper, we provide an efficient algorithm that can handle all RNA secondary structures found in Rfam database. We designed a new stochastic context-free grammar named Rectangle Tree Grammar (RTG) which significantly expands the classes of structures that can be modelled. Our algorithm runs in $O(n^6)$ time and the accuracy is reasonably high, with average PPV and sensitivity over 75%. In addition, the structures that RTG predicts are very similar to the real ones.

1 Introduction

Secondary structures of RNA molecules play important roles in their functionalities [1, 2]. Many methods have been proposed to predict RNA secondary structures. Although the majority of RNAs have simple secondary structures, pseudoknots (base pairs crossing each other) are found in almost all classes of RNAs. Pseudoknots are known to be involved in biological functions such as stimulating ribosomal frameshifting [3, 4]. The existence of pseudoknots make the secondary structure prediction an NP-hard problem, in general [5, 6]. Existing algorithms attempt to solve the problem by considering a restricted set of pseudoknots [7–18]. Not all existing pseudoknots can be modelled. In terms of prediction accuracy, CentroidAlifold[9] generalized a centroid estimator that maximizes the expected accuracy of structure prediction. Tabei, Yasuo and Kiryu[16] proposed a fast multiple sequence alignment method named MXScarna in which the optimal structure that maximized a heuristic scoring function was found during the group alignments of stem component sequences. RNAaliFold[17] pre-computed alignments using a combination of free-energy and a covariation measures, whilst TurboFold[18] utilized an iterative probabilistic method to predict secondary structures for multiple RNA sequences.
Despite of so many algorithms to predict RNA secondary structures, there exist secondary structures of known RNAs in Rfam [19] that cannot be covered by existing efficient algorithms\(^1\). Figure 1 shows such an example.

Fig. 1. A structure in Rfam which cannot be handled by existing efficient algorithms

In this paper, we proposed a grammar-based machine learning method to predict secondary structures for all RNA sequences in Rfam. Enlightened by [20], we designed a new stochastic context-free grammar called Rectangular Tree Grammar (RTG), which can model all possible secondary structures of known RNA sequences in the Rfam database. Each structure can be generated by a unique operation path, that is, the only sequence of operations that yields this sequence. A set of paths is obtained using some real RNA sequences with known structures. Rule transition probabilities and base emission probabilities are calculated based on this set. In order to determine the unknown secondary structure of a RNA sequence, dynamic programming is adopted to generate the most probable structure. This procedure takes \(O(n^6)\) time, where \(n\) is the length of input RNA sequence.

The proposed approach was evaluated using several sets of sequences with one containing pseudoknot-free structures and the others with different types of pseudoknots. We compared the performance of RTG with popular prediction algorithms including gfold[7], CentroidAlifold[9], pknotsRG[21], NUPACK[22], MXScarna[16], RNAaliFold[17] and TurboFold[18]. The experimental results have shown that our approach outperforms others substantially with high PPV and sensitivity, especially on highly-pseudoknotted sequences.

\[\text{2 Method}\]

\[\text{2.1 RNA Secondary Structure Definitions}\]

Let \(S = s_1s_2 \ldots s_n\) be an RNA sequence of length \(n\). \(M_{x,y}\) is the set of base pairs in the range \([x,y]\), \(M_{x,y} = \{(i,j)|x \leq i < j \leq y, (s_i, s_j)\text{ is a base pair}\}\).

**Banding:** The secondary structure of \(s_x \ldots s_y\) is a banding if it satisfies the following conditions:
(i) for any \(i, j, k, l \in [x, y], i \neq k, j \neq l, \text{ if } (i, j) \in M_{x,y} \text{ and } (k, l) \in M_{x,y}, \text{ then } i < k < l < j \text{ or } k < i < j < l. \)
(ii) \((x, y) \in M_{x,y}\).

\(^1\) We only consider algorithms which run in \(O(n^6)\) time.