Subcubic Time Algorithms for RNA Secondary Structure Prediction

Tatsuya Akutsu
takutsu@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

A lot of studies have been done on RNA secondary structure prediction [6], which is a problem of, given an RNA sequence of length $n$, finding its correct secondary structure (an outerplanar graph like structure). Usually, it is modeled as a free-energy minimization problem, for which simple DP (dynamic programming) algorithms have been proposed [6]. However, from a viewpoint of computational complexity, there had been no improvement on global free-energy minimization for 20 years (although there had been significant improvements on finding locally stabilizing substructures) [6].

In a basic and simplest version, global free-energy minimization of an RNA secondary structure is defined as a problem of maximizing the number of complementary base pairs. This problem is denoted by $RNA_0$ in this article. Even for $RNA_0$, a simple $O(n^3)$ time DP algorithm had been the fastest algorithm for 20 years. Recently, we have developed slightly improved algorithms for $RNA_0$: an $O(n^3(\log \log n)^{1/2}/(\log n)^{1/2})$ time exact algorithm and an $O(n^{2.776})$ time approximation algorithm [1]. We briefly describe outline of the algorithms in this short article. Details of the algorithms and extensions to more practical versions of the problem will appear in Ref. [1].

2 Exact Algorithm

It is well known that RNA secondary structure prediction can be formalized as a problem of constructing an optimal parse tree for a stochastic context-free grammar (SCFG, in short) [2, 4]. For this problem and context-free recognition (i.e., deciding whether or not there exists a parse tree for a given sequence), simple $O(n^3)$ time DP algorithms were well known. However, Valiant developed $O(n^{2.376})$ time algorithm for context-free recognition [5], by using fast matrix multiplication. Recently, we found that Valiant’s algorithm can be modified for the construction of an optimal parse tree for SCFG by replacing matrix multiplication with funny matrix multiplication. Using the current fastest algorithm for funny matrix multiplication [3], we have:

**Theorem 1.** For $RNA_0$, an optimal RNA secondary structure can be computed in $O(n^3(\log \log n)^{1/2}/(\log n)^{1/2})$ time.

3 Approximation Algorithm

Although the above improvement is very slight and is not practical, it seems difficult to develop faster algorithms. Thus, we developed an $O(n^{2.776})$ time approximation algorithm which always outputs a secondary structure whose score (i.e., the number of base pairs) is at least $1 - \epsilon$ of the optimal, where $\epsilon > 0$ is any fixed constant.

This approximation algorithm is a combination of an exact algorithm $A_{\text{exact}}$ and an approximation algorithm $A_{\text{approx}}$: $A_{\text{exact}}$ is used when the optimal score is small (precisely, the optimal score is $O(n^?)$
where $\gamma$ is a constant), otherwise $A_{\text{approx}}$ is used. $A_{\text{exact}}$ is similar to the exact algorithm in Section 1 and details are omitted here. $A_{\text{approx}}$ is obtained by modifying the original $O(n^3)$ time DP algorithm.

Let $a_1 \ldots a_n$ be an input RNA sequence. Then, it is well known that the optimal score $S(i, j)$ for subsequence $a_i \ldots a_j$ can be computed by the following simple DP procedure:

$$S(i, j) = \max \left\{ \begin{array}{c} S(i+1, j-1) + \mu(a_i, a_j), \\
\max_{i+1 \leq k \leq j} \{ S(i, k-1) + S(k, j) \} \end{array} \right\},$$

where we let $S(i, j) = 0$ for all $i \geq j$, and $\mu(x, y) = 1$ if $(x, y)$ is a base pair, otherwise $\mu(x, y) = 0$.

In $A_{\text{approx}}$, we do not compute $\max_{i+1 \leq k \leq j} \{ S(i, k-1) + S(k, j) \}$ exactly. Instead, we compute the maximum of $S(i, k-1) + S(k, j)$ for $O(n^\alpha + n^{1-\beta})$ values of $k$’s (see Fig. 1), where $\alpha$ and $\beta$ ($0 < \alpha, \beta < 1$) are appropriate constants.

Making detailed analysis on $A_{\text{exact}}$ and $A_{\text{approx}}$, we can show the following:

**Theorem 2.** For $\mathcal{RN}A_0$, an RNA secondary structure with the score at least $1 - \epsilon$ of the maximum can be computed in $O(n^{2.776})$ time, where $\epsilon$ is any fixed positive number.

![Figure 1: In $A_{\text{approx}}$, $\max_k S(i, k-1) + S(k, j)$ is computed not for all $k$, but for $O(n^\alpha + n^{1-\beta})$ values of $k$’s, where such $k$’s are represented by white circles in this figure.](image_url)

**References**


