A_007

Approximation Algorithms for Optimal RNA Secondary Structures Common to Multiple Sequences

Takeyuki Tamura* Tatsuya Akutsu*

1 Introduction

RNA secondary structure prediction is an important problem in *computational biology* and thus many computational studies have been done. This is a problem of, given an RNA sequence of length n, finding its correct secondary structure. Usually, RNA secondary structure prediction is modeled as a *free energy minimization* problem [7, 9]. For this problem, [10] and [11] proposed simple DP (dynamic programming) algorithms. The time complexities of those DP algorithms were $O(n^3)$ if we ignore the destabilizing energy due to loop regions, otherwise they were at least $O(n^4)$.

In a basic and simplest version, free-energy minimization of an RNA secondary structure is defined as a problem of *maximizing the number of complementary base pairs*, which is denoted by $\mathbf{RNA}_0$ in this paper. Even for $\mathbf{RNA}_0$, only an $O(n^3)$ time simple DP algorithm had been known [7, 9].

An $O(n^{2.776} + (1/\epsilon)^{O(1)})$ time approximation algorithm was also shown for $\mathbf{RNA}_0$ in [1], which always outputs an RNA secondary structure with the score at least $1-\epsilon$ of the maximum, where ϵ is any positive constant number and the score denotes the number of complementary base pairs in $\mathbf{RNA}_0$. Although this algorithm can be considered as a PTAS (polynomial time approximation scheme), it is different from usual PTAS since the problem is not NP-hard but belongs to P. This algorithm is a combination of an approximation algorithm $\mathcal{A}_{approx}$ and an exact algorithm $\mathcal{A}_{exact}$, where $\mathcal{A}_{approx}$ is obtained by modifying the original $O(n^3)$ time DP algorithm for $\mathbf{RNA}_0$, and $\mathcal{A}_{exact}$ is obtained by combining Valiant's algorithm with fast funny matrix multiplication.

In order to improve the prediction accuracy, an approach using multiple RNA sequences from the same RNA family was proposed [6]. An $O(n^6)$ time exact algorithm was shown in [6] which can optimize structure and alignments when two RNA sequences are given. Though some efforts have been done [2, 3], the worst case time complexity has not been improved. In this paper, we show an $O(n^5)$ time approximation algorithm for optimizing structure and alignments of two RNA sequences with assuming that the optimal number of base-pairs is more than $O(n^{0.75})$. We also show that the problem to optimize structure and alignments for given N sequences is NP-hard and introduce a constant-factor approximation algorithm.

2 RNA secondary structure when multiple sequences are given

Let $A_1 = a_{1,1}a_{1,2}\ldots a_{1,n_1}$, $A_2 = a_{2,1}a_{2,2}\ldots a_{2,n_2}$, ..., and $A_N = a_{N,1}a_{N,2}\ldots a_{N,n_N}$ be RNA sequences, where $\max\{n_1, n_2, \ldots, n_N\} = n$. Thus, A_1, A_2, $\cdots$, A_N are strings over an alphabet $\Sigma = \{a, u, g, c\}$. A family of pairs of indices $M_N = \langle\{(1_i, 1_j) | 1 \le 1_i < 1_j \le n_1, (a_{1,1_i}, a_{1,1_j})$ is a base pair$\}$, $\{(2_i, 2_j) | 1 \le 2_i < 2_j \le n_2, (a_{2,2_i}, a_{2,2_j})$ is a base pair$\}$,..., $\{(N_i, N_j) | 1 \le N_i < N_j \le n_N, (a_{N,N_i}, a_{N,N_j})$ is a base pair$\}\rangle$ is called an *N-common RNA secondary structure* if $a_{1,1_i} = a_{2,2_i} = \cdots = a_{N,N_i}$ and $a_{1,1_j} = a_{2,2_j} = \cdots = a_{N,N_j}$, and no distinct pairs (x_i, x_j), (x_h, x_k) in M_N satisfy $x_i \le x_h \le x_j \le x_k$ for all x $(1 \le x \le N)$. The score of M_N is defined as the number of base pairs in each element of M_N (i.e., $|e|$ for any e in M_N), and denoted by $score(M_N)$. Then, $\mathbf{RNA}_0(N)$ is defined as follows: given N RNA sequence $A_1, A_2, \ldots, A_N$, to find an N-common RNA secondary structure M with the maximum score. In $\mathbf{RNA}_0(N)$, such a structure is also called an *optimal N-common RNA secondary structure*, and denoted by $OPT(\mathbf{RNA}_0(N))$.

3 $1-\epsilon$ approximation algorithm for $\mathbf{RNA}_0(2)$

As mentioned above, in $\mathbf{RNA}_0(2)$, two sequences are given. Let (i_1, j_1) be a pair of indices which correspond to the leftmost and rightmost residues of the first sequence respectively. Similarly, let (i_2, j_2) be a pair of indices which correspond to the leftmost and rightmost residues of the other sequence respectively. $\mathbf{RNA}_0(2)$ can be solved in $O(n^6)$ time by the following DP procedure [6]:

$$D(i_1, j_1, i_2, j_2) = \max \begin{cases} D(i_1+1, j_1, i_2, j_2) \\ D(i_1, j_1-1, i_2, j_2) \\ D(i_1, j_1, i_2+1, j_2) \\ D(i_1, j_1, i_2, j_2-1) \\ D(i_1+1, j_1-1, i_2+1, j_2-1) + f(i_1, j_1, i_2, j_2) \\ \max_{i_1<k_1<j_1, i_2<k_2<j_2}\{D(i_1, k_1, i_2, k_2) + D(k_1+1, j_1, k_2+1, j_2)\}, \end{cases}$$

where $f(a,u,a,u) = 1$, $f(u,a,u,a) = 1$, $f(g,c,g,c) = 1$, $f(c,g,c,g) = 1$, otherwise f is zero.

The technique for the approximation algorithm of $\mathbf{RNA}_0$ [1] can be applied to $\mathbf{RNA}_0(2)$ with assuming $OPT(\mathbf{RNA}_0(N))$ is large, where additional ideas are required for analysis of the approximation ratio. As in [1], we do not compute $\max_{i_1<k_1<j_1, i_2<k_2<j_2}\{D(i_1, k_1, i_2, k_2) + D(k_1+1, j_1, k_2+1, j_2)\}$ exactly. Instead, we compute the maximum of $D(i_1, k_1, i_2, k_2) + D(k_1+1, j_1, k_2+1, j_2)\}$ for $O(\{n^\alpha + n^{1-\beta}\}^2)$ values of (k_1, k_2), where α and β $(0 < \alpha, \beta < 1)$ are appropriate constants to be determined later. We define a sequence of indices $f^+_{i_1}(h)$ and $f^+_{j_1}(h)$ for $h_1 = 0, 1, 2, \ldots$ by

$$f^+_{i_1}(0) = i_1 + \lceil n^\alpha \rceil, \qquad f^-_{j_1}(0) = j_1 - \lceil n^\alpha \rceil$$
$$f^+_{i_1}(h_1+1) = f^+_{i_1}(h_1) + \lceil (f^+_{i_1}(h_1) - i_1)^\beta \rceil, \quad f^-_{j_1}(h_1+1) = f^-_{j_1}(h_1) - \lceil (j_1 - f^-_{j_1}(h_1))^\beta \rceil$$

Next, we define $\mathcal{I}_1(i, j)$ by

$$\mathcal{I}_1(i_1, j_1) = \{k_1 | i_1 < k_1 \le n^\alpha \quad or \quad j_1 - n^\alpha \le k_1 \le j_1\} \cup \{f^+_{i_1}(h_1) | f^+_{i_1}(h_1) \le (i_1 + j_1)/2\} \cup \{f^-_{j_1}(h_1) | f^-_{j_1}(h_1) \ge (i_1 + j_1)/2\}.$$

*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan.

Similarly, we define a sequence of indices $f^+_{i_2}(h_2)$ and $f^+_{j_2}(h_2)$ for $h_2 = 0, 1, 2, \ldots$ and $\mathcal{I}_1(i,j)$ by

$$f^+_{i_2}(0) = i_2 + \lceil n^\alpha \rceil, \qquad f^-_{j_2}(0) = j_2 - \lceil n^\alpha \rceil$$
$$f^+_{i_2}(h_2+1) = f^+_{i_2}(h_2) + \lceil (f^+_{i_2}(h_2) - i_2)^\beta \rceil, \quad f^-_{j_2}(h_2+1) = f^-_{j_2}(h_2) - \lceil (j_2 - f^-_{j_2}(h_2))^\beta \rceil$$
$$\mathcal{I}_2(i_2, j_2) = \{k_2 | i_2 < k_2 \leq n^\alpha \quad or \quad j_2 - n^\alpha \leq k_2 \leq j_2\} \cup \{f^+_{i_2}(h_2) | f^+_{i_2}(h_2) \leq (i_2 + j_2)/2\}$$
$$\cup \{f^-_{j_2}(h_2) | f^-_{j_2}(h_2) \geq (i_2 + j_2)/2\}.$$

Then, the approximation algorithm $\mathcal{A}_{approx}(2)$ is expressed by the following DP procedure:

$$D'(i_1, j_1, i_2, j_2) = \max \begin{cases} D'(i_1+1, j_1, i_2, j_2) \\ D'(i_1, j_1-1, i_2, j_2) \\ D'(i_1, j_1, i_2+1, j_2) \\ D'(i_1, j_1, i_2, j_2-1) \\ D'(i_1+1, j_1-1, i_2+1, j_2-1) + f(i_1, j_1, i_2, j_2) \\ \max_{k_1 \in \mathcal{I}_1(i_1,j_1), k_2 \in \mathcal{I}_2(i_1,j_1)} \{D'(i_1, k_1, i_2, k_2) + D'(k_1+1, j_1, k_2+1, j_2)\}, \end{cases}$$

where $f(a,u,a,u) = 1$, $f(u,a,u,a) = 1$, $f(g,c,g,c) = 1$, $f(c,g,c,g) = 1$, otherwise f is zero.

Lemma 1 *$\mathcal{A}_{approx}(2)$ works in $O(n^{2\alpha+4} + n^{6-2\beta} + n^{5+\alpha-\beta})$ time.*

Here, we define the error of an N-common secondary structure M_N to $OPT(\mathbf{RNA}_0(N))$ to be $score(\mathbf{RNA}_0(N))) - score(M_N)$ (note that this value must be non-negative).

Lemma 2 *The error of a secondary structure M_N computed by $\mathcal{A}_{approx}$ is $O(n^{1+\alpha\beta-\alpha})$.*

Theorem 1 *When $OPT(\mathbf{RNA}_0(N)) > O(n^{0.75})$, an N-common RNA secondary structure with the score at least $1-\epsilon$ of the maximum can be computed in $O(n^5)$ time, where ϵ is any positive constant number.*

It is known that the Longest Common Subsequence problem (LCS) over an alphabet of size 2 is NP-hard [4, 5]. Using a reduction from LCS, we have:

Theorem 2 *$\mathbf{RNA}_0(N)$ is NP-hard if N is not fixed.*

Theorem 3 *There is an $O(Nn)$ time approximation algorithm for $\mathbf{RNA}_0(N)$ with the score at least $1/4$ of the maximum.*

4 Concluding remarks

In this paper, we proposed an $O(n^5)$ time (1-ϵ)-approximation algorithm for optimal RNA secondary structures common to two sequences with assuming that the optimal score is more than $O(n^{0.75})$. In order to delete this assumption, we should combine Valiant's algorithm [8] with the proposed algorithm. We also showed that the problem is NP-hard for general N and introduced an $O(Nn)$ time 1/4-approximation algorithm. Improvement of this approximation ratio is left as an open problem.

References

[1] T. AKUTSU, Approximation and exact algorithms for RNA secondary structure prediction and recognition of stochastic context-free languages, *Journal of Combinatorial Optimization* (1999) **3**:321–336

[2] V. BAFNA, S. MUTHUKRISHNAN, AND R. RAVI, Computing similarity between RNA strings, *Proc. 6th Symp. Combinatorial Pattern Matching* (1995) 1–16

[3] V. BAFNA, H. TANG, AND S. ZHANG, Consensus folding of unaligned RNA sequences revisited, *Journal of Computational Biology* (2006) **13(2)**:283–295

[4] D. MAIER , The complexity of some problems on subsequences and supersequences, *Journal of the ACM* (1978) **25**(2):322–336

[5] M. MIDDENDORF, On finding various minimal, maximal, and consistent sequences over a binary alphabet, *Theoretical Computer Science* (1995) **145**:317–327

[6] D. SANKOFF, Simultaneous solution of the RNA folding, alignment and protosequence problems, *SIAM Journal on Applied Mathematics* (1985) **45(5)**:810–825

[7] J. SETUBAL AND J. MEIDANIS, *Introduction to Computational Molecular Biology*, PWS Publishing Company (1997)

[8] L. G. VALIANT, General context-free recognition in less than cubic time, *Journal of Computer and System Science* (1975) **10**:308–315

[9] M. S. WATERMAN, Introduction to computational biology, *Chapman and Hall* (1995)

[10] M. S. WATERMAN AND T. F. SMITH, RNA secondary structure: A complete mathematical analysis, *Mathematical Biosciences* (1978) **41**:257–266

[11] M. ZUKER AND P. STIEGLER, Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Research* (1981) **9**:133–148