Theoretical Computer Science 556 (2014) 63–70

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Approximating the maximum multiple RNA interaction problem

Weitian Tong [a], Randy Goebel [a], Tian Liu [b], Guohui Lin [a,*]

[a] Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada
[b] Key Laboratory of High Confidence Software Technologies, Ministry of Education Institute of Software, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 30 January 2014
Received in revised form 27 March 2014
Accepted 12 April 2014
Available online 18 April 2014

Keywords:
RNA interaction
Maximum weight *b*-matching
Acyclic 2-matching
Approximation algorithm
Worst case performance ratio

ABSTRACT

RNA interactions are fundamental in many cellular processes, where two or more RNA molecules can be involved. Multiple RNA interactions are also believed to be much more complex than pairwise interactions. Recently, multiple RNA interaction prediction has been formulated as a maximization problem. Here we extensively examine this optimization problem under several biologically meaningful interaction models. We present a polynomial time algorithm for the problem when the order of interacting RNAs is known and pseudoknot interactions are allowed; for the general problem without an assumed RNA order, we prove the NP-hardness for both variants (allowing and disallowing pseudoknot interactions), and present a constant ratio approximation algorithm for each of them.

1. Introduction

RNA interaction is one of the fundamental mechanisms underlying many cellular processes, in particular genome regulatory code processes, such as mRNA translation, editing, gene silencing, and synthetic self-assembling RNA design. In the literature, pairwise RNA interaction prediction has been independently formulated as a computational problem by a number of groups of researchers [15,2,12]. While these variants are all motivated by specific biological considerations, the general formulation is usually NP-hard and many special scenarios have been extensively studied [13,4,5,16,8,9].

In more complex instances, biologists have found that multiple small nucleolar RNAs (snoRNAs) interact with ribosomal RNAs (rRNAs) in guiding the methylation of the rRNAs [11], and multiple small nuclear RNAs (snRNA) interact with an mRNA in the splicing of introns [17]. Multiple RNA interactions are believed much more complex than pairwise RNA interactions, where only two RNA molecules are involved. In fact, even if we have a perfect computational framework for pairwise RNA interactions, it might still be difficult to deal with multiple RNA interactions since, for a given pool of RNA molecules, it is non-trivial to predict their interaction order without sufficient prior biological knowledge.

Motivated by biological goals, Ahmed et al. have developed a system for multiple RNA interaction prediction, denoted as MRIP [1]. Here we provide basic definitions to formally introduce the MRIP problem. An RNA molecule is a sequence of nucleotides (A, C, G, and U). A basepair in the RNA is presented as (i, j), where $i < j$, indicating that the i-th nucleotide and the j-th nucleotide form a canonical pairing (*i.e.*, the two nucleotides are either A and U or C and G). The molecule folds into a *structure* which is described as a set of basepairs. In general, every nucleotide can participate in at most one basepair, and if not, it is a *free* base (or nucleotide). The set of basepairs is *nested* (a.k.a. secondary structure), if for any two basepairs

* Corresponding author.
E-mail addresses: weitian@ualberta.ca (W. Tong), rgoebel@ualberta.ca (R. Goebel), lt@pku.edu.cn (T. Liu), guohui@ualberta.ca (G. Lin).

http://dx.doi.org/10.1016/j.tcs.2014.04.018
0304-3975/

(i_1, j_1) and (i_2, j_2) with $i_1 < i_2$, either $j_1 < i_2$ or $j_2 < j_1$; otherwise the set is *crossing* (a.k.a. tertiary structure) containing *pseudoknots*. An interaction between two RNAs is a basepair which consists of one free base from each RNA. In the sequel, we use interaction and basepair interchangeably.

In the MRIP problem, we are given a pool of RNA sequences denoted as $\mathcal{R} = \{R_1, R_2, \ldots, R_m\}$. Without loss of generality, we assume m is even and these RNA sequences have the same length n. We use $R_{i\ell}$ to denote the ℓ-th base of R_i. Following the formulation by Ahmed et al. [1], the possible interactions between every pair of RNAs are assumed known. In fact, these possible interactions can be predicted using existing pairwise RNA interaction predictors [13,4,5,16,8,9]. For a possible interaction $(R_{i_1\ell_1}, R_{i_2\ell_2})$, its weight $w(R_{i_1\ell_1}, R_{i_2\ell_2})$ can be set using a probabilistic model or using an energy model or simply at 1 to indicate its contribution to the structure stability. The problem goal is to find out the order of RNAs in which they interact, that the first RNA interacts with the second RNA, which in turn interacts with the third RNA, and so on, and how every two consecutive RNAs interact, so as to maximize the total weight of the interactions (to achieve the most stable structure). Throughout this paper, we consider the uni-weight case, that is to maximize the total number of interactions. Two interactions $(R_{i_1\ell_1}, R_{i_2\ell_2})$ and $(R_{i_1k_1}, R_{i_2k_2})$ are pseudoknot-like if $\ell_1 < k_1$ but $\ell_2 > k_2$. The MRIP problem can allow or disallow pseudoknot-like interactions, depending on application details similar to RNA structure prediction.

For a very special case of MRIP (the *Pegs and Rubber Bands* problem in [1]), where the order of interacting RNAs is assumed and pseudoknot-like interactions are disallowed, Ahmed et al. proved its NP-hardness and presented a polynomial-time approximation scheme [1]. Given that predicting the interaction order is difficult, they also proposed a heuristic for the more general case with unknown interacting order but still disallowing pseudoknot-like interactions.

In this paper, we first observe that the MRIP allowing pseudoknot-like interactions and with an assumed RNA interaction order can be solved in polynomial time.

Secondly, notice that the interactions are nucleotide basepairs and thus follow the Watson–Crick basepairing rule. For four RNA sequences $R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}$, when there are possible interactions $(R_{i_1\ell_1}, R_{i_2\ell_2})$, $(R_{i_2\ell_2}, R_{i_3\ell_3})$, $(R_{i_3\ell_3}, R_{i_4\ell_4})$ (for example, if they are basepairs (A, U), (U, A), (A, U), respectively), then it is naturally to assume another possible interaction $(R_{i_1\ell_1}, R_{i_4\ell_4})$ between RNAs R_{i_1} and R_{i_4}. If the given interactions satisfy the above property then the MRIP problem is said to have the "*transitivity*" property, or *transitive* MRIP. We show that the transitive MRIP problem without an assumed RNA interaction order, either allowing or disallowing pseudoknot-like interactions, is NP-hard. On the positive side, we present a constant ratio approximation algorithm for each variant.

2. Algorithmic and hardness results

2.1. MRIP with a known RNA interaction order

In this subsection, we consider the MRIP problem with a known RNA interaction order, and we assume the order is $(R_1, R_2, \ldots, R_m)$. When disallowing pseudoknot-like interactions, Ahmed et al. [1] showed that the problem is NP-hard via a reduction from the *Longest Common Subsequence* problem.

Theorem 1. *(See [1].) The MRIP problem disallowing pseudoknot-like interactions is NP-hard.*

When allowing pseudoknot-like interactions, we first construct a graph $H = (U, F)$ where every vertex $u_{i\ell}$ corresponds to nucleotide $R_{i\ell}$ and two vertices are connected by an edge if there is a given possible interaction between them. Clearly, one can see that a matching M of graph H gives a feasible solution to the MRIP problem allowing pseudoknot-like interactions, and vice versa. Therefore, the *MRIP* problem allowing pseudoknot-like interactions can be solved in polynomial time.

2.2. The general MRIP

By general MRIP, we mean the MRIP problem in which no RNA interaction order is assumed. The possible interactions are still given for every pair of RNAs and the problem goal is to find an interaction order and to achieve the maximum number of interactions.

Theorem 2. *The general* MRIP *problem, either allowing or disallowing pseudo-knot-like interactions, is NP-hard.*

Proof. Given a 0–1 matrix $A_{m\times n}$, two consecutive 1's in a column of the matrix is said to form a *bandpass*. When counting the total number of bandpasses in the matrix, no two bandpasses in the same column are allowed to share any common 1. The *Bandpass* problem is to find a row permutation for the input matrix to achieve the maximum total number of bandpasses. Lin proved that the Bandpass problem is NP-hard via a reduction from the *Hamiltonian Path* problem [10].

Let the i-th RNA be the i-th row of matrix A, so there is a possible interaction between $R_{i_1\ell_1}$ and $R_{i_2\ell_2}$ if and only if both positions have a 1. Though such constructed RNAs and interactions are not necessarily biologically meaningful, this reduction shows the general MRIP problem is NP-hard. Furthermore, no two possible interactions between a pair of RNAs are *crossing* each other, and thus there are no pseudoknot-like interactions in this reduction. Hence, the general MRIP problem, either allowing or disallowing pseudoknot-like interactions, is NP-hard. □

Input:	m RNAs R_i, $i = 1, 2, \ldots, m$;
Output:	a permutation π of $[m]$ and interactions between RNAs $R_{\pi(i)}$ and $R_{\pi(i+1)}$, for $i = 1, 2, \ldots, m-1$

1. for each RNA pair R_i and R_j,
   1.1. construct bipartite graph $BG(i, j)$;
   1.2. compute $w(R_i, R_j)$;
2. construct edge-weighted complete graph G;
3. compute a maximum weight matching M^* of G;
4. stack RNA pairs in M^* arbitrarily to form a permutation π;
5. output π and the interactions in $w(R_{\pi(i)}, R_{\pi(i+1)})$.

Fig. 1. A high-level description of Approx I.

Given an instance I of a maximization problem Π, let $C^*(I)$ ($C(I)$, respectively) denote the value of the optimal solution (the value of the solution produced by an algorithm, respectively). The performance ratio of the algorithm on I is $\frac{C(I)}{C^*(I)}$. The algorithm is a ρ-approximation if $\inf_I \frac{C(I)}{C^*(I)} \geq \rho$, that is, it guarantees, on any instance, a solution of value at least a fraction ρ of the optimum.

Using the possible interactions between the pair of RNAs R_i and R_j, we construct a bipartite graph $BG(i, j) = (V_i \cup V_j, E(i, j))$, where the vertex subset V_i (V_j, respectively) corresponds to the set of nucleotides in R_i (R_j, respectively) and the edge set $E(i, j)$ corresponds to the set of given possible interactions between R_i and R_j. That is, $(R_{i\ell_1}, R_{j\ell_2})$ is a possible interaction if and only if there is an edge between $R_{i\ell_1}$ and $R_{j\ell_j}$ in $BG(i, j)$. One clearly sees that, when allowing pseudoknot-like interactions, the size of the maximum matching in $BG(i, j)$ is exactly the maximum total number of interactions between RNAs R_i and R_j; when pseudoknot-like interactions are not allowed, the maximum total number of interactions between RNAs R_i and R_j can be computed by a dynamic programming algorithm similar to one for computing the longest common subsequence between two given sequences. Either way, this maximum number of interactions can be computed in polynomial time, and it is set as the weight between RNAs R_i and R_j, denoted as $w(R_i, R_j)$.

We next construct an edge-weighted complete graph G, in which a vertex corresponds to an RNA sequence and the weight between two vertices (RNAs) R_i and R_j is $w(R_i, R_j)$ computed above. Since the optimal solution to the MRIP problem, either allowing or disallowing pseudoknot-like interactions, can be decomposed into two matchings in graph G by including alternate edges in the solution, the maximum weight matching M^* of graph G has a weight that is at least half of the total number of interactions in the optimal solution. It follows that this maximum weight matching-based algorithm, described in Fig. 1, is a 0.5-approximation to the MRIP problem.

Theorem 3. Approx I *is a* 0.5*-approximation algorithm for the general MRIP problem, either allowing or disallowing pseudoknot-like interactions.*

Proof. When allowing pseudoknot-like interactions, $w(R_i, R_j)$ can be computed by a maximum matching algorithm in $O(n^3)$ time, where n is the (common) length of the given RNAs.

When disallowing pseudoknot-like interactions, $w(R_i, R_j)$ can be computed by a dynamic programming algorithm in $O(n^2)$ time.

It follows that the time for constructing graph G is $O(m^2n^3)$. Graph G contains m vertices, and this its maximum weight matching M^* can be computed in $O(m^3)$ time. Subsequent construction of the solution permutation π takes linear time.

Therefore, Approx I is an $O(\max\{m^3, m^2n^3\})$-time 0.5-approximation algorithm for the MRIP problem allowing pseudoknot-like interactions. For the MRIP problem disallowing pseudoknot-like interactions, it is an $O(\max\{m^3, m^2n^2\})$-time 0.5-approximation algorithm. □

3. Better approximations for general transitive MRIP

In the previous section, we proved the NP-hardness for the general MRIP problem, and presented a 0.5-approximation algorithm. One can imagine that this performance ratio of 0.5 must be tight, if the given possible interactions are arbitrary. Consider for example an instance of m RNAs $\mathcal{R} = \{R_1, R_2, \ldots, R_m\}$ each of length n ($n \geq m$). The only possible interactions are $(R_{i,i}, R_{i+1,i})$, for $i = 1, 2, \ldots, m-1$. One can see that the optimal permutation is $\langle 1, 2, \ldots, m\rangle$, which contains $m-1$ interactions; while algorithm Approx I can produce a permutation containing only $\lfloor \frac{m}{2} \rfloor$ interactions.

Now we consider a biologically meaningful special case where the given possible interactions are *transitive*, that is, for any four RNA sequences $R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}$, when there are possible interactions $(R_{i_1\ell_1}, R_{i_2\ell_2}), (R_{i_2\ell_2}, R_{i_3\ell_3}), (R_{i_3\ell_3}, R_{i_4\ell_4})$ (for example, they are basepairs (A, U), (U, A), (A, U), respectively), then $(R_{i_1\ell_1}, R_{i_4\ell_4})$ is also a possible interaction between RNAs R_{i_1} and R_{i_4}. We call it the general transitive MRIP problem. Note that in the proof of NP-hardness in Theorem 2, the constructed instance of the MRIP problem satisfies the transitivity property. Therefore the general transitive MRIP problem, either allowing or disallowing pseudoknot-like interactions, is still NP-hard. We next show that we can explore the transitivity property to design approximation algorithms with performance ratios better than 0.5.

Input:	m RNAs R_i, $i = 1, 2, \ldots, m$, with transitivity;
Output:	a permutation π of $[m]$ and interactions between RNAs $R_{\pi(i)}$ and $R_{\pi(i+1)}$, for $i = 1, 2, \ldots, m-1$

1. for each RNA pair R_i and R_j,
   - 1.1. construct bipartite graph $BG(i, j)$;
   - 1.2. compute $w(R_i, R_j)$ disallowing pseudoknot-like interactions;
2. construct edge-weighted complete graph G using edge weight function w;
3. compute a maximum weight matching M^* of G;
   - 3.1. delete nucleotides involved in the interactions of M^*;
   - 3.2. reconstruct bipartite graph $BG(i, j)$;
   - 3.3. compute $w'(R_i, R_j)$ disallowing pseudoknot-like interactions;
4. construct edge-weighted complete graph G' using edge weight function w';
   - 4.1. compute a maximum weight 4-matching $\mathcal{C}$ of G';
   - 4.2. compute an approximate acyclic 2-matching $\mathcal{P}$ of G';
   - 4.3. compute a matching M out of $\mathcal{C}$ and $\mathcal{P}$ to extend M^*;
5. stack RNA paths in $G[M^* \cup M]$ arbitrarily to form a permutation π;
6. output π and the interactions in $w(R_{\pi(i)}, R_{\pi(i+1)}) + w'(R_{\pi(i)}, R_{\pi(i+1)})$.

Fig. 2. A high-level description of Approx II.

3.1. A 0.5328-approximation for disallowing pseudoknots

The improved approximation algorithm for the general transitive MRIP disallowing pseudoknot-like interactions is denoted as Approx II, and its high-level description is provided in Fig. 2.

Note that to compute the maximum number of interactions between two RNAs R_i and R_j (Step 1.2) while disallowing pseudoknot-like interactions, we can use the same dynamic programming algorithm used in Approx I, which runs in $O(n^2)$-time. In Step 4.2, the best approximation algorithm for the Maximum-TSP (with a performance ratio of $\frac{7}{9}$ [14]) is used to compute an acyclic 2-matching. In Step 4.3, to compute a matching M to extend M^*, the union of the edge sets of M and M^*, *i.e.* $G[M \cup M^*]$, is an acyclic 2-matching (sub-tour is an alternative terminology often used in the literature). Basically algorithm Approx II adds to the maximum weight matching M^* of graph G a subset of edges that contains a proven fraction of interactions.

Let I denote the set of interactions in the optimal solution. Let J be the set of interactions extracted from the weights of the edges in the maximum weight matching M^* of graph G. Note that neither I nor J contains crossing interactions. Similarly as in the MRIP problem with a known RNA interaction order (Section 2), we construct another graph $H = (U, F)$ for the instance where every vertex $u_{i\ell}$ corresponds to nucleotide $R_{i\ell}$ and two vertices are connected by an edge if there is a given possible interaction between them. With respect to graph H, both I and J are non-crossing matchings. Therefore, the subgraph of H induced by the interactions of I and J, $H[I \cup J]$, is a 2-matching of graph H, denoted by T. Using this 2-matching T, we partition I into 4 subsets of interactions, $I = I_1 \cup I_2 \cup I_3 \cup I_4$, and at the same time partition J into 4 subsets of interactions, $J = J_1 \cup J_2 \cup J_3 \cup J_4$.

Here is the partitioning scheme. Since T is a 2-matching, there are only alternating paths and cycles in T. First we consider paths. For a path of length 1, say $P = \langle u_1, u_2 \rangle$, if its only edge/interaction is in $I \cap J$, then the edge belongs to I_1 and belongs to J_1 too; if the edge is in $I - J$, then the edge belongs to I_4; if the edge is in $J - I$, then the edge belongs to J_4. For a path of length 3, say $P = \langle u_1, u_2, u_3, u_4 \rangle$, if $(u_1, u_2), (u_3, u_4) \in I$, then they belong to I_2 and edge (u_2, u_3) belongs to J_2. For a path other than the above cases, the edges of I all belong to I_3 and the edges of J all belong to J_3. And for each cycle, the edges of I all belong to I_3 and the edges of J all belong to J_3.

Lemma 1. *Let $|X_i|$ denote the size of, that is the number of interactions in, set X_i, for $X = I$, J and $i = 1, 2, 3, 4$. We have $|J_1| = |I_1|$, $|J_2| = \frac{1}{2}|I_2|$, and $|J_3| \geq \frac{2}{3}|I_3|$.*

Proof. By the definition of I_1, J_1, I_2, J_2, we can easily see $|J_1| = |I_1|$ and $|J_2| = \frac{1}{2}|I_2|$. For I_3 and J_3, from each path or cycle, the number of edges assigned to J_3 is either greater than or equal to the number of edges assigned to I_3, or 1 less; but in the latter case the length of the path must be at least 5. Therefore, the worst case happens when two and three edges are assigned to J_3 and I_3 respectively, which implies $|J_3| \geq \frac{2}{3}|I_3|$. □

Corollary 1. *We have the following relationships:*

$$|I| = |I_1| + |I_2| + |I_3| + |I_4|, \tag{3.1}$$

$$w(M^*) = |J_1| + |J_2| + |J_3| + |J_4|, \tag{3.2}$$

$$w(M^*) \geq |I_1| + \frac{1}{2}|I_2| + \frac{2}{3}|I_3|, \tag{3.3}$$

$$w(M^*) \geq \frac{1}{2}|I| = \frac{1}{2}\big(|I_1| + |I_2| + |I_3| + |I_4|\big). \tag{3.4}$$

Proof. The first two equations are straightforward, following the description of partitioning process and that $w(M^*) = |J|$. The last two inequalities follow from Lemma 1 and Theorem 3, respectively. □

After deleting the bases involved in the interactions of the maximum weight matching M^*, graph G' is constructed the same as graph G except using the residual weight function w'. For a path of length 3, $P = \langle u_1, u_2, u_3, u_4\rangle$, such that (u_1, u_2), $(u_3, u_4) \in I_2$, the transitivity property ensures that there is a possible interaction between u_1 and u_4. Clearly, this interaction is left in graph G', and such an interaction is called an *induced* interaction. Let G'_s be the subgraph of G' that contains exactly those edges each of which is contributed by at least one induced interaction.

Lemma 2. *G'_s is a 4-matching in G', and its weight $w'(G'_s) \geq \frac{1}{2}|I_2|$.*

Proof. To prove the first part, we only need to prove that every RNA can have induced interactions with at most 4 other RNAs. By the definition of I_2, there is an induced interaction (u_1, u_4) if and only if there is an alternating length-3 path $P = \langle u_1, u_2, u_3, u_4\rangle$, such that $(u_1, u_2), (u_3, u_4) \in I$ and $(u_2, u_3) \in J$. Suppose $u_k \in R_{i_k}$, for $k = 1, 2, 3, 4$. It follows that R_{i_1}, R_{i_2} (R_{i_3}, R_{i_4}, respectively) are adjacent in the optimal permutation and R_{i_2}, R_{i_3} are matched in M^*. Since each RNA can be adjacent to at most two other RNAs in the optimal solution, R_{i_1} and every RNA can have induced interactions with at most 4 other RNAs.

The second part of the lemma follows directly from the definition of an induced interaction, which corresponds to a distinct pair of interactions of I_2. □

It is known that in $O(n^{2.5})$ time, a 4-matching can be decomposed into two 2-matchings [6,7], and a 2-matching can be further decomposed for our purpose in the next few lemmas.

Lemma 3. *(See [3,18].) Let $\mathcal{C}$ be a 2-matching of graph G such that $M^* \cap \mathcal{C} = \emptyset$. Then, we can partition the edge set of $\mathcal{C}$ into 4 matchings X_0, X_1, X_2, X_3 each of which extends M^*. Moreover, the partitioning takes $O(n\alpha(n))$ time, where $\alpha(n)$ is the inverse Ackerman function.*

The maximum weight 4-matching $\mathcal{C}$ of graph G' can be decomposed into two 2-matchings $\mathcal{C}_1$ and $\mathcal{C}_2$. By Lemma 3, $\mathcal{C}_1$ can be partitioned into 4 matchings X_0, X_1, X_2, X_3 and $\mathcal{C}_2$ can be partitioned into 4 matchings Y_0, Y_1, Y_2, Y_3, each of which extends M^*.

Lemma 4. *(See [18].) Let $\mathcal{C}$ be a 4-matching of graph G such that $M^* \cap \mathcal{C} = \emptyset$. Then, we can partition the edge set of $\mathcal{C}$ into 8 matchings such that each of them extends M^* and the maximum weight among them is at least $\frac{2}{15}w'(\mathcal{C})$. Moreover, the partitioning takes $O(n^{2.5})$ time.*

Lemma 5. *The maximum weight acyclic 2-matching $\mathcal{D}$ of graph G' has weight $w'(\mathcal{D}) \geq |I_4|$.*

Proof. Note that graph G' contains all interactions of I_4 because the only bases deleted are those involved in the interactions of M^*. The subgraph of graph G' containing exactly the edges contributed by at least one interaction of I_4 is a subgraph of the optimal solution, and thus it is an acyclic 2-matching in graph G'. Therefore,

$$w'(\mathcal{D}) \geq |I_4|.$$

This proves the lemma. □

Lemma 6. *(See [3,18].) Let $\mathcal{P}$ be an acyclic 2-matching of G such that $M^* \cap \mathcal{P} = \emptyset$. Then, we can partition the edge set of $\mathcal{P}$ into three matchings Y_0, Y_1, Y_2 each of which extends M^*. Moreover, the partitioning takes $O(n\alpha(n))$ time.*

Lemma 7. *(See [14].) The* Max-TSP *admits an $O(n^3)$-time $\frac{7}{9}$-approximation algorithm, where n is the number of vertices in the graph.*

Corollary 2. *The weight of the second matching M in* Approx II *to extend M^* has weight $w'(M) \geq \max\{\frac{1}{15}|I_2|, \frac{7}{27}|I_4|\}$.*

Proof. Using Lemmas 2 and 4, we have

$$w'(M) \geq \frac{2}{15}w'(\mathcal{C}) \geq \frac{1}{15}|I_2|.$$

Using Lemmas 5–7, we have

$$w'(M) \geq \frac{1}{3}w'(\mathcal{P}) \geq \frac{7}{27}w'(\mathcal{D}) \geq \frac{7}{27}|I_4|.$$

The corollary holds. □

Input: m RNAs R_i, $i = 1, 2, \ldots, m$, with transitivity;
Output: a permutation π of $[m]$ and interactions between RNAs $R_{\pi(i)}$ and $R_{\pi(i+1)}$, for $i = 1, 2, \ldots, m-1$

1. for each RNA pair R_i and R_j,
   1.1. construct bipartite graph $BG(i, j)$;
   1.2. compute $w(R_i, R_j)$ allowing pseudoknot-like interactions;
2. construct edge-weighted complete graph G using edge weight function w;
3. compute a maximum weight matching M^* of G;
   3.1. delete nucleotides involved in the interactions of M^*;
   3.2. reconstruct bipartite graph $BG(i, j)$;
   3.3. compute $w'(R_i, R_j)$ allowing pseudoknot-like interactions;
4. construct edge-weighted complete graph G' using edge weight function w';
   4.1. compute a maximum weight 4-matching $\mathcal{C}$ of G';
   4.2. compute a matching M out of $\mathcal{C}$ to extend M^*;
5. stack RNA paths in $G[M^* \cup M]$ arbitrarily to form a permutation π;
6. output π and the interactions in $w(R_{\pi(i)}, R_{\pi(i+1)}) + w'(R_{\pi(i)}, R_{\pi(i+1)})$.

Fig. 3. A high-level description of Approx III.

Theorem 4. *Algorithm* Approx II *is a* 0.5328*-approximation for the general transitive MRIP problem disallowing pseudoknot-like interactions.*

Proof. Combining Corollaries 1 and 2, we have for any real $x, y \in [0, 1]$,

$$\begin{aligned}
w(\pi) &= w\big(M^*\big) + w'(M) \\
&\geq x\left(|I_1| + \frac{1}{2}|I_2| + \frac{2}{3}|I_3|\right) + (1-x)\frac{1}{2}\big(|I_1| + |I_2| + |I_3| + |I_4|\big) + y\frac{1}{15}|I_2| + (1-y)\frac{7}{27}|I_4| \\
&= \frac{1+x}{2}|I_1| + \frac{15+2y}{30}|I_2| + \frac{3+x}{6}|I_3| + \frac{41-27x-14y}{54}|I_4| \\
&\geq \frac{255}{426}|I_1| + \frac{227}{426}|I_2| + \frac{227}{426}|I_3| + \frac{227}{426}|I_4| \\
&\geq \frac{227}{426}|I| \\
&> 0.5328|I|,
\end{aligned}$$

where the third last inequality holds by setting $x = \frac{14}{71}$ and $y = \frac{35}{71}$. □

3.2. A 0.5333-approximation for allowing pseudoknots

The improved approximation algorithm for the general transitive MRIP allowing pseudoknot-like interactions is denoted as Approx III, and its high-level description is provided in Fig. 3.

Approx III is very similar to Approx II, and only differs at two places. First, since the problem allows pseudoknot-like interactions, we run a maximum weight bipartite matching algorithm to compute those edge weights, in Steps 1.2 and 3.3. Second, computing a matching M to extend M^* is now based only on the maximum weight 4-matching $\mathcal{C}$, of which the weight can be better estimated because pseudoknot-like interactions are allowed.

The analysis of the algorithm is similar to that of the previous section. We do exactly the same interaction partitioning for the optimal solution, of which the interaction set is I, and the maximum weight matching M^*, of which the interaction set is J. One can easily verify that all Lemma 1, Corollary 1, and Lemma 2 hold again. The following lemma is the key to the improvement of the performance analysis, which provides an improved lower bound on the weight of the maximum weight 4-matching.

Lemma 8. *The weight of the maximum weight* 4*-matching* $\mathcal{C}$ *of graph* G' *is*

$$w'(\mathcal{C}) \geq \max\left\{\frac{1}{2}|I_2|, \frac{1}{4}|I_2| + |I_4|\right\}. \tag{3.5}$$

Proof. The first component straightly follows from Lemma 2 since G'_s is a 4-matching in graph G'. Note also that graph G' contains all the edges of the optimal solution, each of which is contributed by at least one interaction of I_4. This remaining optimal solution, denoted as $\mathcal{P}$, is an acyclic 2-matching in G', and has weight $w'(\mathcal{P}) \geq |I_4|$.

Since G'_s is a 4-matching, it can be decomposed into two 2-matchings denoted as $\mathcal{D}_1$ and $\mathcal{D}_2$. One clearly see that both $\mathcal{P} \cup \mathcal{D}_1$ and $\mathcal{P} \cup \mathcal{D}_2$ are 4-matchings in graph G'. The interactions of I_4 counted toward $\mathcal{P}$ are not counted toward G'_s. Therefore, we have

$$\begin{aligned} w'(\mathcal{C}) &\geq \max\{w'(\mathcal{P}\cup\mathcal{D}_1), w'(\mathcal{P}\cup\mathcal{D}_2)\} \\ &\geq \frac{1}{2}\big(w'(\mathcal{D}_1)+w'(\mathcal{D}_2)\big)+w'(\mathcal{P}) \\ &= \frac{1}{2}w'\big(G'_s\big)+|I_4| \\ &\geq \frac{1}{4}|I_2|+|I_4|. \end{aligned}$$

This proves the lemma. □

Corollary 3. *The weight of the second matching M in* Approx III *to extend M^* has weight $w'(M) \geq \frac{1}{30}|I_2| + \frac{2}{15}|I_4|$.*

Proof. Using Lemmas 2 and 8, we have

$$w'(M) \geq \frac{2}{15}w'(\mathcal{C}) \geq \frac{1}{30}|I_2| + \frac{2}{15}|I_4|.$$

The corollary holds. □

Theorem 5. *Algorithm* Approx III *is a 0.5333-approximation for the general transitive MRIP problem allowing pseudoknot-like interactions.*

Proof. Combining Corollaries 1 and 3, we have for any real $x \in [0,1]$,

$$\begin{aligned} w(\pi) &= w\big(M^*\big)+w'(M) \\ &\geq x\left(|I_1|+\frac{1}{2}|I_2|+\frac{2}{3}|I_3|\right)+(1-x)\frac{1}{2}\big(|I_1|+|I_2|+|I_3|+|I_4|\big)+\frac{1}{30}|I_2|+\frac{2}{15}|I_4| \\ &= \frac{1+x}{2}|I_1|+\frac{8}{15}|I_2|+\frac{3+x}{6}|I_3|+\frac{19-15x}{30}|I_4| \\ &\geq \frac{3}{5}|I_1|+\frac{8}{15}|I_2|+\frac{8}{15}|I_3|+\frac{8}{15}|I_4| \\ &\geq \frac{8}{15}|I| \\ &> 0.5333|I|, \end{aligned}$$

where the third last inequality holds by setting $x = \frac{1}{5}$. □

Acknowledgements

Weitian Tong, Randy Goebel, and Guohui Lin are supported in part by NSERC, AITF and iCORE. Weitian Tong, Tian Liu, and Guohui Lin thank the Open Fund of Top Key Discipline of Computer Software and Theory in Zhejiang Provincial Colleges at the Zhejiang Normal University for sponsoring a workshop where this work was started.

References

[1] S. Ahmed, S. Mneimneh, N. Greenbaum, A combinatorial approach for multiple RNA interaction: formulations, approximations, and heuristics, in: Computing and Combinatorics, in: LNCS, vol. 7936, 2013, pp. 421–433.
[2] C. Alkan, E. Karakoç, J.H. Nadeau, S.C. Sahinalp, K. Zhang, RNA–RNA interaction prediction and antisense RNA target search, J. Comput. Biol. 13 (2006) 267–282.
[3] Z.-Z. Chen, L. Wang, An improved approximation algorithm for the bandpass-2 problem, in: Combinatorial Optimization and Applications, in: LNCS, vol. 7402, 2012, pp. 188–199.
[4] H. Chitsaz, R. Backofen, S.C. Sahinalp, biRNA: fast RNA–RNA binding sites prediction, in: Workshop on Algorithms in Bioinformatics, in: LNCS/LNBI, vol. 5724, 2009, pp. 25–36.
[5] H. Chitsaz, R. Salari, S.C. Sahinalp, R. Backofen, A partition function algorithm for interacting nucleic acid strands, Bioinformatics 25 (2009) 365–373.
[6] R. Diestel, Graph Theory, Graduate Texts in Mathematics, Springer, 2005.
[7] F. Harary, Graph Theory, Addison-Wesley, 1969.
[8] F.W.D. Huang, J. Qin, C.M. Reidys, P.F. Stadler, Partition function and base pairing probabilities for RNA–RNA interaction prediction, Bioinformatics 25 (2009) 2646–2654.
[9] A.X. Li, M. Marz, J. Qin, C.M. Reidys, RNA–RNA interaction prediction based on multiple sequence alignments, Bioinformatics 27 (2011) 456–463.
[10] G. Lin, On the Bandpass problem, J. Comb. Optim. 22 (2011) 71–77.
[11] I.M. Meyer, Predicting novel RNA–RNA interactions, Current Opinion Struct. Biology 18 (2008) 387–393.
[12] S. Mneimneh, On the approximation of optimal structures for RNA–RNA interaction, IEEE/ACM Trans. Comput. Biology and Bioinformatics 6 (2009) 682–688.

[13] U. Mückstein, H. Tafer, J. Hackermüller, S.H. Bernhart, P.F. Stadler, I.L. Hofacker, Thermodynamics of RNA–RNA binding, Bioinformatics 22 (2006) 1177–1182.
[14] K. Paluch, M. Mucha, A. Madry, A 7/9-approximation algorithm for the maximum traveling salesman problem, in: Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques, in: LNCS, vol. 5687, 2009, pp. 298–311.
[15] D.D. Pervouchine, Iris: intermolecular RNA interaction search, Genome Inform. 15 (2004) 92–101.
[16] R. Salari, R. Backofen, S.C. Sahinalp, Fast prediction of RNA–RNA interaction, Algorithms Molecular Biology 5 (5) (2010).
[17] J.S. Sun, J.L. Manley, A novel U2–U6 snRNA structure is necessary for mammalian mRNA splicing, Genes Dev. 9 (1995) 843–854.
[18] W. Tong, Z.-Z. Chen, L. Wang, Y. Xu, J. Xu, R. Goebel, G. Lin, An approximation algorithm for the bandpass-2 problem, arXiv:1307.7089, July 2013.