A Combinatorial Approach for Multiple RNA Interaction: Formulations, Approximations, and Heuristics

Syed Ali Ahmed*, Saad Mneimneh**,***,†, and Nancy L. Greenbaum‡

The Graduate Center and Hunter College, City University of New York (CUNY), New York, USA
sahmed3@gc.cuny.edu,{saad,ngreenba}@hunter.cuny.edu

Abstract. The interaction of two RNA molecules involves a complex interplay between folding and binding that warranted recent developments in RNA-RNA interaction algorithms. However, biological mechanisms in which more than two RNAs take part in an interaction exist.

We formulate multiple RNA interaction as a computational problem, which not surprisingly turns out to be NP-complete. Our experiments with approximation algorithms and heuristics for the problem suggest that this formulation is indeed useful to determine interaction patterns of multiple RNAs when information about which RNAs interact is not necessarily available (as opposed to the case of two RNAs where one must interact with the other), and because the resulting RNA structure often cannot be predicated by existing algorithms when RNAs are simply handled in pairs. We show instances of multiple RNA interaction that are accurately predicted by our algorithms.

Keywords: multiple RNA interaction, dynamic programming, approximation algorithms, structure prediction.

1 Introduction

The interaction of two RNA molecules has been independently formulated as a computational problem in several works, e.g. [1,2,3]. In their most general form, these formulations lead to NP-hard problems. To overcome this hurdle, researchers have been either reverting to approximation algorithms, or imposing algorithmic restrictions; for instance, analogous to the avoidance of pseudoknot formation in the folding of RNAs.

While these algorithms had limited use in the beginning, they became important venues for (and in fact popularized) an interesting biological fact: RNAs

* Supported by NSF Award CCF-AF 1049902 and a CUNY GC Science Fellowship.

** Supported by NSF Award CCF-AF 1049902 and in part by NIH Award GM07800.

*** Corresponding author.

† The first two authors contributed equally to this work.

‡ Supported by NSF Award MCB 0929394.

D.-Z. Du and G. Zhang (Eds.): COCOON 2013, LNCS 7936, pp. 421–433, 2013.
© Springer-Verlag Berlin Heidelberg 2013

interact. For instance, micro-RNAs (miRNAs) bind to a complementary part of messenger RNAs (mRNAs) and inhibit their translation [4]. One might argue that such a simple interaction does not present a pressing need for RNA-RNA interaction algorithms; however, more complex forms of RNA-RNA interaction exist. In E. Coli, CopA binds to the ribosome binding site of CopT, also as a regulation mechanism to prevent translation [5]; so does OxyS to fhlA [6]. In both of these structures, the simultaneous folding (within the RNA) and binding (to the other RNA) are non-trivial to be predicted as separate events. To account for this, most of the RNA-RNA interaction algorithms calculate the probability for a pair of subsequences (one of each RNA) to participate in the interaction, and in doing so they generalize the energy model used for the partition function of a single RNA to the case of two RNAs [7,8,9,10,11,12]. This generalization takes into consideration the simultaneous aspect of folding and binding.

Not surprisingly, there exist other mechanisms in which more than two RNA molecules take part in an interaction. Typical scenarios involve the interaction of multiple small nucleolar RNAs (snoRNAs) with ribosomal RNAs (rRNAs) in guiding the methylation of the rRNAs [4], and multiple small nuclear RNAs (snRNA) with mRNAs in the splicing of introns [13]. Even with the existence of a computational framework for a single RNA-RNA interaction, it is reasonable to believe that interactions involving multiple RNAs are generally more complex to be treated pairwise. In addition, given a pool of RNAs, it is not trivial to predict which RNAs interact without some prior biological information.

We formulate the problem of multiple RNA interaction by bringing forward an optimization perspective where each part of an RNA will contribute certain weights to the entire interaction when binding to different parts of other RNAs. We seek to maximize the total weight. This notion of weight can be obtained by using existing RNA-RNA interaction algorithms on pairs of RNAs. We call our formulation the Pegs and Rubber Bands problem. We show that under certain restrictions, which are similar to those against pseudoknots, the problem remains NP-hard (in fact it becomes equivalent to a special instance of the interaction of two RNAs). We describe a polynomial time approximation scheme PTAS for the problem, some heuristics, and experimental results. For instance, given a pool of RNAs in which interactions between pairs of RNAs are known, our algorithm is capable of identifying those pairs and predicting satisfactorily the pattern of interaction between them [8]. Moreover, our algorithm finds the correct interaction of a given instance of splicing consisting of two snRNAs (a modified U2-U6 human snRNA complex) and two structurally autonomous parts of an intron [14], a total of four RNAs. When (partially) mixing the two examples in one pool, our algorithm structurally separates them.

2 Pegs and Rubber Bands: A Formulation

We introduce an optimization problem we call Pegs and Rubber Bands that will serve a base framework for the multiple RNA interaction problem. The link between the two problems will be made shortly after the description of Pegs and Rubber Bands.

Consider m levels numbered 1 to m with n_l pegs in level l numbers 1 to n_l. There is an infinite supply of rubber bands that can be placed around two pegs in consecutive levels. For instance, we can choose to place a rubber band around peg i in level l and peg j in level $l+1$; we call it a rubber band at $[l, i, j]$. Every such pair of pegs $[l, i]$ and $[l+1, j]$ contribute their own weight $w(l, i, j)$. The Pegs and Rubber Bands problem is to maximize the total weight by placing rubber bands around pegs in such a way that no two rubber bands intersect. In other words, each peg can have at most one rubber band around it, and if a rubber band is placed at $[l, i_1, j_1]$ and another at $[l, i_2, j_2]$, then $i_1 < i_2 \Leftrightarrow j_1 < j_2$. We assume without loss of generality that $w(l, i, j) \neq 0$ to avoid the unnecessary placement of rubber bands and, therefore, either $w(l, i, j) > 0$ or $w(l, i, j) = -\infty$. Figure 1 shows an example.

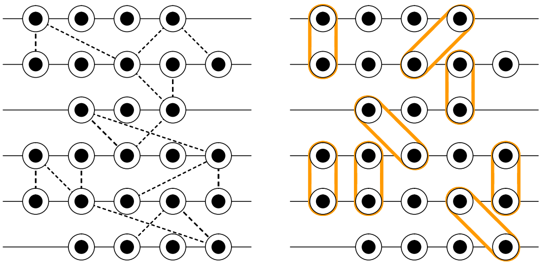

Fig. 1. Pegs and Rubber Bands. All positive weights are equal to 1 and are represented by dashed lines. The optimal solution achieves a total weight of 8.

Given an optimal solution, it can always be reconstructed from left to right by repeatedly placing some rubber band at $[l, i, j]$ such that, at the time of this placement, no rubber band is around peg $[l, k]$ for $k > i$ and no rubber band is around peg $[l+1, k]$ for $k > j$. This process can be carried out by a dynamic programming algorithm to compute the maximum weight (in exponential time), by defining $W(i_1, i_2, \ldots, i_m)$ to be the maximum weight when we truncate the levels at pegs $[1, i_1], [2, i_2], \ldots, [m, i_m]$ (see Figure 2). The maximum weight is given by $W(n_1, n_2, \ldots, n_m)$ and the optimal solution can be obtained by standard backtracking. When all levels have $O(n)$ pegs, this algorithm runs in $O(mn^m)$ time and $O(n^m)$ space.

2.1 Multiple RNA Interaction as Pegs and Rubber Bands

To provide some initial context we now describe how the formulation of Pegs and Rubber Bands, though in a primitive way, captures the problem of multiple RNA interaction. We think of each level as an RNA and each peg as one base of the RNA. The weight $w(l, i, j)$ corresponds to the negative of the energy

$$W(i_1, i_2, \ldots, i_m) = \max \begin{cases} W(i_1 - 1, i_2, \ldots, i_m) \\ W(i_1, i_2 - 1, i_3, \ldots, i_m) \\ \vdots \\ W(i_1, \ldots, i_{m-1}, i_m - 1) \\ \\ W(i_1 - 1, i_2 - 1, i_3, \ldots, i_m) + w(1, i_1, i_2) \\ W(i_1, i_2 - 1, i_3 - 1, i_4, \ldots, i_m) + w(2, i_2, i_3) \\ \vdots \\ W(i_1, \ldots, i_{m-2}, i_{m-1} - 1, i_m - 1) + w(m - 1, i_{m-1}, i_m) \end{cases}$$

where $W(0, 0, \ldots, 0) = 0$.

Fig. 2. Dynamic programming algorithm for Pegs and Rubber Bands

contributed by the binding of the i^{th} base of RNA l to the j^{th} base of RNA $l+1$. It should be clear, therefore, that an optimal solution for Pegs and Rubber Bands represents the lowest energy conformation in a base-pair energy model, when a pseudoknot-like restriction is imposed on the RNA interaction (rubber bands cannot intersect). In doing so, we obviously assume that an order on the RNAs is given with alternating sense and antisense, and that the first RNA interacts with the second RNA, which in turn interacts with the third RNA, and so on. We later relax this ordering and condition on the interaction pattern of the RNAs. While a simple base-pairing model is not likely to give realistic results, our goal for the moment is simply to establish a correspondence between the two problems.

2.2 Complexity of the Problem and Approximations

With the above correspondence in mind, the problem of Pegs and Rubber Bands can be viewed as a instance of a classical RNA-RNA interaction, involving only two RNAs that is: We construct the first as RNA 1 followed by RNA 4 reversed followed by RNA 5 followed by RNA 8 reversed and so on, and the second as RNA 2 followed by RNA 3 reversed followed by RNA 6 followed by RNA 7 reversed and so on, as shown in Figure 3.

Therefore, Pegs and Rubber Bands can be solved as an RNA-RNA interaction problem. While this RNA-RNA interaction represents a restricted instance of the more general NP-hard problem, it is still NP-hard. In fact, Pegs and Rubber Bands itself is NP-hard.

Theorem 1. *Pegs and Rubber Bands is NP-hard.*

Proof: We make a reduction from the longest common subsequence (LCS) for a set of binary strings, which is an NP-hard problem. In this reduction, pegs are labeled and $w(l, i, j)$ depends only on the label of peg $[l, i]$ and the label of peg $[l+1, j]$. We describe this weight as a function of labels shortly. Each binary string is modified by adding the symbol b between every two consecutive bits. A string of original length n is then transformed into two consecutive (identical) levels of $2n - 1$ pegs each, where each peg is labeled by the corresponding symbol in

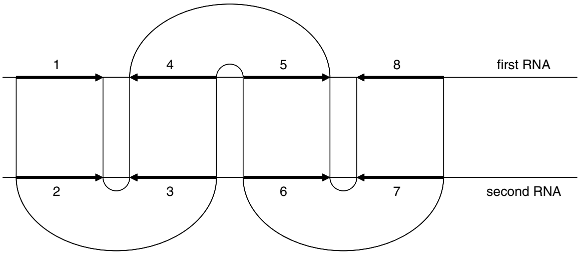

Fig. 3. Pegs and Rubber Bands as a special instance of RNA-RNA interaction, vertical lines indicate regions where only interaction (binding of the two RNAs) is allowed, and curved lines indicate regions where only folding within each RNA is allowed

$\{0, 1, b\}$. For any given integer k, the first and last levels consist of k pegs labeled $*$. We now define the weight as a function of labels: $w(0,0) = w(1,1) = w(b,b) = w(*,0) = w(*,1) = w(0,*) = w(1,*) = 1$ and $w(x,y) = -\infty$ otherwise. It is easy to verify that the strings have a common subsequence of length k if and only if the optimal solution has a weight of $\sum_i (2n_i - 1) + k = 2\sum_i n_i - m + k$ (when every peg has a rubber band around it), where n_i is the original length of string i and m is the number of strings. ■

```
    * * * *
    | | | |
  0b0b1b0b1b1b1
  ||  | | | ||||
  0b0b1b0b1b1b1
    | | | |
    0b1b0b1b0
     | | | ||
    0b1b0b1b0
    | | | |
1b0b0b1b0b1
|||| | | |
1b0b0b1b0b1
    | | | |
    * * * *
```

Fig. 4. Reduction from LCS for $\{0010111, 01010, 100101\}$ to Pegs and Rubber Bands (the symbol $|$ denotes a rubber band). The optimal solution with weight $2(7+5+6) - 3 + 4 = 37$ corresponds to a common subsequence of length 4, namely 0101.

While our problem is NP-hard, we can show that the same formulation can be adapted to obtain a polynomial time approximation. A maximization problem admits a polynomial time approximation scheme (PTAS) iff for every fixed $\epsilon > 0$ there is an algorithm with a running time polynomial in the size of the input that finds a solution within $(1-\epsilon)$ of optimal [15]. We show below that we can find a solution within $(1-\epsilon)$ of optimal in time $O(m\lceil\frac{1}{\epsilon}\rceil n^{\lceil\frac{1}{\epsilon}\rceil})$, where m is the number of levels and each level has $O(n)$ pegs.

Theorem 2. *Pegs and Rubber Bands admits a PTAS.*

Proof: Let OPT be the weight of the optimal solution and denote by $W[i\dots j]$ the weight of the optimal solution when the problem is restricted to levels $i, i+1, \dots, j$ (a sub-problem). For a given $\epsilon > 0$, let $k = \lceil\frac{1}{\epsilon}\rceil$. Consider the following k solutions (weights), each obtained by a concatenation of optimal solutions for sub-problems consisting of at most k levels.

$$W_1 = W[1\dots 1] + W[2\dots k+1] + W[k+2\dots 2k+1] + \dots$$

$$W_2 = W[1\dots 2] + W[3\dots k+2] + W[k+3\dots 2k+2] + \dots$$

$$\vdots$$

$$W_k = W[1\dots k] + W[k+1\dots 2k] + W[2k+1\dots 3k] + \dots$$

While each $W_i \leq OPT$, it is easy to verify that every pair of consecutive levels appear in exactly $k-1$ of the above sub-problems. Therefore,

$$\sum_{i=1}^{k} W_i \geq (k-1)OPT$$

$$\Rightarrow \max_i W_i \geq \frac{k-1}{k}OPT \geq (1-\epsilon)OPT$$

If m is the total number of levels, then there are $O(m)$ sub-problems of at most k levels each and, therefore, the running time required to find $\max_i W_i$ when every level has $O(n)$ pegs is $O(mkn^k) = O(m\lceil\frac{1}{\epsilon}\rceil n^{\lceil\frac{1}{\epsilon}\rceil})$. ■

For a given integer k, the $(1-1/k)$-factor approximation algorithm is to simply choose the best $W_i = W[1\dots i] + W[i+1\dots i+k] + W[i+k+1\dots i+2k] + \dots$ as a solution, where $W[i\dots j]$ denotes the weight of the optimal solution for the sub-problem consisting of levels $i, i+1\dots, j$. However, as a practical step, and instead of using the W_i's for the comparison, we can fill in for each W_i some additional rubber bands (interactions) between (RNAs) level i and level $i+1$, between level $i+k$ and level $i+k+1$, and so on, by identifying the pegs of these levels (regions of RNAs) that are not part of the solution. This does not affect the theoretical guarantee but gives a larger weight to the solution. We call it *gap filling.*

3 Windows and Gaps: A Better Formulation for RNA Interaction

In the previous section, we described our initial attempt to view the interaction of m RNAs as a Pegs and Rubber Bands problem with m levels, where the first

RNA interacts with the second RNA, and the second with the third, and so on (so they alternate in sense and antisense). This used a simple base-pair energy model, which is not too realistic. We now address this issue (and leave the issues of the ordering and the interaction pattern to Section 3.3). A better model for RNA interaction will consider windows of interaction instead of single bases. In terms of our Pegs and Rubber Bands problem, this translates to placing rubber bands around a stretch of contiguous pegs in two consecutive levels, e.g. around pegs $[l, i_1]$, $[l, i_2]$, $[l+1, j_1]$, and $[l+1, j_2]$, where $i_2 \geq i_1$ and $j_2 \geq j_1$. The weight contribution of placing such a rubber band is now given by $w(l, i_2, j_2, u, v)$, where i_2 and j_2 are the last two pegs covered by the rubber band in level l and level $l+1$, and $u = i_2 - i_1 + 1$ and $v = j_2 - j_1 + 1$ represent the length of the two windows covered in level l and level $l+1$, respectively.

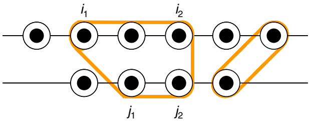

Fig. 5. A rubber band can now be placed around a window of pegs, here $u = 3$ and $v = 2$ in the big window.

As a *heuristic*, we also allow for the possibility of imposing a gap $g \geq 0$ between windows to establish a distance at which windows may be considered energetically separate. This gap is also taken into consideration when we perform the gap filling procedure described at the end of Section 3.1. The modified algorithm is shown in Figure 6, and if we set $u = v = 1$ and $g = 0$, then we retrieve the original algorithm of Figure 2.

$$W(i_1, i_2, \ldots, i_m) = \max \begin{cases} W(i_1 - 1, i_2, \ldots, i_m) \\ W(i_1, i_2 - 1, i_3, \ldots, i_m) \\ \vdots \\ W(i_1, \ldots, i_{m-1}, i_m - 1) \\ W((i_1 - u - g)^+, (i_2 - v - g)^+, i_3, \ldots, i_m) + w(1, i_1, i_2, u, v) \\ W(i_1, (i_2 - u - g)^+, (i_3 - v - g)^+, i_4, \ldots, i_m) + w(2, i_2, i_3, u, v) \\ \vdots \\ W(i_1, \ldots, i_{m-2}, (i_{m-1} - u - g)^+, (i_m - v - g)^+) + \\ w(m-1, i_{m-1}, i_m, u, v) \end{cases}$$

where x^+ denotes $\max(0, x)$, $w(l, i, j, u, v) = -\infty$ if $u > i$ or $v > j$, $0 < u, v \leq w$ (the maximum window size), $g \geq 0$ (the gap), and $W(0, 0, \ldots, 0) = 0$.

Fig. 6. Modified dynamic programming algorithm for Pegs and Rubber Bands with the windows and gaps formulation

The running time of the modified algorithm is $O(mw^2n^m)$ and $O(mw^2\lceil\frac{1}{\epsilon}\rceil n^{\lceil\frac{1}{\epsilon}\rceil})$ for the approximation scheme, where w is the maximum window length. If we impose that $u = v$, then those running times become $O(mwn^m)$ and $O(mw\lceil\frac{1}{\epsilon}\rceil n^{\lceil\frac{1}{\epsilon}\rceil})$ respectively.

For the correctness of the algorithm, we now have to assume that windows are *sub-additive*. In other words, we require the following condition (otherwise, the algorithm may compute an incorrect optimum due to the possibility of achieving the same window by two or more smaller ones with higher total weight):

$$w(l,i,j,u_1,v_1) + w(l,i-u_1,j-v_1,u_2,v_2)$$

$$\leq w(l,i,j,u_1+u_2,v_1+v_2)$$

In our experience, most existing RNA-RNA interaction algorithms produce weights (the negative of the energy values) of RNA interaction windows that mostly conform to the above condition. In rare cases, we filter the windows to eliminate those that are not sub-additive. For instance, if the above condition is not met, we set $w(l,i,j,u_1,v_1) = w(l,i-u_1,j-v_1,u_2,v_2) = -\infty$ (recursively starting with smaller windows).

4 Interaction Pattern and Permutations: A Heuristic

We now describe how to relax the ordering and the condition on the interaction pattern of the RNAs. We first identify each RNA as being *even* (sense) or *odd* (antisense), but this convention can obviously be switched. Given m RNAs and a permutation on the set $\{1,\ldots,m\}$, we map the RNAs onto the levels of a Pegs and Rubber Bands problem as follows: We place the RNAs in the order in which they appear in the permutation on the same level as long as they have the same parity (they are either all even or all odd). We then increase the number of levels by one, and repeat. RNAs that end up on the same level are *virtually* considered as one RNA that is the concatenation of all. However, in the corresponding Pegs and Rubber Bands problem, we do not allow windows to span multiple RNAs, nor do we enforce a gap between two windows in different RNAs. For example, if we consider the following permutation of RNAs $\{1,3,4,7,5,8,2,6\}$, where the RNA number also indicates its parity (for the sake of illustration), then we end up with the following placement: RNA 1 and RNA 3 in that order on the first level, followed by RNA 4 on the second level, followed by RNA 7 and RNA 5 in that order on the third level, followed by RNA 8, RNA 2, and RNA 6 in that order on the fourth level, resulting in four virtual RNAs on four levels of pegs as shown in Figure 7.

But what is the best placement as a Pegs and Rubber Bands problem for a given set of RNAs? Figure 8 shows a possible (greedy) heuristic that tackles this question by starting with a random permutation and then searching for the best one via neighboring permutations (and recall that the permutation uniquely determines the placement).

```
        ---RNA 1---RNA 3---

            ---RNA 4---

        ---RNA 7---RNA 5---

  ---RNA 8---RNA 2---RNA 6---
```

Fig. 7. Placement of the permutation $\{1, 3, 4, 7, 5, 8, 2, 6\}$ where the RNA number also indicates its parity. The interaction pattern is less restrictive then before; for instance, RNA 7 can interact with RNA 2, RNA 4, RNA 6, and RNA 8.

```
Given $\epsilon = 1/k$ and $m$ RNAs
   produce a random permutation $\pi$ on $\{1, \ldots, m\}$
   let $W$ be the weight of the $(1 - \epsilon)$-optimal solution given $\pi$
   **repeat**
      better←false
      generate a set $\Pi$ of neighboring permutations for $\pi$
      **for** every $\pi' \in \Pi$ (in any order)
         let $W'$ be the weight of the $(1 - \epsilon)$-optimal solution given $\pi'$
         **if** $W' > W$
            **then** $W \leftarrow W'$
                  $\pi \leftarrow \pi'$
                  better←true
   **until not** better
```

Fig. 8. A heuristic for multiple RNA interaction using the PTAS algorithm

To generate neighboring permutations for this heuristic algorithm one could adapt a standard 2-opt method used in the Traveling Salesman Problem (or other techniques). For instance, given permutation π, a neighboring permutation π' can be obtained by dividing π into three parts and making π' the concatenation of the first part, the reverse of the second part, and the third part. In other words, if $\pi = (\alpha, \beta, \gamma)$, then $\pi' = (\alpha, \beta^R, \gamma)$ is a neighbor of π, where β^R is the reverse of β.

5 Experimental Results

We apply the algorithm of Section 3.3 using the 2-opt method, where the PTAS is based on the Windows and Gaps formulation of Section 3.2, with windows satisfying $2 \le u, v \le w = 26$ (RNAup's default [7]) and a gap $g = 4$. The weights $w(l, i, j, u, v)$ are obtained from RNAup as (negative of energy values):

$$w(l, i, j, u, v) \propto \log p_l(i - u + 1, i) + \log p_{l+1}(j - v + 1, j)$$

$$+ \log Z_l^I(i - u + 1, i, j - v + 1, j)$$

where $p_l(i_1, i_2)$ is the probability that subsequence $[i_1, i_2]$ is free (does not fold) in RNA l, and $Z_l^I(i_1, i_2, j_1, j_2)$ is the partition function of the interaction of subsequences $[i_1, i_2]$ in RNA l and $[j_1, j_2]$ in RNA $l+1$ (subject to no folding within RNAs).

The windows are filtered for sub-additivity as described in Section 3.2. We impose the condition that $u = v$ for every window. We also have the option to compress RNAs on level l by removal of a base i whenever $w(l, i, j, u, u)$ is less than some threshold for every j and every u; however, peg $[l, i]$ can still be part of some window, e.g. if $w(l, i+x, j, x+y, x+y)$ is added to the solution, where $x, y > 0$. We did not use that option here. We pick the largest weight solution among several runs of the algorithm. The value of k and the gap filling criterion depend on the scenario, as described below.

5.1 Fishing for Pairs

Six RNAs of which three pairs are known to interact are used [8]. We are interested in identifying the three pairs. For this purpose, it will suffice to set $k = 2$ and to ignore gap filling. Furthermore, we only consider solutions in which each RNA interacts with at most one other RNA. The solution with the largest weight identifies the three pairs correctly (Figure 9). In addition, the interacting sites in each pair are consistent (not surprisingly) with the predictions of existing RNA-RNA interaction algorithms, e.g. [10].

```
              OxyS   5' ...CCCUUG...  ...GUG...  ...UCCAG... 3'
                           ||||||        |||        |||||
              fhlA   3' ...GGGAAC...  ...CAC...  ...AGGUC... 5'


CopA 5' ...CGGUUUAAGUGGG... ...UUUCGUACUCGCCAAAGUUGAAGA... ...UUUUGCUU 3'
           |||||||||||||       ||||||||||||||||||||||||       ||||||||
CopT  3' GCCAAAUUCACCC...  ...AAAGCAUGAGCGGUUUCAACUUCU...  ...AAAACGAA 5'


             MicA   5' ...GCGCA...  ...CUGUUUUC...  ...CGU... 3'
                          |||||        ||||||||        |||
             lamB   3' ...CGCGU...  ...GAUAGAGG...  ...GCA... 5'
```

Fig. 9. Known pairs of interacting RNAs with reasonable solutions

5.2 Structure Prediction

The human snRNA complex U2-U6 is necessary for the splicing of a specific mRNA intron [14]. Only the preserved regions of the intron are considered, which consist of two structurally autonomous parts, resulting in an instance with a total of four RNAs. The algorithm is performed with $k = 2, 3, 4$ and gap filling. In all three cases, the solution with the largest weight consistently finds the structure shown in Figure 10. This structure reveals a correct pattern described in [13,14], and cannot be easily predicted by considering the RNAs in pairs; for instance, AUAC in U6 will bind to UAUG in both U2 and I1, and it is not immediately obvious which one to break without a global view, e.g. that AUGAU in U2 binds with UACUA in I2 as well. This is a typical issue of using local information to produce a globally optimal solution.

```
I1  3' UGUAUG 5'
         ||||
U6    5' AUACA.....GAUUa...      ...cGUGAAGCGU 3'
                   ||||              |||||||||
U2   3' UAUGAUg....CUAGAAu..........gCACUUCGCA 5'
         |||||
I2    5' UACUAAc 3'
```

Fig. 10. A modified human snRNA U2-U6 complex in the splicing of an intron, as reported in [14]. Bases indicated by small letters are missing from the interaction. From left to right: g-c and a-u are missing due to the condition $2 \leq u = v \leq 26$, but also due to the added instability of a bulge loop when this condition is relaxed; c-g ends up being not favored by RNAup. I1 is shifted (UGU should interact with ACA instead) but this is a computational artifact of optimization that is hard to avoid. Overall, the structure is accurate and cannot be predicted by a pairwise handling of the RNAs.

5.3 Structural Separation

Six RNAs are used: CopA, CopT, and the four RNAs of the previous scenario. The algorithm is performed with $k = 3$ and gap filling. The solution with the largest weight results in a successful prediction that separates the RNA complex CopA-CopT of Figure 9 from the RNA structure shown in Figure 10.

5.4 Making Improvements

In this section, we try to eliminate some heuristics, an approach we did not attempt in a previous work on the subject [16]. We relax the condition that $u = v$ so we allow arbitrary window sizes and, furthermore, we drop the gap heuristic so we set $g = 0$. Some unwanted interactions now start to appear.

To correct for this, we modify RNAup weights in a reasonable way. For simplicity of notation, let A denote the subsequence $[i - u + 1, i]$ in RNA l and B the subsequence $[j - v + 1, j]$ in RNA $l + 1$. We now have

$$w(l, i, j, u, v) \propto \log p_A + \log p_B + \log q_A + \log q_B - \log(1 - p^I_{AB})$$

where p_A and p_B are as before the probabilities that the corresponding subsequences are free, Z^I is replaced by $(1 - p^I_{AB})^{-1}$, and p^I_{AB} is the probability that the two subsequences will interact (as opposed to individually fold) given they are free (in the following, Z_A is the partition function for folding subsequence A).

$$p^I_{AB} = \frac{Z^I_{AB}}{Z^I_{AB} + Z_A Z_B}$$

The probabilities q_A and q_B are additional corrective factors that reflect the preferential choice of the subsequences given they will interact.

$$q_A = \frac{p_B p^I_{AB}}{\sum_X p_X p^I_{AX}} \qquad q_B = \frac{p_A p^I_{AB}}{\sum_Y p_Y p^I_{YB}}$$

where X and Y are subsequences in RNA l and RNA $l+1$ respectively.

With these newly defined weights, we obtain similar results for Section 5.1 and the exact same results for Section 5.2.

6 Conclusion

While RNA-RNA interaction algorithms exist, they are not suitable for predicting RNA structures in which more than two RNA molecules interact. For instance, the interaction pattern may not be known, in contrast to the case of two RNAs where one must interact with the other. Moreover, even with some existing knowledge on the pattern of interaction, treating the RNAs pairwise may not lead to the best global structure. In this work, we formulate multiple RNA interaction as an optimization problem, prove it is NP-complete, and provide approximation and heuristic algorithms. We explore three scenarios: 1) fishing for pairs: given a pool of RNAs, we identify the pairs that are known to interact; 2) structure prediction: we predict a correct complex of two snRNAs (modified human U2 and U6) and two structurally autonomous parts of an intron, a total of four RNAs; and 3) structural separation: we successfully divide the RNAs into independent groups of multiple interacting RNAs.

References

1. Pervouchine, D.D.: Iris: Intermolecular RNA interaction search. In: 15th International Conference on Genome Informatics (2004)
2. Alkan, C., Karakoc, E., Nadeau, J.H., Sahinalp, S.C., Zhang, K.: RNA-RNA interaction prediction and antisense RNA target search. Journal of Computational Biology 13(2) (2006)
3. Mneimneh, S.: On the approximation of optimal structures for RNA-RNA interaction. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009)
4. Meyer, I.M.: Predicting novel RNA-RNA interactions. Current Opinions in Structural Biology 18 (2008)
5. Kolb, F.A., Malmgren, C., Westhof, E., Ehresmann, C., Ehresmann, B., Wagner, E.G.H., Romby, P.: An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. RNA Society (2000)

6. Argaman, L., Altuvia, S.: fhla repression by oxys: Kissing complex formation at two sites results in a stable antisense-target RNA complex. Journal of Molecular Biology 300 (2000)
7. Muckstein, U., Tafer, H., Hackermuller, J., Bernhart, S.H., Stadler, P.F., Hofacker, I.L.: Thermodynamics of RNA-RNA binding. Journal of Bioinformatics (2006)
8. Chitsaz, H., Backofen, R., Sahinalp, S.C.: biRNA: Fast RNA-RNA binding sites prediction. In: Salzberg, S.L., Warnow, T. (eds.) WABI 2009. LNCS, vol. 5724, pp. 25–36. Springer, Heidelberg (2009)
9. Chitsaz, H., Salari, R., Sahinalp, S.C., Backofen, R.: A partition function algorithm for interacting nucleic acid strands. Journal of Bioinformatics (2009)
10. Salari, R., Backofen, R., Sahinalp, S.C.: Fast prediction of RNA-RNA interaction. Algorithms for Molecular Biology 5(5) (2010)
11. Huang, F.W.D., Qin, J., Reidys, C.M., Stadler, P.F.: Partition function and base pairing probabilities for RNA-RNA interaction prediction. Journal of Bioinformatics 25(20) (2009)
12. Li, A.X., Marz, M., Qin, J., Reidys, C.M.: RNA-RNA interaction prediction based on multiple sequence alignments. Journal of Bioinformatics (2010)
13. Sun, J.S., Manley, J.L.: A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. Genes and Development 9 (1995)
14. Zhao, C., Bachu, R., Popovic, M., Devany, M., Brenowitz, M., Schlatterer, J.C., Greenbaum, N.L.: Conformational heterogeneity of the protein-free human spliceosomal U2-U6 snRNA complex. RNA 19, 561–573 (2013), doi:10.1261/rna.038265.113; These two autors contributed equally to the manuscript
15. Cormen, T., Leiserson, C.E., Rivest, R.L., Stein, C.: Approximation Algorithms in Introduction to Algorithms. MIT Press (2010)
16. Mneimneh, S., Ahmed, S.A., Greenbaum, N.L.: Multiple RNA interaction: Formulations, approximations, and heuristics. In: Fourth International Conference on Bioinformatics Models, Methods, and Algorithms (2013)