Gibbs/MCMC Sampling for Multiple RNA Interaction with Sub-optimal Solutions

Saad Mneimneh(✉) and Syed Ali Ahmed

Hunter College and The Graduate Center City University of New York, New York, NY, USA
saad@hunter.cuny.edu, sahmed3@gradcenter.cuny.edu

Abstract. The interaction of two RNA molecules involves a complex interplay between folding and binding that warranted the development of RNA-RNA interaction algorithms. However, these algorithms do not handle more than two RNAs. We note our recent successful formulation for the multiple (more than two) RNA interaction problem based on a combinatorial optimization called *Pegs and Rubber Bands*. Even then, however, the optimal solution obtained does not necessarily correspond to the actual biological structure. Moreover, a structure produced by interacting RNAs may not be unique to start with. Multiple solutions (thus sub-optimal ones) are needed. Here, a sampling approach that extends our previous formulation for multiple RNA interaction is developed. By clustering the sampled solutions, we are able to reveal representatives that correspond to realistic structures. Specifically, our results on the U2-U6 complex and its introns in the spliceosome of yeast, and the CopA-CopT complex in E. Coli are consistent with published biological structures.

Keywords: Multiple RNA interaction · RNA structure · Gibbs sampling · Metropolis-Hastings algorithm · Clustering

1 Introduction

The role of interaction between two or more RNA molecules has been increasingly recognized in regulatory mechanisms, including gene expression, methylation, and splicing. Pairwise interaction has been noted for regulating gene expression, e.g. when one RNA binds to the ribosome binding site of another mRNA, thus blocking its translation to protein [18]. Typical scenarios of multiple RNA interaction involve the interaction of multiple small nucleolar RNAs (snoRNAs) with ribosomal RNAs (rRNAs) in guiding the methylation of the rRNAs [24], and multiple small nuclear RNAs (snRNA) with mRNAs in the splicing of introns [34].

The prediction of structures resulting from pairwise interactions is now somewhat understood, due to successful efforts in generalizing the *partition function* of a single RNA to the case of two.

S. Mneimneh—Supported by a Research Starter Award in Informatics from the PhRMA Foundation. Partially supported by CoSSMO CUNY.

S.A. Ahmed—Supported by a PSC CUNY Award 68671-00 46.

M. Botón-Fernández et al. (Eds.): AlCoB 2016, LNBI 9702, pp. 78–90, 2016.
DOI: 10.1007/978-3-319-38827-4_7

Algorithms for pairwise interaction of RNAs can be found in [3,7,8,15,19,24,25,29,31,32]. However, when carried over to multiple RNAs (more than two), generalizing the partition function further does not necessarily lead to efficient algorithms for computing it. Consequently, structure prediction in the context of multiple RNAs was almost non-existent; with just a few attempts that lack the ability to produce realistic structures. The de facto approach for multiple RNAs has been to account for their interaction by concatenating the RNAs into a single long RNA, which is then folded in order to predict the structure [4,10]. On the one hand, this presents a challenge to existing folding algorithms, which are far less reliable when the RNA is too long. On the other hand, most folding algorithms prevent the formation of pseudoknots due to their increased computational complexity. While pseudoknots are rare in folded structures, they translate into kissing loops when spanning multiple RNAs, which are quite frequent in interacting RNA structures. There are a few attempts for introducing kissing loops into the concatenation model, e.g. [6], but advances in pairwise interaction algorithms based on the generalized partition function suggest that the latter are more adequate, so they remain the state-of-the-art for two RNAs.

Therefore, a promising approach is to adapt existing pairwise interaction algorithms to the case of multiple RNAs. This generally leads to a computational hurdle: when RNAs are treated pairwise, an immediate consequence is the *greedy* nature of the algorithm. The best interacting pair of RNAs will dominate the solution, as in [35,36]. Since the pair of RNAs is required to fully interact, this will "lock" the interaction pattern of the whole ensemble into a sub-optimal state; thus preventing the correct structure from presenting itself as a solution.

We have recently proposed in a series of works [1,2,26,28] a mathematical formulation based on combinatorial optimization that overcomes the issues outlined above. The model handles multiple RNAs without having to generalize the partition function beyond pairs. The resulting algorithms are not based on the concatenation paradigm, so they allow the formation of kissing loops, as well as other structures. And while they are still primarily based on an adaptation of pairwise interaction, they avoid the "locking" problem mentioned earlier.

Even then, obtaining one (optimal) solution for a multiple RNA interaction problem is not completely satisfactory. Many biological factors are hard to account for computationally. In addition, correct biological structures are often not unique. Therefore, some realistic solutions are ought to be sub-optimal, which is what we address here.

2 Preliminaries

2.1 The Model: Pegs and Rubber Bands

We advocate a combinatorial optimization problem called Pegs and Rubber Bands as a framework for multiple RNA interaction. The link between the two will be made shortly following a formal description of Pegs and Rubber Bands.

Consider m levels numbered 1 to m with n_l pegs in level l numbered 1 to n_l. There is an infinite supply of rubber bands, and a rubber band can be placed

around pegs in consecutive levels. For instance, we may choose to place a rubber band around pegs $[i_1, i_2]$ (i.e., the set of pegs from i_1 to i_2, where $i_1 \leq i_2$), in level l, and pegs $[j_1, j_2]$ in level $l+1$. In this case, the rubber band defines a window with a given weight $w(l, i_2, j_2, u, v)$, where $u = i_2 - i_1 + 1$ and $v = j_2 - j_1 + 1$ represent the lengths of the intervals covered by the window in levels l and $l+1$, respectively (as in Fig. 1). For convenience, we will use $w(l, i, j, u, v)$ interchangeably to denote both the window and its weight, depending on context. As such, each window $w(l, i, j, u, v)$ defines two intervals, $[i - u + 1, i]$ in level l and $[j - v + 1, j]$ in level $l + 1$. Two windows overlap if any of their intervals overlap on the same level. In addition, $w(l, i, j, u, v)$ and $w(l, i', j', u', v')$ overlap if $\text{sgn}(i - i') \neq \text{sgn}(j - j')$ (their rubber bands cross).

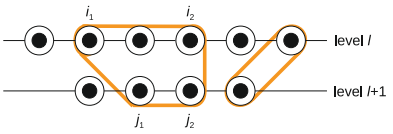

Fig. 1. A rubber band around pegs defines a window. The lengths $u = i_2 - i_1 + 1$ and $v = j_2 - j_1 + 1$ of the corresponding intervals may be different.

The Pegs and Rubber Bands problem is to maximize the total weight by placing rubber bands around pegs in such a way that none of their corresponding windows overlap.

To make the connection with multiple RNA interactions: RNA sequences become the levels, the ordered pegs in each level represent RNA bases $\{A, G, C, U\}$ in the order of occurrence in their sequence, a window $w(l, i, j, u, v)$ is an interaction between bases $[i - u + 1, i]$ in RNA l and bases $[j - v + 1, j]$ in RNA $l + 1$, and the weight $w(l, i, j, u, v)$ is chosen based on the energy of that interaction. The energies are obtained using a generalized partition function for pairwise interaction, and account for both intra- and inter- molecular energies. The no overlap condition reflects a typical nature of RNA interactions, and the maximization nature of the problem corresponds to energy minimization.

2.2 An Approximation Algorithm

A polynomial time approximation scheme (PTAS) for Pegs and Rubber Bands based on dynamic programming was described in [2,28], where $n = \max_l n_l$.

Theorem 1. *Polynomial Time Approximation Scheme (PTAS) Pegs and Rubber Bands is NP-hard; however, for every $\epsilon > 0$, it admits a polynomial time algorithm that runs in $O(\lceil \frac{1}{\epsilon} \rceil mn^{\lceil \frac{1}{\epsilon} \rceil})$ time and achieves a total weight within a $(1 - \epsilon)$-factor of optimal.*

The mapping of RNAs to levels can be obtained as in [2,28]. Figure 2 shows an example of a structure predicted using the Pegs and Rubber Bands formulation as reported in [2,28], where windows are replaced by bonds between their

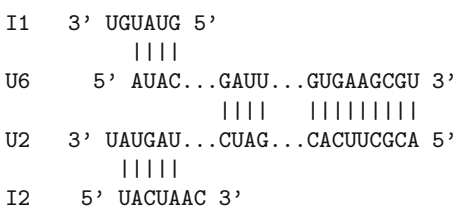

Fig. 2. Multiple RNA interaction within the eukaryotic spliceosome, a large ribonucleoprotein assembly responsible for the excision of intervening sequences in precursor messenger (pre-mRNA) molecules. Showing is the spliceosomal U2-U6 small nuclear (snRNA) and introns I1 and I2. The resulting structure is consistent with biological experiments [34,38].

corresponding intervals. The formulation avoids the "locking" problem, since treating the RNAs pairwise would have favored the full binding of U2-U6 to include their left extremities in Fig. 2, leaving I1 and I2 detached.

3 Realistic Biological Factors and Sub-optimal Solutions

Most algorithms for RNA-RNA interaction compute a partition function for the two RNAs based on loop energies in ways inspired by the basic algorithm of McCaskill for a single RNA [21]. Thus, when it comes to multiple RNA interaction, the maximization of weight in the Pegs and Rubber Bands problem is somewhat equivalent to minimization of energy.

We have successfully used weights obtained from the tool RNAup [29] as follows: $w(l,i,j,u,v) \propto \log P_l(free[i-u+1,i]) + \log P_{l+1}(free[j-v+1,j]) + \log Z_l^I(i-u+1,i,j-v+1,j)$ where $P_l(free[i,j])$ is the probability that subsequence $[i,j]$ is free (does not fold) in RNA l, and $Z_l^I(i_1,i_2,j_1,j_2)$ is the generalized 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 the RNAs subsequences). Therefore, the method may be categorized as an MFE-like approach (Minimum Free Energy). It is clear that such an approach does not capture "everything".

Many biological factors affect the observed structure of interacting RNA molecules. For instance, reversible kissing loops (where some hydrogen bonds of the interaction between hairpins unwind) [17] are generally not captured by MFE since a kissing loop is energetically more favorable than a partial one. We observe such artifacts within the pairwise interaction of CopA-CopT in E. Coli, as shown in Fig. 3.

Another example is the U2-U6 snRNA complex. There seems to be a lack of consensus whether the U2-U6 snRNA complex forms a 4-way or a 3-way junction (most likely both structures co-exist [5,30,33,38]). Figure 4 shows the two possibilities. It has been conjectured in [5] that co-axial stacking is essential for the stabilization of helix I in U2-U6 and, therefore, inhibition of the co-axial stacking, possibly by protein binding, may activate the second conformation (with helices Ia and Ib).

```
(a)
CopA 5' CGGUUUAAGUGGG...UUUCGUACUCGCCAAAGUUGAAGA...UUUUGCUU 3'
        ||||||||||||||   |||||||||||||||||||||||||   ||||||||
CopT 3' GCCAAAUUCACCC...AAAGCAUGAGCGGUUUCAACUUCU...AAAACGAA 5'


(b)
CopA 5' CGGUUUAAGUGGG...UUUCGUACUCGCCAAAGUUGAAGA...UUUUGCUU 3'
        ||||||||||||||     |||||||||      ||||||
CopT 3' GCCAAAUUCACCC...AAAGCAUGAGCGGUUUCAACUUCU...AAAACGAA 5'
```

Fig. 3. The pairwise interaction of CopA-CopT: (a) computational prediction with artifact interactions due to the maximization nature of the problem, and (b) the actual biologically known interaction [18], where the last window is dropped and the middle window is split (reversible kissing loop).

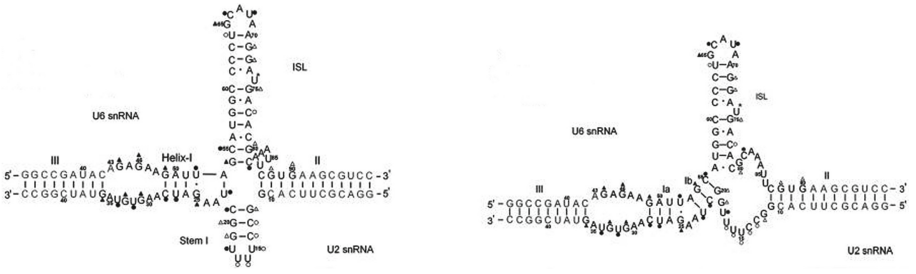

Fig. 4. U2-U6 snRNA complex in humans obtained by Greenbaum's lab [38]. The 4-way junction appears on the left hand side with Helix I, and the 3-way junction appears on the right hand side with Helices Ia and Ib.

Therefore, correct biological structures are not always "optimal" (from the computational perspective), and often are not unique. Sub-optimal solutions are needed to cover the biological ground. To that end, we consider in this paper two main modifications to our original approach based on Pegs and Rubber Bands:

- Sampling is used to produce multiple (sub-optimal) solutions instead of a single solution (this is described in Sect. 4).
- Windows are considered to be either *single* or *dependent*. Single windows contribute a weight equal to a sum of three terms as described above (our original formulation). Recall that each window $w(l, i, j, u, v)$ defines two intervals, $[i - u + 1, i]$ in level l and $[j - v + 1, j]$ in level $l + 1$. If a solution contains two windows that define intervals $[a, b]$ and $[c, d]$ in level l with $b < c$ and no other intervals in between, then we may consider them dependent in level l (windows can be dependent in one or two levels) and thus replace $\log P_l(free[a, b]) + \log P_l(free[c, d])$ (each of these two terms is coming from the *single* contribution of each window) with $\log P_l(free[a, d])$, if the latter is larger than the former sum. This allows window splits such as the one shown

in Fig. 3(b) to be not so detrimental to the total weight of the solution. Given a solution, its total weight is then obtained by the optimal determination of single and dependent windows in each level to maximize that weight (this is achieved by a dynamic programming algorithm for each level). We denote this modified weight of a solution S by $w(S)$.

4 A Sampling Approach

Sampling is more efficient than exhaustive enumeration of solutions within a certain threshold of optimal, especially that many of these solutions will be similar. Furthermore, sampling has been successfully used in the context of a single RNA; for instance, in [9,23,37] to mention a few examples. For the multiple RNA interaction, we propose below an approach based on Gibbs sampling and the Metropolis-Hastings algorithm.

4.1 The Gibbs Sampler

The described model for multiple RNA interaction, viewed as Pegs and Rubber Bands with m levels, lends itself quite naturally to Gibbs sampling [13,20]. As a random variable, let S_l be a set of non-overlapping windows of the form $w(l, i, j, u, v)$, so S_l represents a valid interaction pattern between RNA l and RNA $l+1$. A Gibbs sampler works by sampling each random variable individually in order, conditioned on the current values of the other variables. In other words, we work with $P(S_l|S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_{m-1})$. Therefore, if we start with $S_1^0 = \ldots = S_{m-1}^0 = \emptyset$, we sample S_1^1 using $P(S_1|S_2^0, \ldots, S_{m-1}^0)$, then S_2^1 using $P(S_2|S_1^1, S_3^0, \ldots S_{m-1}^0)$, then S_3^1 using $P(S_3|S_1^1, S_2^1, S_4^0, \ldots, S_{m-1}^0)$, and so on until we sample S_{m-1}^1 using $P(S_{m-1}|S_1^1, \ldots, S_{m-2}^1)$. We call $(S_1^1, \ldots, S_{m-1}^1)$ our first sample, and we repeat to obtain $(S_1^t, \ldots, S_{m-1}^t)$ for every t. Under typical conditions of ergodicity [11], the Gibbs guarantee is that $(S_1^t, \ldots, S_{m-1}^t)$ for large t is a sample from $P(S_1, \ldots, S_{m-1})$, which is not necessarily a known distribution, in contrast to $P(S_l|S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_{m-1})$ which is reasonably accessible.

This is interesting because, conditioned on $S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_{m-1}$, the permissible windows of the form $w(l, i, j, u, v)$ are exactly those which do not overlap with windows in S_{l-1} and S_{l+1}. As such, we assume that:

$$P(S_l|S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_{m-1}) = P(S_l|S_{l-1}, S_{l+1})$$

$$P(S_l|S_{l-1}, S_{l+1}) \propto \begin{cases} 0 & S_l \text{ contains a window that overlaps in } S_{l-1} \text{ or } S_{l+1} \\ e^{w(S_l)} & \text{otherwise} \end{cases}$$

The exponential term is similar in spirit to the standard Boltzman distribution used for RNAs, knowing that $w(S_l)$ represents the negative of the energy.

If $P(S_l|S_{l-1}, S_{l+1})$ is easy to sample from, then the Gibbs sampler works nicely given a fixed mapping of RNAs to levels 1 to m. We describe in the next section how to sample from $P(S_l|S_{l-1}, S_{l+1})$.

4.2 Gibbs Sampling with Metropolis-Hastings

The Metropolis-Hastings algorithm for sampling (also known as the Markov Chain Monte Carlo method) was described in [14,22], and since then has been utilized extensively in the literature. To sample from $P(S_l|S_{l-1}, S_{l+1})$, we first drop all the windows of the form $w(l, i, j, u, v)$ that overlap in S_{l-1} or S_{l+1}. We only work with the remaining windows of the form $w(l, i, j, u, v)$. We then construct a random sequence $S_l^0, S_l^1, \ldots$, where S_l^t is a set of non-overlapping windows of the form $w(l, i, j, u, v)$. This can be done with a Metropolis-Hastings strategy: Given S_l^t, we randomly generate S_l^{t+1} with some proposal probability $Q(S_l^{t+1}|S_l^t)$, and either accept S_l^{t+1} with probability

$$\min\left\{1, \frac{Q(S_l^t|S_l^{t+1})}{Q(S_l^{t+1}|S_l^t)} \times \frac{e^{w(S_l^{t+1})}}{e^{w(S_l^t)}}\right\}$$

or reject it and let $S_l^{t+1} = S_l^t$.

It is well known and easy to show that such a strategy results in a Markov chain which converges to the desired probability distribution if the proposal chain $Q(S_l^{t+1}|S_l^t)$ satisfies $Q(S_l^{t+1} = y|S_l^t = x) > 0 \Leftrightarrow Q(S_l^{t+1} = x|S_l^t = y) > 0$; this also makes it irreducible [12].

For practical purposes, we limit S_l^t to contain only windows $w(l, i, j, u, v)$ where $u = v$. We also do not allow two adjacent windows $w(l, i, j, u, v)$ and $w(l, i - u, j - v, u', v')$ to co-exists (since together they represent one bigger window). With that in mind, a simple strategy is to make $Q(S_l^{t+1}|S_l^t)$ **uniform** among all the neighbors of S_l^t (including S_l^t itself), where a neighbor other than S_l^t can be obtained by one of the following three operations:

- a window $w(l, i, j, u, v) \in S_l^t$ is removed from S_l^t
- a window $w(l, i, j, u, v) \notin S_l^t$ that does not overlap in S_l^t is added to S_l^t
- a window $w(l, i, j, u, v) \in S_l^t$ is replaced by a window $w(l, i', j', u', v') \notin S_l^t$ that only overlaps with $w(l, i, j, u, v)$ in S_l^t

Therefore, for every S_l^{t+1} that is a neighbor of S_l^t, $Q(S_l^{t+1}|S_l^t)$ is the inverse of the number of neighbors of S_l^t. This proposal probability defines an irreducible Markov chain since every pair of solutions can be reached from one another through a sequence of neighbors.

4.3 A Notion of Distance for Sub-optimal Solutions

Many of the sampled sub-optimal solutions will be similar. To quantify this similarity/dissimilarity, we need to describe a distance function. To motivate our approach, we first define the notion of a *terminal* window: Given a solution S, the terminal window $w(l, i, j, u, v) \in S$ is the window with the largest l such that no windows appear on its right in levels $l - 1$, l, and $l + 1$:

- no window $w(l - 1, i', j', u', v') \in S$ has $j' > i$
- no window $w(l, i', j', u', v') \in S$ has $i' > i$
- no window $w(l + 1, i', j', u', v') \in S$ has $i' > j$

By recursively eliminating the terminal window from a solution, we obtain a total order on the windows of that solution.

Our approach builds on the idea that if two solutions are similar, we expect them to have a similar set of windows; furthermore, these windows should exhibit the same order. In more detail, given a solution S, define $|S|$ as the number of windows in S, and let $w(l_1, i_1, j_1, u_1, v_1), \ldots, w(l_{|S|}, i_{|S|}, j_{|S|}, u_{|S|}, v_{|S|})$ be the $|S|$ windows in the order defined by terminal windows. Each of these windows, say $w(l, i, j, u, v)$, defines the two intervals, $[i - u + 1, i]$ in level l and $[j - v + 1, j]$ in level $l + 1$. Define the set of interaction intervals

$$I(S) = (I_1, \ldots, I_{2|S|}) = ([i_1 - u_1 + 1, i_1], [j_1 - v_1 + 1, j_1], \ldots$$

$$\ldots, [i_{|S|} - u_{|S|} + 1, i_{|S|}], [j_{|S|} - v_{|S|} + 1, j_{|S|}])$$

as an ordered sequence of $2|S|$ intervals, and $L(S) = (l_1, \ldots, l_{|S|})$ as an ordered sequence of $|S|$ levels, where l_i is the level defining the i^{th} window. Therefore, $L(S)$ means that we have the following set of pairwise interactions (not necessarily unique in terms of RNAs): RNA l_1 with RNA $l_1 + 1$, RNA l_2 with RNA $l_2 + 1$, ..., RNA $l_{|S|}$ with RNA $l_{|S|} + 1$. Two solutions that do not agree on this set, or do not define overlapping interaction intervals, are considered completely dissimilar; otherwise, their distance is given by the amount of overlap in their interaction intervals (as in the Jaccard metric [16]), hence the following definition of distance:

Given two solutions S_1 with $I(S_1) = (I_1, I_2, \ldots)$ and S_2 with $I(S_2) = (T_1, T_2, \ldots)$, the distance between S_1 and S_2 is

$$d(S_1, S_2) = \begin{cases} 1 - \frac{\sum_i |I_i \cap T_i|}{\sum_i |I_i \cup T_i|} & L(S_1) = L(S_2) \text{ and } I_i \cap T_i \neq \emptyset \text{ for all } i \\ 1 & \text{otherwise} \end{cases}$$

where $\cap$ and $\cup$ represent the standard intersection and union operations on sets respectively, and intervals are treated as sets of integers. This distance is modified from our previous metric in [26,27], and is not a metric; however, it works well with the clustering algorithm described below.

4.4 Clustering the Samples

The sampled sub-optimal solutions are generally more than what we need. In addition, as mentioned above, many of them will be similar. Therefore, we use clustering to reduce their number. To cluster the samples, we first remove duplicates, so we only work with unique samples. We then drop all solutions with a weight below 1/3 of the best. Finally, we sort the solutions to make the output of the clustering deterministic. We adopt hierarchical agglomerative clustering with complete linkage, and we obtain the clusters by "cutting" the tree where distance between clusters is 1. Given the clusters, the optimal solution in each cluster acts as a "representative" of the cluster. The representatives should reveal some of the structures that are observed in biological experiments [1,26,27].

5 Experimental Results

We perform 50 iterations of the Metropolis-Hastings algorithm **without** rejection. This allows us to start at some random solution. We then allow 50 iterations (with rejection) for the "burn-in" time of the Metropolis-Hastings algorithm. Finally, we generate 50 samples in 50 iterations and select one uniformly at random. We generate 1000 solutions (Gibbs samples) by repeating this procedure, as described in Sect. 4.1.

After clustering, we sort the representatives of the clusters by decreasing weight. We consider the first k representatives, for a given k. To assess our approach, we repeat the experiment 100 times. Given a set of candidate structures in mind; for instance, Fig. 5 shows four candidates for the yeast spliceosome, we then count how many times (in the 100 runs) each candidate is found among the first k representatives, as a percentage hit. We also compute the "rank" of each candidate, which is the first time[1] that candidate is seen as representative, averaged over the 100 experiments.

5.1 Experiment 1: Structural Variation

The U2-U6 complex in the spliceosome of yeast has been reported to have two distinct experimental structures, e.g. [33]. In one conformation, U2 and U6 interact to form a helix known as helix Ia. In another conformation, the interaction reveals a structure containing an additional helix, known as helix Ib. Section 3 describes possible underlying mechanisms that are responsible for this conformational switch. We consider the set of four candidates in Fig. 5. The results are summarized in Table 1.

Table 1. Results for the yeast spliceosome. Each entry lists the percentage hit followed by the average rank.

k	1		2		3		4		5		6		7		8		9		10	
Helices Ia+Ib	100	1	100	1	100	1	100	1	100	1	100	1	100	1	100	1	100	1	100	1
Helices Ia+Ib, I1 detached	0	–	100	2	100	2	100	2	100	2	100	2	100	2	100	2	100	2	100	2
Helix Ia	0	–	0	–	0	–	40	4	85	4.5	100	4.8	100	4.8	100	4.8	100	4.8	100	4.8
Helix Ia, I1 detached	0	–	0	–	0	–	0	–	40	5	85	5.5	100	5.8	100	5.8	100	5.8	100	5.8

5.2 Experiment 2: Artifact Interactions

Due to the optimization nature of the problem, it is sometimes easy to pick up interactions that are not biologically real. This is because dropping these interactions from the solution would make it sub-optimal (even when preferred biologically, as described in Sect. 3). The last interaction window of CopA-CopT

[1] We use "first time" because many solutions can represent the same candidate; for instance, a window can split in different ways, but we still refer to it as a window split.

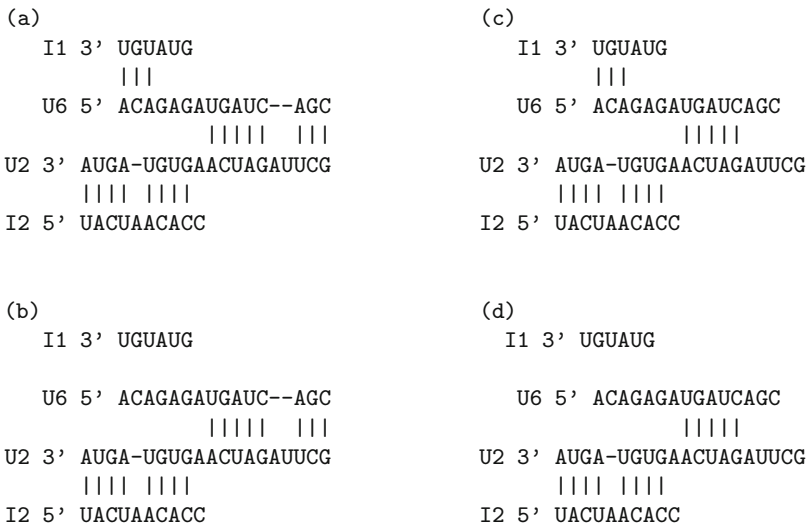

Fig. 5. The yeast spliceosome with 4 RNAs (I1 and I2 are functionally independent stretches of the same much longer messenger RNA). (a) Helix Ia and helix Ib with both introns attached. (b) Helix Ia and helix Ib with I1 detached. (c) Helix Ia with both introns attached. (d) Helix Ia with I1 detached.

in Fig. 3 is an example of such an artifact. We consider six candidate solutions based on presence/absence of windows and window splits, as described in Table 2. For each of the three interaction windows in Fig. 3, we consider whether the window is present, dropped, or split. Typically, we detect a window split when the two portions happen to be treated as *dependent* in some level l (see Sect. 3). Therefore, to correctly capture reversible kissing loops, undesired splits can be ignored if the corresponding window does not represent a kissing loop. Given the RNA structures of CopA and CopT, only the middle window is a kissing loop.

Table 2. Results for CopA-CopT. For each of the three interaction windows in Fig. 3, we consider whether the window is present, dropped, or split. Each entry lists the percentage hit followed by the average rank.

k	1		2		3		4		5		6		7		8		9		10	
First, middle, last	89.6	1	93.8	1	97.9	1.1	100	1.2	100	1.2	100	1.2	100	1.2	100	1.2	100	1.2	100	1.2
First, middle split, last	4.2	1	52.1	1.9	77.1	2.3	93.8	2.6	97.9	2.7	100	2.8	100	2.8	100	2.8	100	2.8	100	2.8
First, middle, last dropped	4.2	1	10.4	1.6	16.7	2.1	20.8	2.5	27.1	3.1	29.2	3.3	31.2	3.5	31.2	3.5	35.4	4.2	35.4	4.2
First, middle split, last dropped	0	–	2.1	2	2.1	2	4.2	3	12.5	4.3	18.8	4.9	25	5.4	29.2	5.8	37.5	6.5	41.7	6.8
First split, middle, last	2.1	1	8.3	1.8	20.8	2.5	27.1	2.8	43.8	3.7	54.2	4.1	70.8	4.8	79.2	5.1	83.3	5.3	83.3	5.3
First split, middle, last dropped	0	–	0	–	0	–	0	–	2.1	5	4.2	5.5	6.2	6	6.2	6	10.4	7.2	10.4	7.2

6 Conclusion

In RNA interaction, the optimal structure may not be the real structure, and the real structure may not be unique. In this work, we build on our previous approach for multiple RNA interaction using the Pegs and Rubber Bands formulation to generate multiple sub-optimal solutions. This is developed using Gibbs sampling and the Metropolis-Hastings algorithm.

Our sampling approach successfully computes sub-optimal solutions for the multiple RNA interaction problem that are truthful representations of the actual biological structures. For instance, it can provide several candidate structures when they exist, e.g. the U2-U6 complex and its introns in the spliceosome of yeast, and find structures that agree with the literature, but are not necessarily optimal in the computational sense, e.g. CopA-CopT in E. Coli.

References

1. Ahmed, S.A., Mneimneh, S.: Multiple RNA interaction with sub-optimal solutions. In: Basu, M., Pan, Y., Wang, J. (eds.) ISBRA 2014. LNCS, vol. 8492, pp. 149–162. Springer, Heidelberg (2014)
2. Ahmed, S.A., Mneimneh, S., Greenbaum, N.L.: A combinatorial approach for multiple RNA interaction: formulations, approximations, and heuristics. In: Du, D.-Z., Zhang, G. (eds.) COCOON 2013. LNCS, vol. 7936, pp. 421–433. Springer, Heidelberg (2013)
3. Alkan, C., Karakoc, E., Nadeau, J.H., Sahinalp, S.C., Zhang, K.: RNA-RNA interaction prediction and antisense RNA target search. J. Comput. Biol. **13**(2), 267–282 (2006)
4. Andronescu, M., Zhang, Z.C., Condon, A.: Secondary structure prediction of interacting RNA molecules. J. Mol. Biol. **345**(5), 987–1001 (2005)
5. Cao, S., Chen, S.J.: Free energy landscapes of RNA/RNA complexes: with applications to snRNA complexes in spliceosomes. J. Mol. Biol. **357**(1), 292–312 (2006)
6. Chen, H.L., Condon, A., Jabbari, H.: An $o(n^5)$ algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. J. Comput. Biol. **16**(6), 803–815 (2009)
7. 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)
8. Chitsaz, H., Salari, R., Sahinalp, S.C., Backofen, R.: A partition function algorithm for interacting nucleic acid strands. Bioinformatics **25**(12), i365–i373 (2009)
9. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31**(24), 7280–7301 (2003)
10. Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E., Pierce, N.A.: Thermodynamic analysis of interacting nucleic acid strands. SIAM Rev. **49**(1), 65–88 (2007)
11. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998). Chap. 11
12. Gallager, R.G.: Discrete Stochastic Processes, vol. 321. Springer Science & Business Media, Newyork (2012). Chap. 4

13. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI **6**(6), 721–741 (1984)
14. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
15. Huang, F.W., Qin, J., Reidys, C.M., Stadler, P.F.: Partition function and base pairing probabilities for RNA-RNA interaction prediction. Bioinformatics **25**(20), 2646–2654 (2009)
16. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr. Corbaz (1901)
17. Kolb, F.A., Engdahl, H.M., Slagter-Jäger, J.G., Ehresmann, B., Ehresmann, C., Westhof, E., Wagner, E.G.H., Romby, P.: Progression of a loop-loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. EMBO J. **19**(21), 5905–5915 (2000)
18. Kolb, F.A., Malmgren, C., Westhof, E., Ehresmann, C., Ehresmann, B., Wagner, E., 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 **6**(3), 311–324 (2000)
19. Li, A.X., Marz, M., Qin, J., Reidys, C.M.: RNA-RNA interaction prediction based on multiple sequence alignments. Bioinformatics **27**(4), 456–463 (2011)
20. Liu, J.S.: The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89**(427), 958–966 (1994)
21. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**(6–7), 1105–1119 (1990)
22. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
23. Metzler, D., Nebel, M.E.: Predicting RNA secondary structures with pseudoknots by MCMC sampling. J. Math. Biol. **56**(1–2), 161–181 (2008)
24. Meyer, I.M.: Predicting novel RNA-RNA interactions. Curr. Opin. Struct. Biol. **18**(3), 387–393 (2008)
25. Mneimneh, S.: On the approximation of optimal structures for RNA-RNA interaction. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **6**(4), 682–688 (2009)
26. Mneimneh, S., Ahmed, S.A.: Multiple RNA interaction: beyond two. To appear in IEEE Trans. Nanobiosci. (2015)
27. Mneimneh, S., Ahmed, S.A.: A sampling approach for multiple RNA interaction: finding sub-optimal solutions fast. In: BIOINFORMATICS 2015 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, Rome, Italy, 21–23 February 2015
28. Mneimneh, S., Ahmed, S.A., Greenbaum, N.L.: Multiple RNA interaction - formulations, approximations, and heuristics. In: BIOINFORMATICS 2013 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, Barcelona, Spain, 11–14 February 2013, pp. 242–249 (2013)
29. Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S.H., Stadler, P.F., Hofacker, I.L.: Thermodynamics of RNA-RNA binding. Bioinformatics **22**(10), 1177–1182 (2006)
30. Newby, M.I., Greenbaum, N.L.: A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. RNA **7**(6), 833–845 (2001)
31. Pervouchine, D.D.: IRIS: intermolecular RNA interaction search. Genome Inform. Ser. **15**(2), 92 (2004)

32. Salari, R., Backofen, R., Sahinalp, S.C.: Fast prediction of RNA-RNA interaction. Algorithms Mol. Biol. **5**(5) (2010)
33. Sashital, D.G., Cornilescu, G., Butcher, S.E.: U2–U6 RNA folding reveals a group II intron-like domain and a four-helix junction. Nat. Struct. Mol. Biol. **11**(12), 1237–1242 (2004)
34. Sun, J.S., Manley, J.L.: A novel U2–U6 snRNA structure is necessary for mammalian mRNA splicing. Genes Dev. **9**(7), 843–854 (1995)
35. Tong, W., Goebel, R., Liu, T., Lin, G.: Approximation algorithms for the maximum multiple RNA interaction problem. In: Widmayer, P., Xu, Y., Zhu, B. (eds.) COCOA 2013. LNCS, vol. 8287, pp. 49–59. Springer, Heidelberg (2013)
36. Tong, W., Goebel, R., Liu, T., Lin, G.: Approximating the maximum multiple RNA interaction problem. Theoret. Comput. Sci. **556**, 63–70 (2014)
37. Wei, D., Alpert, L.V., Lawrence, C.E.: Rnag: a new gibbs sampler for predicting RNA secondary structure for unaligned sequences. Bioinformatics **27**(18), 2486–2493 (2011)
38. Zhao, C., Bachu, R., Popović, 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**(4), 561–573 (2013)