A Sampling Approach for Multiple RNA Interaction
Finding Sub-optimal Solutions Fast

Saad Mneimneh[1,*] and Syed Ali Ahmed[2,†]

[1] *Department of Computer Science, Hunter College of the City University of New York, New York, U.S.A.*

[2] *Department of Computer Science, The Graduate Center of the City University of New York, New York, U.S.A.*

Keywords: Multiple RNA Interaction, RNA Structure, Gibbs Sampling, Metropolis-Hastings Algorithm, Clustering.

Abstract: The interaction of two RNA molecules involves a complex interplay between folding and binding that warranted recent developments in RNA-RNA interaction algorithms. These algorithms cannot be used to predict interaction structures when the number of RNAs is more than two. Our recent formulation of the multiple RNA interaction problem is based on a combinatorial optimization called *Pegs and Rubber Bands*, and has been successful in predicting structures that involve more than two RNAs. 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. We extend our previous approach to generate multiple sub-optimal solutions that was based on exhaustive enumeration. Here, a sampling approach for multiple RNA interaction is developed. Since not too many samples are needed to reveal solutions that are sufficiently different, sampling provides a much faster alternative. By clustering the sampled solutions, we are able to obtain representatives that correspond to the biologically observed structures. Specifically, our results for 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.

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. 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 (Meyer, 2008), and multiple small nuclear RNAs (snRNA) with mRNAs in the splicing of introns (Sun and Manley, 1995).

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

$$Z = \sum_S e^{-\beta E(S)}$$

where S is a structure, $E(S)$ is its free energy, and β is typically $1/RT$, where R is the Boltzman constant and T is temperature. The probability of a structure S is then given by:

$$P(S) = \frac{e^{-\beta E(S)}}{Z}$$

and the probability of an event (e.g. the event $free[i,j]$ that bases $i, i+1, \ldots, j$ do not bind) is

$$\sum_{S \text{ has this event}} P(S)$$

Algorithms for pairwise interaction of RNAs can be found in (Pervouchine, 2004; Alkan et al., 2006; Mneimneh, 2009; Meyer, 2008; Mückstein et al., 2006; Chitsaz et al., 2009a; Salari et al., 2010; Chitsaz et al., 2009b; Huang et al., 2009; Li et al., 2011) (the last six of these deal with some form of a generalized partition function for the two RNAs as a whole). 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

[*] Partially supported by CoSSMO at CUNY.

[†] Supported by a PSC CUNY Award 68671-00 46.

Mneimneh, S. and Ahmed, S.
A Sampling Approach for Multiple RNA Interaction - Finding Sub-optimal Solutions Fast.
DOI: 10.5220/0005707900750084
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 3: BIOINFORMATICS*, pages 75-84
ISBN: 978-989-758-170-0

in the context of multiple RNAs was almost nonexistent; 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 (Andronescu et al., 2005), (Dirks et al., 2007). 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. (Chen et al., 2009), 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 (Tong et al., 2013; Tong et al., 2014). which are a spin off of our work (Ahmed et al., 2013a). 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 been recently proposing in a series of works (Mneimneh et al., 2013; Ahmed et al., 2013b; Ahmed and Mneimneh, 2014; Mneimneh and Ahmed, 2015) 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, biological structures are often not unique. Therefore, some correct 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 that we call 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 Figure 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).

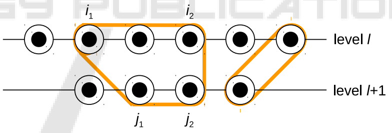

Figure 1: A rubber band around pegs defines a window. The lengths $u = i_2 - i_1 + 1$ and $v = j_2 - j_1 + 1$ of the corresponding intervals may be different.

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. We have successfully used weights obtained from the tool RNAup

(Mückstein et al., 2006) as follows:

$$
\begin{aligned}
w(l,i,j,u,v) \propto \log P_l(free[i-u+1,i]) \\
+\log P_{l+1}(free[j-v+1,j]) \\
+\log Z_l^I(i-u+1,i,j-v+1,j)
\end{aligned}
$$

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 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).

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 (Mneimneh et al., 2013; Ahmed et al., 2013b), where $n = \max_l n_l$.

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

Viewing the interaction of m RNAs as Pegs and Rubber Bands with m levels dictates that the first RNA interacts with the second RNA, and the second with the third, and so on. This not only imposes a specific order on the interaction, but it also restricts each RNA to interact with at most two others. Therefore, this rather arbitrary choice for the order can be eliminated: We first identify each RNA as being *even* (sense) or *odd* (antisense). Given m RNAs and a permutation (order) on the set $\{1,\ldots,m\}$, we map the RNAs onto the levels as follows: Starting with the first RNA, and moving in order, we place RNAs on the first level as long as they have the same parity. We then move to the next level, and perform this process for the remaining set. This is repeated until all RNAs have been placed. 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 a rubber band to span multiple RNAs on the same level. Given a solution, random perturbations of the permutation are then used to find better solutions. A heuristic algorithm is shown in Figure 2.

Figure 3 shows an example of a structure predicted using the Pegs and Rubber Bands formulation and the above algorithm as reported in (Mneimneh et al., 2013; Ahmed et al., 2013b), where windows are replaced by bonds between their corresponding intervals.

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

Figure 2: A heuristic algorithm to find one solution for multiple RNA interaction using the PTAS algorithm.

```
I1   3′ UGUAUG 5′
          ||||
U6     5′ AUAC...GAUU...GUGAAGCGU 3′
                 ||||   |||||||||
U2   3′ UAUGAU...CUAG...CACUUCGCA 5′
         |||||
I2    5′ UACUAAC 3′
```

Figure 3: 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 (Sun and Manley, 1995; Zhao et al., 2013).

The algorithm avoids the "locking" problem, since treating the RNAs pairwise would have favored the full binding of U2-U6 to include their left extremities in Figure 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 (McCaskill, 1990). Thus, when it comes to multiple RNA interaction, our maximization of weight in the Pegs and Rubber Bands problem is somewhat equivalent to minimization of energy. Recall,

$$w(l,i,j,u,v) \propto -E_1 - E_2 - E_3$$

where $E_1 = -(1/\beta)\log P_l(free[i-u+1,i])$ and $E_2 = -(1/\beta)\log P_{l+1}(free[j-v+1,j])$ are the free energies associated with exposing the binding sites in both RNAs respectively, and $E_3 = -(1/\beta)\log Z_l^I(i-u+1,i,j-v+1,j)$ is the free energy associated with their hybridization (interaction). Therefore, our 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) (Kolb et al., 2000a) 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 Figure 4.

```
(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′
```

Figure 4: 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 (Kolb et al., 2000b).

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 (Newby and Greenbaum, 2001; Zhao et al., 2013; Cao and Chen, 2006; Sashital et al., 2004)). Figure 5 shows the two possibilities. It has been conjectured in (Cao and Chen, 2006) 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).

Therefore, correct biological structures are not always "optimal" (from the computational perspective), and often are not unique. While the algorithm of Figure 2 will produce a solution within $(1-\varepsilon)$ factor of optimal (and hopefully the optimal), multiple sub-optimal solutions are needed to cover the biological ground. To substantiate this claim, a previous work in (Ahmed and Mneimneh, 2014; Mneimneh and Ahmed, 2015) has successfully identified biologically meaningful structures by first exhaustively enumerating all solutions that fall within a threshold of the best, and then clustering them to identify the crucial differences. When each cluster is represented by its optimal solution, the few best representatives turned out to be biologically relevant candidates. In particular, for the yeast U2-U6 snRNA complex (with introns) reported in (Newby and Greenbaum, 2001), the two configurations involving either helix Ia or both helices Ia and Ib have been successfully retrieved (the sequences of U2 and U6 have been truncated up to helix Ib, as done in (Newby and Greenbaum, 2001)). Similarly, multiple solutions move CopA-CopT closer toward the actual solution we know. The results are shown in Figures 6 and 7, respectively.

4 A SAMPLING APPROACH

The exhaustive enumeration for sub-optimal solutions based on setting a minimum threshold, as described in the previous section, suffers from a major drawback: many sub-optimal solutions are similar and, therefore, to obtain sufficiently different solutions, a large number of sub-optimal solutions must be generated. This will not only generate hundreds of thousands of solutions, but will also slow down the clustering algorithm. To alleviate this problem, we consider a sampling approach. In fact, sampling has been successfully used in the context of a single RNA; for instance, in (Ding and Lawrence, 2003), (Metzler and Nebel, 2008), and (Wei et al., 2011) 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

Our model for multiple RNA interaction, viewed as Pegs and Rubber Bands with m levels, lends itself quite naturally to Gibbs sampling (Geman and Geman, 1984; Liu, 1994). 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 $P(S_l|S_1,\ldots,S_{l-1},S_{l+1},\ldots,S_m)$. Therefore, if we start with $S_1^0 = \ldots = S_m^0 = \emptyset$, we sample S_1^1 using $P(S_1|S_2^0,\ldots,S_m^0)$, then S_2^1 using $P(S_2|S_1^1,S_3^0,\ldots S_m^0)$, then S_3^1 using $P(S_3|S_1^1,S_2^1,S_4^0,\ldots,S_m^0)$, and so on until we sample S_m^1 using $P(S_m|S_1^1,\ldots,S_{m-1}^1)$. We call $(S_1^1,\ldots,S_m^1)$ our first sample, and we repeat to obtain $(S_1^t,\ldots,S_m^t)$ for every t. Under typical conditions of

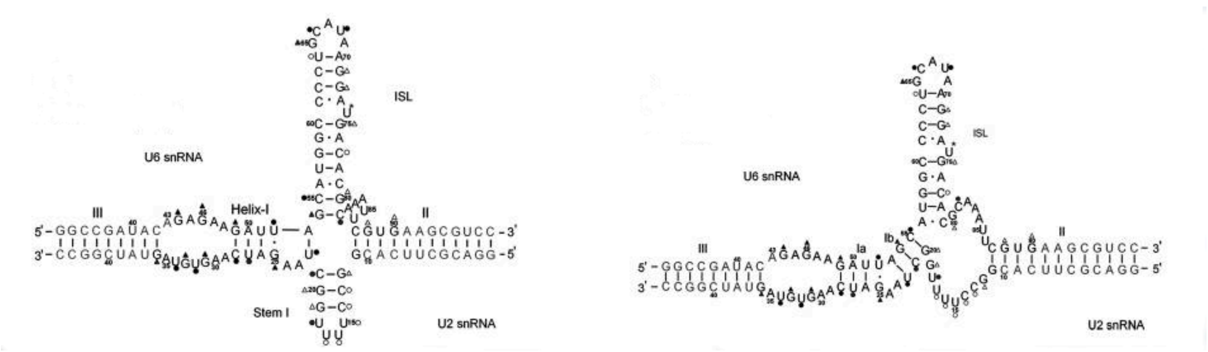

Figure 5: U2-U6 snRNA complex in humans obtained by Greenbaum's lab (Zhao et al., 2013). The 4-way junction appears on the left hand side with Helix I, and the 3-way junction appears on the right hand side with Helices Ia and Ib.

ergodicity (Durbin et al., 1998), the Gibbs guarantee is that $(S_1^t, \ldots, S_m^t)$ for large t is a sample from $P(S_1, \ldots, S_m)$, which is not necessarily a known distribution, in contrast to $P(S_l|S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_m)$ which is reasonably accessible.

This is interesting because, conditioned on $S_1, \ldots, S_{l-1}, S_{l+1}, \ldots, S_m$, 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 (recall that we use $w(l,i,j,u,v)$ to denote both a window and its weight, depending on context):

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

$$P(S_l|S_{l-1}, S_{l+1}) \propto \begin{cases} 0 & S_l \text{ contains a window that overlaps in } S_{l-1} \text{ or } S_{l+1} \\ \exp\left[\sum_{w(l,i,j,u,v)\in S_l} w(l,i,j,u,v)\right] & \text{otherwise} \end{cases}$$

The exponential term is similar in spirit to the standard Boltzman distribution used for RNAs, knowing that $w(l,i,j,u,v)$ 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. Such mapping may be obtained from the heuristic algorithm of Figure 2 (we typically use 1/2, 1/3 or 1/4 for ε). 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 (Metropolis et al., 1953) and (Hastings, 1970), 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

$$\min\left\{1, \frac{Q(S_l^t|S_l^{t+1})}{Q(S_l^{t+1}|S_l^t)} \times \frac{\exp[\sum_{w(l,i,j,u,v)\in S_l^{t+1}} w(l,i,j,u,v)]}{\exp[\sum_{w(l,i,j,u,v)\in S_l^t} w(l,i,j,u,v)]}\right\}$$

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 (Gallager, 2012).

For practical purposes, we limit S_l^t to contain only windows $w(l,i,j,u,v)$ where $u = v$ and $w(l,i,j,u,v) > 0$. 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, where a neighbor 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

```
(a) candidate 1

   I1 3′ UGUAUG
         |||
   U2 5′ ACAGAGAUGAUC--AGC
                |||||  |||
U6 3′ AUGA-UGUGAACUAGAUUCG
      |||| ||||
I2 5′ UACUAACACC
```

```
(b) candidate 2

   I1 3′ UGUAUG

   U2 5′ ACAGAGAUGAUC--AGC
                |||||  |||
U6 3′ AUGA-UGUGAACUAGAUUCG
      |||| ||||
I2 5′ UACUAACACC
```

```
(c) candidate 3

   I1 3′ UGUAUG
         |||
   U2 5′ ACAGAGAUGAUCAGC
                |||||
U6 3′ AUGA-UGUGAACUAGAUUCG
      |||| ||||
I2 5′ UACUAACACC
```

```
(d) candidate (4)

   I1 3′ UGUAUG
         |||
   U2 5′ ACAGAGAUGAUC--AGC
                |||||  |||
 U6 3′ AUGAUGUGAACUAGAUUCG
           ||||
I2 5′ UACUAACACC
```

```
(e) candidate 5

  I1 3′ UGUAUG

   U2 5′ ACAGAGAUGAUCAGC
                |||||
U6 3′ AUGA-UGUGAACUAGAUUCG
      |||| ||||
I2 5′ UACUAACACC
```

```
(f) candidate 6

  I1 3′ UGUAUG

  U2 5′ ACAGAGA-UGAUC--AGC
        ||      |||||  |||
 U6 3′ AUGAUGUGAACUAGAUUCG
           ||||
I2 5′ UACUAACACC
```

Figure 6: 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 and helix Ib with I2 partially detached. (e) Helix Ia with I1 detached. (f) Helix Ia and helix Ib with I1 detached and I2 partially detached, moving towards detaching both introns, as would happen when U2 and U6 are bound optimally in a full pairwise interaction.

can be reached from one another through a sequence of neighbors.

4.3 A Distance Metric for Sub-optimal Solutions

Many of the sampled sub-optimal solutions will be similar. To quantify this similarity/dissimilarity, we need a notion of a distance. We adopt an approach inspired by the Jaccard metric (Jaccard, 1901).

To motivate this 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 $i' > j$

```
(a) candidate 1

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

```
(b) candidate 2

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

Figure 7: CopA-CopT in E. Coli. (a) The best solution found. (b) A solution closer to the one observed in biological experiments in which the third interaction window in non-existent.

By recursively eliminating the terminal window from a solution, we obtain a total order on the windows of that solution. 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, are considered completely dissimilar; otherwise, their distance is given by the amount of overlap in their interaction intervals (as in the Jaccard metric), 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

$$d(S_1,S_2) = \begin{cases} 1 & L(S_1) \neq L(S_2) \\ 1 - \frac{\sum_i |I_i \cap T_i|}{\sum_i |I_i \cup T_i|} & L(S_1) = L(S_2) \end{cases}$$

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 a metric in $[0,1]$ (Mneimneh and Ahmed, 2015).

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 adopt hierarchical agglomerative clustering with single linkage and the silhouette index (Rousseeuw, 1987) to determine the optimal number of clusters. Given a solution S, let $c(S)$ be its cluster. Let $b_j(S)$ be the average distance from S to all solutions in cluster j, and let $b(S) = \min_{j \neq c(S)} b_j(S)$. We assume that the number of clusters is at least 2, so $b(S)$ is defined. Let $a(S)$ be the average distance from S to all other solutions in $c(S)$. If S is a singleton in its cluster, we make $a(S) = b(S)$. The silhouette of a solution S is given by

$$\frac{b(S) - a(S)}{\max[a(S), b(S)]}$$

and is always in the interval $[-1, 1]$. A silhouette close to 1 means that solution S is well situated in its cluster since $a(S) \ll b(S)$. The silhouette of a cluster is the average silhouette of all the solutions in the cluster. The silhouette index is the average of all the cluster silhouettes. We seek the number of clusters that maximizes this index. The beauty of this index follows from that it is always bounded, works for arbitrary notions of distance (dissimilarity), and does not require the use of a cluster centroid, which is typically not trivial to find for non-Euclidean distances.

Given a number of 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 (Ahmed and Mneimneh, 2014; Mneimneh and Ahmed, 2015).

5 EXPERIMENTAL RESULTS

We allow 100 iterations for the "burn-in" time of the Metropolis-Hastings algorithm to obtain the first sample, and 50 iterations between consecutive samples thereafter. We note that we can generate a 1000 solutions (Gibbs samples) in just a few seconds, which is several orders of magnitude improvement over the previous exhaustive approach in (Ahmed and Mneimneh, 2014; Mneimneh and Ahmed, 2015).

After clustering, we sort the representatives of the clusters by decreasing weight. Then to assess our approach, for each experiment we have a number k of candidate structures in mind; for instance, Figure 6 shows six candidates ($k = 6$), and Figure 7 shows two candidates ($k = 2$). Given that set of candidates, we only consider the first $\leq k$ representatives, but we still record the total number of clusters obtained. For each candidate S, and using the distance metric described earlier, we find the representative R that is closest to S. If $S = R$, that's a direct hit (a distance of zero). Otherwise, if among the k candidates, the closest to R is S itself, we declare this as a "close" hit with the given distance. In this case, and after converting the windows of S and R to bonds, we also compute their $F1$-score (Powers, 2011) given by:

$$F1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision is defined as the number of bonds in R that are also in S divided by the number of bonds in R, and recall is defined as the number of bonds in R that are also in S divided by the number of bonds in S. If there are no direct or close hits for candidate S, we declare a miss.

We repeat the experiment 100 times and compute the average number of unique samples and the average number of clusters obtained, and for each of the k candidates, the percentage of direct hits, the average distance and average $F1$-score of close hits, and the percentage of misses.

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. (Sashital et al., 2004). 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 six candidates in Figure 6. Tables 1 and 2 summarize the results of this experiment using 1000 samples for $k = 4$ and $k = 6$ respectively, supporting the fact that the two conformations show up. With 1000 samples, the first k candidates always show up with a 100% hit.

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 less optimal (even when preferred biologically, as described in Section 3). The third interaction window of CopA-CopT in Figure 7 is an example of such an artifact.

Table 1: Yeast spliceosome. 100 runs, 100 samples in each run, avg. number of unique samples: 70.82, avg. number of clusters: 11.07, $k = 6$.

candidate	%hit	avg. distance	avg. $F1$-score	%miss
1	98	0.1	0.947	0
2	90	0.118	0.938	1
3	69	0.106	0.944	3
4	68	0.106	0.937	13
5	52	0.123	0.934	13
6	44	0.079	0.959	47

Table 2: Yeast spliceosome. 100 runs, 100 samples in each run, avg. number of unique samples: 70.82, avg. number of clusters: 11.07, $k = 4$.

candidate	%hit	avg. distance	avg. $F1$-score	%miss
1	98	0.1	0.947	0
2	90	0.118	0.938	1
3	69	0.118	0.937	17
4	68	0.067	0.966	30
5	30	0.121	0.936	67
6	7	-	-	93

For the two candidates of Figure 7 ($k = 2$), Tables 3 and 4 summarize the results of this experiment using 100 and 1000 samples respectively, showing that we succeed in dropping the undesired window.

Table 3: CopA-CopT. 100 runs, 100 samples in each run, avg. number of unique samples: 72.26, avg. number of clusters: 2.55, $k = 2$.

candidate	%hit	avg. distance	avg. $F1$-score	%miss
1	39	0.082	0.980	0
2	2	0.094	0.950	60

Table 4: CopA-CopT. 100 runs, 1000 samples in each run, avg. number of unique samples: 505.1, avg. number of clusters: 2.37, $k = 2$.

candidate	%hit	avg. distance	avg. $F1$-score	%miss
1	99	0.022	0.989	0
2	30	0.052	0.973	3

6 CONCLUSION

In multiple RNA interaction, the best structure may not be the real structure, and the real structure may not be unique. In this work, we build on a previous approach (exhaustive enumeration) to generate multiple sub-optimal solutions using the Pegs and Rubber Bands formulation. Here, an approach using Gibbs sampling and the Metropolis-Hastings algorithm is developed, and provides a much faster alternative to exhaustive enumeration.

This new 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

Ahmed, S. A. and Mneimneh, S. (2014). Multiple rna interaction with sub-optimal solutions. In *Bioinformatics Research and Applications*, pages 149–162. Springer.

Ahmed, S. A., Mneimneh, S., and Greenbaum, N. L. (2013a). A combinatorial approach for multiple rna interaction: Formulations, approximations, and heuristics. In *Computing and Combinatorics*, pages 421–433. Springer.

Ahmed, S. A., Mneimneh, S., and Greenbaum, N. L. (2013b). A combinatorial approach for multiple rna interaction: Formulations, approximations, and heuristics. In *Computing and Combinatorics*, pages 421–433. Springer Berlin Heidelberg.

Alkan, C., Karakoc, E., Nadeau, J. H., Sahinalp, S. C., and Zhang, K. (2006). Rna-rna interaction prediction and antisense rna target search. *Journal of Computational Biology*, 13(2):267–282.

Andronescu, M., Zhang, Z. C., and Condon, A. (2005). Secondary structure prediction of interacting rna molecules. *Journal of molecular biology*, 345(5):987–1001.

Cao, S. and Chen, S.-J. (2006). Free energy landscapes of rna/rna complexes: with applications to snrna complexes in spliceosomes. *Journal of molecular biology*, 357(1):292–312.

Chen, H.-L., Condon, A., and Jabbari, H. (2009). An $o(n^5)$ algorithm for mfe prediction of kissing hairpins and 4-chains in nucleic acids. *Journal of Computational Biology*, 16(6):803–815.

Chitsaz, H., Backofen, R., and Sahinalp, S. C. (2009a). birna: Fast rna-rna binding sites prediction. In *Algorithms in Bioinformatics*, pages 25–36. Springer.

Chitsaz, H., Salari, R., Sahinalp, S. C., and Backofen, R. (2009b). A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373.

Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for rna secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301.

Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and Pierce, N. A. (2007). Thermodynamic analysis of interacting nucleic acid strands. *SIAM review*, 49(1):65–88.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids, Chapter 11*. Cambridge university press.

Gallager, R. G. (2012). *Discrete stochastic processes, Chapter 4*, volume 321. Springer Science & Business Media.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Huang, F. W., Qin, J., Reidys, C. M., and Stadler, P. F. (2009). Partition function and base pairing probabilities for rna–rna interaction prediction. *Bioinformatics*, 25(20):2646–2654.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Kolb, F. A., Engdahl, H. M., Slagter-Jäger, J. G., Ehresmann, B., Ehresmann, C., Westhof, E., Wagner, E. G. H., and Romby, P. (2000a). Progression of a loop–loop complex to a four-way junction is crucial for the activity of a regulatory antisense rna. *The EMBO journal*, 19(21):5905–5915.

Kolb, F. A., Malmgren, C., Westhof, E., Ehresmann, C., Ehresmann, B., Wagner, E., and Romby, P. (2000b). 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.

Li, A. X., Marz, M., Qin, J., and Reidys, C. M. (2011). Rna–rna interaction prediction based on multiple sequence alignments. *Bioinformatics*, 27(4):456–463.

Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metzler, D. and Nebel, M. E. (2008). Predicting rna secondary structures with pseudoknots by mcmc sampling. *Journal of mathematical biology*, 56(1-2):161–181.

Meyer, I. M. (2008). Predicting novel rna–rna interactions. *Current opinion in structural biology*, 18(3):387–393.

Mneimneh, S. (2009). On the approximation of optimal structures for rna-rna interaction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):682–688.

Mneimneh, S. and Ahmed, S. A. (2015). Multiple rna interaction: Beyond two. *To appear in IEEE Transactions on NanoBioscience*.

Mneimneh, S., Ahmed, S. A., and Greenbaum, N. L. (2013). 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.*, pages 242–249.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. (2006). Thermodynamics of rna–rna binding. *Bioinformatics*, 22(10):1177–1182.

Newby, M. I. and Greenbaum, N. L. (2001). A conserved pseudouridine modification in eukaryotic u2 snrna induces a change in branch-site architecture. *RNA*, 7(06):833–845.

Pervouchine, D. D. (2004). Iris: intermolecular rna interaction search. *Genome Informatics Series*, 15(2):92.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Salari, R., Backofen, R., and Sahinalp, S. C. (2010). Fast prediction of rna-rna interaction. *Algorithms for molecular Biology*, 5(5).

Sashital, D. G., Cornilescu, G., and Butcher, S. E. (2004). U2–u6 rna folding reveals a group ii intron-like domain and a four-helix junction. *Nature structural & molecular biology*, 11(12):1237–1242.

Sun, J.-S. and Manley, J. L. (1995). A novel u2-u6 snrna structure is necessary for mammalian mrna splicing. *Genes & Development*, 9(7):843–854.

Tong, W., Goebel, R., Liu, T., and Lin, G. (2013). Approximation algorithms for the maximum multiple rna interaction problem. In *Combinatorial Optimization and Applications*, pages 49–59. Springer.

Tong, W., Goebel, R., Liu, T., and Lin, G. (2014). Approximating the maximum multiple rna interaction problem. *Theoretical Computer Science*.

Wei, D., Alpert, L. V., and Lawrence, C. E. (2011). Rnag: a new gibbs sampler for predicting rna secondary structure for unaligned sequences. *Bioinformatics*, 27(18):2486–2493.

Zhao, C., Bachu, R., Popović, M., Devany, M., Brenowitz, M., Schlatterer, J. C., and Greenbaum, N. L. (2013). Conformational heterogeneity of the protein-free human spliceosomal u2-u6 snrna complex. *RNA*, 19(4):561–573.

APPENDIX

RNA Sequences

```
CopA-CopT in E. Coli.

CopA (even)
5′ CGGUUUAAGUGGGCCCCGGUAAUCUUUUCGUACUCGCCA
   AAGUUGAAGAAGAUUAUCGGGGUUUUUGCUU 3′

CopT (odd)
5′ AAGCAAAAACCCCGAUAAUCUUCUUCAACUUUGGCGAGU
   ACGAAAAGAUUACCGGGGCCCACUUAAACCG 3′

Human Spliceosome

I1 (odd)
5′ NNNNNNNNNNGUAUGUNNNNNNNNNN 3′

U6 (even)
5′ AUACAGAGAAGAUUAGCAUGGCCCCUGCGCAAGGAUGAC
   ACGCAAAUUCGUGAAGCGU 3′

U2 (odd)
5′ ACGCUUCACGGCCUUUUGGCUAAGAUCAAGUGUAGUAU 3′

I2 (even)
5′ NNNNNNNNNNUACUAACNNNNNNNNNN 3′

Yeast Spliceosome

I1 (odd)
5′ NNNNGUAUGUNNNN 3′

U6 (even)
5′ ACAGAGAUGAUCAGC 3′

U2 (odd)
5′ GCUUAGAUCAAGUGUAGUA 3′

I2 (even)
5′ NNNNUACUAACACCNNNN 3′
```