On the Approximation of Optimal Structures for RNA-RNA Interaction

Saad Mneimneh

Abstract—The interaction of two RNA molecules is a common mechanism for many biological processes. Small interfering RNAs represent a simple example of such an interaction. But other more elaborate instances of RNA-RNA interaction exist. Therefore, algorithms that predict the structure of the RNA complex thus formed are of great interest. Most of the proposed algorithms are based on dynamic programming. RNA-RNA interaction is generally NP-complete; therefore, these algorithms (and other polynomial time algorithms for that matter) are not expected to produce optimal structures. Our goal is to characterize this suboptimality. We demonstrate the existence of constant factor approximation algorithms that are based on dynamic programming. In particular, we describe 1/2 and 2/3 factor approximation algorithms. We define an *entangler* and prove that 2/3 is a theoretical upper bound on the approximation factor of algorithms that produce entangler-free solutions, e.g., the mentioned dynamic programming algorithms.

Index Terms—RNA-RNA interaction, approximation algorithms.

1 Introduction

The interaction of two RNA molecules involves an interplay between the folding of individual molecules on one hand, and the binding of the two molecules on the other hand. An example of such interaction is RNA interference (RNAi), where a small interfering RNA (known as siRNA) can be used to silence a given gene by targeting its messenger RNA: The siRNA binds to the (possibly folding) messenger RNA and triggers a cascade of events that would eventually destroy it (Post Transcriptional Gene Silencing by RNAi) [4]. While RNAi may be viewed as a special case of RNA-RNA interaction, since siRNAs are 19-21 nucleotides long and do not generally fold, other complex examples of RNA-RNA interaction exist where both RNAs fold (see Section 5).

The problem of individual RNA folding has been studied extensively in the literature, and many polynomial time algorithms for determining the optimal structure (e.g., with maximum number of bonds, or more generally with minimum energy) of a folded RNA molecule have been developed [9], [12], [13]. Only recently, however, there have been several concurrent yet independent efforts (including our own) to mathematically formulate RNA-RNA interaction and develop algorithms that predict the structure of the RNA complex thus formed, e.g., [10], [2], [11], and [1].

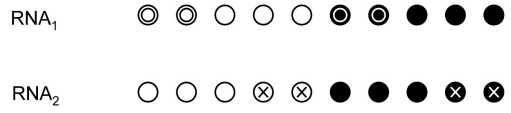
Most of the proposed algorithms (e.g., not [10]) are based on dynamic programming—apparently a "hard to avoid" influence from extensive RNA folding literature. Since mathematical formulations of RNA-RNA interaction generally give rise to NP-complete problems [1], these algorithms (and other polynomial time algorithms for that matter) are not expected to produce optimal structures. Our goal is to characterize this suboptimality.

We demonstrate the existence of constant factor approximation algorithms that are based on dynamic programming. In particular, we describe 1/2 and 2/3 factor approximation algorithms. We introduce the concept of an *entangler*: a special molecular substructure that may exist in the formed RNA complex. We argue that the mentioned dynamic programming algorithms do not produce entanglers, and prove that for some instances, any entangler-free solution is at best a 2/3 factor approximation. However, despite the theoretical suboptimality of these algorithms, they are able to predict some known RNA complexes. In particular, our algorithms predict to a great degree of satisfaction the fhlA-OxyS and the CopA-CopT complexes in the *Escherichia coli* bacteria.

Section 2 motivates the RNA-RNA interaction problem through a toy example. Section 3 gives a precise mathematical formulation of the problem to be solved. Section 4 describes approximation algorithms based on dynamic programming and derives the claimed 2/3 upper bound on the approximation factor. Section 5 provides experimental results. Finally, we conclude in Section 6.

2 A Toy Example

Consider the following two RNAs where nucleotides are represented by patterns. A solid pattern can bond to the same nonsolid pattern.



The two RNAs can independently fold into two optimal structures as shown on the next page. Each RNA can have

• *The author is with the Department of Computer Science, Hunter College, City University of New York, 695 Park Avenue, New York, NY 10065. E-mail: saad@hunter.cuny.edu.*

Manuscript received 25 July 2006; revised 12 June 2007; accepted 13 Sept. 2007; published online 12 Oct. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0148-0706.
Digital Object Identifier no. 10.1109/TCBB.2007.70258.

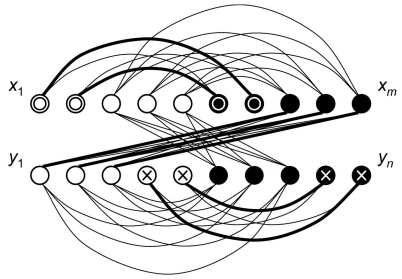
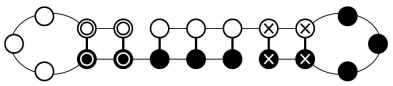


Fig. 1. RNA-RNA interaction graph (solution emphasized).

up to three bonds optimally. The total number of bonds in this case is six.

optimal folding of RNA_1 optimal folding of RNA_2

However, if the two RNAs can interact and form external bonds, then we may obtain a better structure (in terms of the number of bonds), yielding seven bonds in total, as shown below (another possibility is to bind the loops instead of the tails, i.e., kissing loops):



Therefore, to capture the simultaneous folding and binding of this interaction, one can form an RNA-RNA interaction graph where every edge represents a possible bond (either internal to the RNA itself, or external to the other RNA). The RNA-RNA interaction graph for the example above is illustrated in Fig. 1 (with edges of the above structure emphasized).

The problem then becomes to identify a maximum cardinality set (or, more generally, a set with maximum weight as described in Section 3) of nonintersecting edges[1] (edges sharing a node are also intersecting). This particular formulation of avoiding intersection is motivated by two facts:

- Pseudoknots are rare in RNA folded structures [9], [13], [7].
- RNA-RNA binding occurs between a sense ($5'$ to $3'$) molecule and an antisense ($3'$ to $5'$) molecule (so it is also not likely to have knotted interactions).

3 The RNA-RNAi Problem

We generalize the idea explored in Section 2 and include a weight for every possible bond. Therefore, the RNA-RNAi[2] problem is the following: Given an RNA-RNA interaction graph (V, E), where nodes of V are partitioned into two ordered sets $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, and every edge $e \in E$ has a weight $w(e) \in \mathbb{Q}$, find a set of **node disjoint** edges S that maximizes $\sum_{e \in S} w(e)$ such that (intersection is avoided):

- If $(x_i, x_j) \in S$ and $(x_k, x_l) \in S$, then NOT $i < k < j < l$.
- If $(y_i, y_j) \in S$ and $(y_k, y_l) \in S$, then NOT $i < k < j < l$.
- If $(x_i, y_j) \in S$ and $(x_k, y_l) \in S$, then NOT ($i < k$ and $j > l$).

The nucleotides of RNA_1 are represented by the ordered elements of X, and the nucleotides of RNA_2 are represented by the ordered elements of Y. Therefore, it is convenient to consider RNA_1 to be the string $x = x_1 \ldots x_m$, and RNA_2 to be the string $y = y_1 \ldots y_n$. We refer to the edges of the RNA-RNA interaction graph connecting both RNAs, i.e., of the form (x_i, y_j), as *binding edges*. We also refer to the edges of the RNA-RNA interaction graph internal to a given RNA, i.e., of the form (x_i, x_j) or (y_i, y_j), as *folding edges* for that RNA. A solution to RNA-RNAi (whether optimal or not) is said to have a weight equal to its achievable sum. As in the case of RNA folding problems, this weighted formulation provides a **basic** model for real RNA-RNA interaction problems where every bond contributes a specific energy. In an RNA world, weights are negative and the objective is to minimize the energy, i.e., the sum of weights, but this is an equivalent formulation.

A special case of RNA-RNAi is the uniformly weighted RNA-RNAi where the weights are all the same. For a uniformly weighted RNA-RNAi, maximizing $\sum_{e \in S} w(e)$ is equivalent to maximizing the cardinality of S (as described in the toy example of Section 2). Even the special case of a uniformly weighted RNA-RNAi (the decision version) is NP-complete [1].[3]

Theorem 1. *RNA-RNAi (the decision version) is NP-complete (even when uniformly weighted).*

4 Basic Approximation Algorithms

In this section, we provide some basic constant factor approximation algorithms for the RNA-RNAi problem. Recall from Section 3 the definitions of binding edges and folding edges. Also recall that RNA_1 can be represented as the string $x = x_1 \ldots x_m$, and RNA_2 can be represented as the string $y = y_1 \ldots y_n$.

4.1 A 1/2 Factor Approximation Algorithm

Consider the structures obtained from performing the following (using the given weight function):

- optimally solve RNA-RNAi while ignoring binding edges and
- optimally solve RNA-RNAi while ignoring folding edges for both RNAs.

The first step corresponds to optimally folding RNA_1 and RNA_2 independently. The second step corresponds to optimally aligning[4] RNA_1 and RNA_2.

1. Intersection is interpreted here given the particular realization of the graph, i.e., the graph cannot be redrawn to avoid intersection.

2. The terminology RNA-RNAi, which stands for RNA-RNA interaction, is not to be confused with RNA interference (RNAi).

3. The NP-completeness result was also established independently by the author in an unpublished manuscript (initially Southern Methodist University Technical Report 04-CSE-03) in September 2004. The manuscript is available from the author.

4. By alignment, we signify the binding resulting from aligning RNA_1 and the complement of RNA_2 with a zero scoring gap function. For instance, *cgga* and *gccu* align perfectly. This is equivalent to finding the largest weight common subsequence of RNA_1 (of string $x = x_1 \ldots x_m$) and the complement of RNA_2 (of string $y = y_1 \ldots y_n$).

Optimal folding and optimal alignment are both well-studied problems and can be solved in polynomial time. Optimally folding an RNA of length n takes $O(n^3)$ time and $O(n^2)$ space [9], and optimally aligning two RNAs of lengths m and n, respectively, takes $O(mn)$ time [8] and linear space [5].

The important observation is that one of the two obtained structures has a weight equal to at least 1/2 the weight of the optimal solution for the corresponding RNA-RNAi problem. Let w_1 and w_2 be the weights achieved by the two structures, respectively. Let OPT be the weight of the optimal solution S.

Lemma 1. $\max(w_1, w_2) \geq \frac{1}{2}OPT$.

Proof. Consider the optimal solution. Let A be the sum of weights of edges (bonds) formed in the folded part of RNA_1, i.e., edges in S of the form (x_i, x_j). Let B be the sum of weights of edges (bonds) formed by the alignment part of RNA_1 and RNA_2, i.e., edges in S of the form (x_i, y_j). Let C be the sum of weights of edges (bonds) formed in the folded part of RNA_2, i.e., edges in S of the form (y_i, y_j). Then, $OPT = A + B + C$ (the weight of the optimal solution). By the optimality of the two structures, $w_1 \geq A + C$ and $w_2 \geq B$. Therefore, $2\max(w_1, w_2) \geq w_1 + w_2 \geq A + C + B = OPT$. □

Obviously, in the independent folding of the RNAs, the remaining (nonfolded) nucleotides of the RNAs may be aligned. On the other hand, in the alignment, the remaining (nonbinding) nucleotides of the RNAs may be folded independently. These optional steps are shown between parentheses in the description of the algorithm below.[5]

Algorithm 1: 1/2 factor approximation.
1. optimally fold RNA_1 and RNA_2 independently (optional: optimally align their remainders)
2. optimally align RNA_1 and RNA_2 (optional: optimally fold their remainders independently)
3. choose the structure with the maximum weight.

4.2 A 2/3 Factor Approximation Algorithm

Consider the structures obtained from performing the following (using the given weight function):

- optimally solve RNA-RNAi while ignoring binding edges,
- optimally solve RNA-RNAi while ignoring the folding edges for RNA_2, and
- optimally solve RNA-RNAi while ignoring the folding edges for RNA_1.

5. While these steps make it possible for the algorithm to produce nonentangler-free solutions (see Section 4.3), they do not theoretically enhance the approximation factor; for example, consider the following uniformly weighted instance: $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3, y_4\}$, and $E = \{(x_1, x_3), (x_2, x_4), (x_1, y_2), (x_2, y_2), (x_3, y_3), (x_3, y_4), (y_1, y_3), (y_2, y_4)\}$. $S = \{(x_2, x_4), (x_1, y_2), (x_3, y_4), (y_1, y_3)\}$ is optimal, while $S_1 = \{(x_1, x_3), (y_2, y_4)\}$ and $S_2 = \{(x_2, y_2), (x_3, y_3)\}$ are locally optimal and cannot be extended.

As before, the first step corresponds to optimally folding RNA_1 and RNA_2 independently. The second step corresponds to optimally folding RNA_1 while interacting with the nonfolding RNA_2. Similarly, the third step corresponds to optimally folding RNA_2 while interacting with the nonfolding RNA_1.

One of the three obtained structures has a weight equal to at least 2/3 of the weight of the optimal solution for the corresponding RNA-RNAi problem. Let w_i, for $i = 1 \ldots 3$, be the weight achieved by the three structures, respectively. Let OPT be the weight of the optimal solution S.

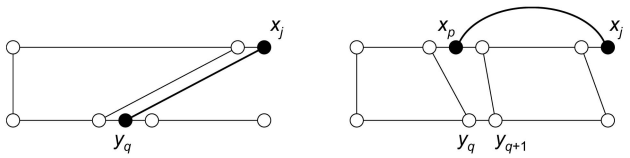
Lemma 2. $\max(w_1, w_2, w_3) \geq \frac{2}{3}OPT$.

Proof. Let A, B, and C be defined as in the proof of Lemma 1, where $OPT = A + B + C$. By the optimality of the three structures, $w_1 \geq A + C$, $w_2 \geq A + B$, and $w_3 \geq B + C$. Therefore, $3\max(w_1, w_2, w_3) \geq w_1 + w_2 + w_3 \geq A + C + A + B + B + C = 2(A + B + C) = 2OPT$. □

As stated in Section 4.1, optimal folding is a well-studied problem and can be solved in $O(n^3)$ time for an RNA of length n [9]. Therefore, the only concern now is to show that it is possible to optimally fold RNA_1 while interacting with a nonfolding RNA_2 in polynomial time. This can be done by a dynamic programming algorithm. Let the strings $x = x_1 \ldots x_m$ and $y = y_1 \ldots y_n$ denote the folding and nonfolding RNAs, respectively. Let $V(i, j, k, l)$ denote the weight that can be achieved in the optimal solution S for the substrings $x[i \ldots j]$ and $y[k \ldots l]$. Then, we have three possibilities for x_j: x_j does not bond, x_j bonds with some y_q (edge $(x_j, y_q) \in S$) and $k \leq q \leq l$, or x_j bonds with some x_p (edge $(x_p, x_j) \in S$) and $i \leq p < j$. Therefore, we have the following dynamic programming algorithm to compute $V(1, m, 1, n)$. The last two cases are also illustrated pictorially:

$$V(i,j,k,l) = \max \begin{cases} V(i, j-1, k, l) \\ V(i, j-1, k, q-1) + w(x_j, y_q) \\ V(i, p-1, k, q) + \\ V(p+1, j-1, q+1, l) + w(x_{p \neq j}, x_j) \end{cases}$$

over all choices of p and q, where $i \leq p \leq j$ and $k \leq q \leq l$ and w is the weight function.



If $k < l$, we set $V(i, j, k, l) = F_x(i, j)$, the weight of optimally folding $x[i \ldots j]$. The actual structure (i.e., S) can be obtained by standard dynamic programming bookkeeping/backtracking methods.

Noting that each case divides the problem into two independent subproblems, i.e., substrings (with one of them being possibly empty), where folding binds only the extremities, the formulation above can be simplified as follows:

Fig. 2. Entangler.

$$V(i,j,k,l) = \max\begin{cases} V(i+1,j-1,k,l) + w(x_{i\neq j}, x_j) \\ V(i,p,k,q) + V(p+1,j,q+1,l) \end{cases}$$

over all choices of p and q, where $i-1 \leq p \leq j$ and $k-1 \leq q \leq l$ and $(p \neq i-1 \vee q \neq k-1) \wedge (p \neq j \vee q \neq l)$, $V(i,j,k,k-1) = F_x(i,j), V(i,i-1,k,l) = 0$, and $V(i,i,k,k) = \max(0, w(x_i, y_k))$.

We have $O(m)$ values for p and $O(n)$ values for q, and hence, $V(i,j,k,l)$ requires $O(mn)$ time to compute. Since we have $O(m^2)$ substrings of x and $O(n^2)$ substrings of y, this algorithm runs in $O(m^3n^3)$ time and $O(m^2n^2)$ space.

In the independent folding of the RNAs, the remaining (nonfolded) nucleotides of the RNAs may be aligned. On the other hand, in a folding/alignment, the remaining (nonbinding) nucleotides of the nonfolding RNA may be folded. These optional steps are shown between parentheses in the description of the algorithm below.[6]

Algorithm 2: 2/3 factor approximation.

1. optimally fold RNA_1 and RNA_2 independently
   (optional: optimally align their remainders)
2. optimally fold RNA_1 while interacting with RNA_2 and ignore folding for RNA_2
   (optional: optimally fold the remainder of RNA_2)
3. optimally fold RNA_2 while interacting with RNA_1 and ignore folding for RNA_1
   (optional: optimally fold the remainder of RNA_1)
4. choose the structure with the maximum weight.

4.3 A Note on the Approximation Factor

In the previous sections, we relied on dynamic programming algorithms (through alignments and foldings) to obtain constant factor approximations for RNA-RNAi. Therefore, a legitimate question is whether better constant approximation factors can be obtained using such algorithms. To answer this question, we introduce the concept of an *entangler* (Fig. 2).

Definition 1 (entangler). *An entangler is a set of five nonintersecting edges that contains two folding edges (x_i, x_j) and (y_k, y_l), and three binding edges e_1, e_2, and e_3, such that:*

- $e_1 = (x_p, y_q) \Rightarrow p \in (i,j), q \notin (k,l)$,
- $e_2 = (x_p, y_q) \Rightarrow p \in (i,j), q \in (k,l)$, *and*
- $e_3 = (x_p, y_q) \Rightarrow p \notin (i,j), q \in (k,l)$,

where (i,j) denotes $\{i+1, \ldots, j-1\}$.

Most dynamic programming algorithms for RNA-RNAi (including the formulations in Section 4) recursively divide the problem into two independent subproblems (i.e. with disjoint substrings), by making a choice for at most one edge to be included in the solution. Such algorithms do not produce entanglers: There is no way to break an entangler into independent subproblems, even after making a choice for one edge (at least another edge must be excluded).[7] We claim that an entangler-free solution cannot achieve a constant approximation factor better than 2/3.

One can definitely design a dynamic programming algorithm that computes the optimal entangler-free solution. It will be similar to the dynamic programming formulation described in the previous section, but performing symmetrically on both RNAs and allowing both RNAs to fold (and interact):

$$V(i,j,k,l) = \max\begin{cases} V(i+1,j-1,k,l) + w(x_{i\neq j}, x_j) \\ V(i,p,k,q) + V(p+1,j,q+1,l) \\ V(i,j,k+1,l-1) + w(y_{k\neq l}, y_l) \end{cases}$$

over all choices of p and q, where $i-1 \leq p \leq j$ and $k-1 \leq q \leq l$ and $(p \neq i-1 \vee q \neq k-1) \wedge (p \neq j \vee q \neq l)$, $V(i,j,k,k-1) = F_x(i,j)$, $V(i,i-1,k,l) = F_y(k,l)$, and $V(i,i,k,k) = \max(0, w(x_i, y_k))$.

The above algorithm appears in [11] and [1]. Its running time and space requirements are still $O(m^3n^3)$ and $O(m^2n^2)$, respectively. It is easy to show that any entangler-free solution can be recursively broken into two independent subproblems as dictated by the above dynamic programming formulation, and hence, this algorithm computes the optimal entangler-free solution (we do not provide a formal argument because it is not needed for the upper bound result). Note that this algorithm is also a 2/3 factor approximation algorithm because all three solutions described at the beginning of Section 4.2 are entangler-free.

We now exhibit an instance of the RNA-RNAi problem where every entangler-free solution is asymptotically at most a 2/3 factor approximation. This proves the claim of this section.

Given an integer $r > 0$, the instance consists of 3^r nonintersecting binding edges partitioned into three groups (of 3^{r-1} edges each) by 2^{r-1} nonintersecting folding edges on each side, i.e., of the form (x_i, x_j) and (y_k, y_l), respectively. Then, each of the three groups is recursively partitioned in the same way. The partitioning stops when we obtain a single entangler, i.e., when $r = 1$. We assume all edges have the same weight (an instance of uniformly weighted RNA-RNAi). Fig. 3 illustrates the instance for $r = 3$.

It is easy to show that the number of folding edges on each side is given by the following expression:

$$\sum_{i=0}^{r-1} 3^{r-1-i} 2^i = 3^{r-1} \sum_{i=0}^{r-1} \left(\frac{2}{3}\right)^i = 3^{r-1} \frac{1-(2/3)^r}{1-2/3} = 3^r - 2^r.$$

6. While these steps make it possible for the algorithm to produce nonentangler-free solutions (see Section 4.3), they do not theoretically enhance the approximation factor; for example, consider the following uniformly weighted instance: $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3, y_4\}$, and $E = \{(x_2, x_3), (x_3, x_4), (x_1, y_3), (x_2, y_2), (x_3, y_1), (y_2, y_3), (y_3, y_4)\}$. $S = \{(x_3, x_4), (x_2, y_2), (y_3, y_4)\}$ is optimal, while $S_1 = \{(x_2, x_3), (y_2, y_3)\}$, $S_2 = \{(x_2, x_3), (x_1, y_3)\}$, and $S_3 = \{(x_3, y_1), (y_2, y_3)\}$ are locally optimal and cannot be extended.

7. We do not attempt to make this notion precise because we will prove a result in terms of entangler-free solutions rather than algorithms that produce them.

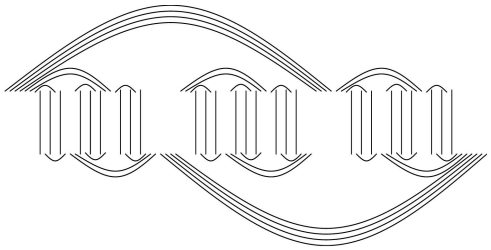


Fig. 3. An instance of recursive entanglers with $r = 3$.

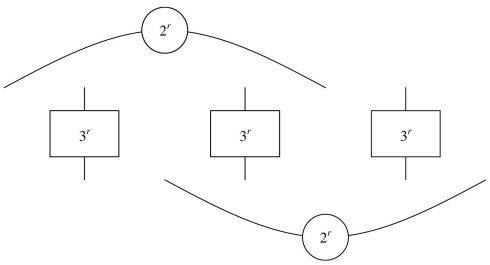
Therefore, the optimal solution contains $3^r + (3^r - 2^r) + (3^r - 2^r) = 3^{r+1} - 2^{r+1}$ edges (all of them). Obviously, there is an entangler-free solution with $3^r + (3^r - 2^r) = 2 \cdot 3^r - 2^r$ edges (it excludes the folding edges on one side). This is not the only possible entangler-free solution; however, we claim that any entangler-free solution must exclude at least $3^r - 2^r$ edges. Assuming this claim is true, the approximation factor of an entangler-free solution is at most

$$\frac{2 \cdot 3^r - 2^r}{3^{r+1} - 2^{r+1}} = \frac{2}{3} + \epsilon,$$

where $\lim_{r \to \infty} \epsilon = 0$.

Theorem 2. *An entangler-free solution for the (uniformly weighted) RNA-RNAi problem is asymptotically at most a 2/3 factor approximation.*[8]

Proof. For the instance corresponding to a given r, we prove that any entangler-free solution must exclude at least $3^r - 2^r$ edges, by induction on r. The base case is when $r = 1$, i.e., the instance is just an entangler. Therefore, at least $3^1 - 2^1 = 1$ edge must be excluded. Now assume the claim is true up to some value r. The instance corresponding to $r + 1$ can be viewed as follows:



The three rectangular sets represent the binding edges. The two circular sets represent the folding edges. Since the solution is entangler-free, at least one of these five sets must be excluded. If a circular set is excluded, the number of excluded edges is at least $3^r - 2^r$ for each of the three subproblems (inductive hypothesis) in addition to 2^r edges for a circular set. Hence, the number of excluded edges is at least $3(3^r - 2^r) + 2^r = 3^{r+1} - 2^{r+1}$. If a rectangular set is excluded, the number of excluded edges is at least $3^r - 2^r$ for two subproblems (inductive hypothesis) in addition to 3^r edges for one rectangular set. Hence, the number of excluded edges is at least $2(3^r - 2^r) + 3^r = 3^{r+1} - 2^{r+1}$. This proves the induction and, hence, the theorem. □

Note that this theorem is a tight characterization of entangler-free solutions because there is always an entangler-free solution for the RNA-RNAi problem that achieves a 2/3 factor approximation.

8. An entangler can be generalized to a set of $2k - 1$ nonintersecting edges that contains k binding edges interleaved by $k - 1$ folding edges ($\lceil \frac{k-1}{2} \rceil$ and $\lfloor \frac{k-1}{2} \rfloor$, respectively, on each side). One can then build an instance of k^r nonintersecting binding edges recursively partitioned into k groups by $k - 1$ sets of $(k-1)^{r-1}$ nonintersecting folding edges to obtain an asymptotic bound of $\frac{k-1}{k}$.

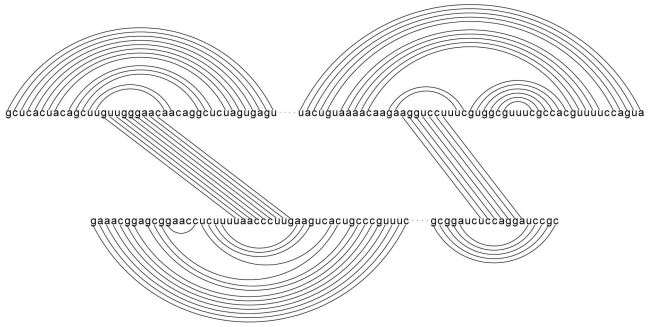


Fig. 4. fhlA-OxyS.

5 Experimental Results for RNA-RNAi

Although the presented algorithms achieve constant approximation factors, not every solution obtained by these algorithms is realistic. For instance, RNAs do not fold sharply and tend to fold locally. Moreover, two RNAs are likely to interact using complementary blocks of certain sizes. Therefore, we performed variants of the basic algorithms of Section 4 on two example RNA-RNAi problems; fhlA-OxyS interaction [3] and CopA-CopT interaction [6] in the *Escherichia coli* bacteria. As heuristics, we constrained the folding and alignment in the following ways: if x_i binds to x_j (edge $(x_i, x_j) \in S$), then $4 \leq |i - j| \leq 50$. Moreover, if x_p binds to y_q (edge $(x_p, y_q) \in S$), then $p \in [i, j]$ and $q \in [k, l]$ such that

- $j - i = l - k = B - 1$,
- x_{i+r} binds to y_{k+r} for all $r = 0 \ldots B - 1$, and
- x_{i-1} and x_{j+1} do not bind to y, and y_{k-1} and y_{l+1} do not bind to x.

Therefore, the algorithms are modified to compute block alignments. The details of the modified algorithms are not included (a variation on the first dynamic programming formulation of Section 4.2 to allow lower and upper bounds on B), but the modifications do not affect the theoretical complexity of the algorithms.

For weights, we used $w(g, u) = 1$, $w(a, u) = 2$, and $w(g, c) = 3$ (which are reasonably proportional to the energy values at 37 degrees [14]). We performed the algorithm of Section 4.1 on fhlA-OxyS with $7 \leq B < \infty$ as acceptable block sizes. We obtained the structure illustrated in Fig. 4 which is almost identical to the known structure of fhlA-OxyS [3] (small differences in folding around the first binding site). Stretches in the middle of the RNAs (9 nucleotides for fhlA, and 43 nucleotides for OxyS) were ignored for better prediction, because they were not reported to fold or bind [3]. Keeping those stretches maintains the same binding sites and loops of Fig. 4;

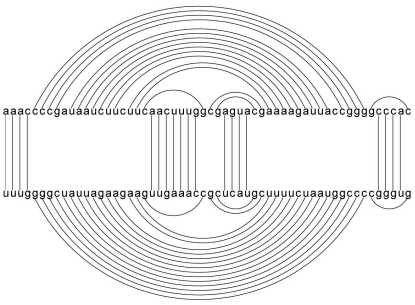


Fig. 5. CopA-CopT.

however, results in one additional binding site and a number of additional loops, which cannot be avoided computationally due to the optimization nature of the problem.

Since CopA and CopT are complementary, performing the algorithm of Section 4.1 will produce the trivial solution where both RNAs bind completely to form a double strand. One can possibly multiply $w(x_i, y_j)$ by an appropriate value $\alpha < 1$ (reducing the weight contribution of external bonds) to cancel this effect. Doing this, however, will explicitly bias the solution toward the independent folding of the two RNAs first. Therefore, the use of such a multiplicative factor is more appropriate with the algorithm of Section 4.2 (or that of Section 4.3). We performed the algorithm of Section 4.2 with $\alpha = 1/3$ and $4 \leq B < \infty$. We obtained the structure illustrated in Fig. 5 which is very close to the known structure of CopA-CopT [6]. Namely, the folding in the middle parts should be replaced by binding, and the folding and binding of the extremities should be ignored. Again, the latter cannot be avoided computationally due to the optimization nature of the problem.

Although the choice of a block size B is important, several block sizes may be tried in the neighborhood of some expected or desired size. Note that a smaller block size (less constrained) does not necessarily imply a better result because the algorithms perform a local optimization followed by a completion on the remaining parts (see description of algorithms in Sections 4.1 and 4.2). Fig. 6 below illustrates the variation in weight for the solutions of fhlA-OxyS and CopA-CopT (using the corresponding algorithms described above) when changing the block size (the lower bound) from 1 to 12, and shows that the choices made above ($B \geq 7$ and $B \geq 4$) are reasonable.

More generally, the stacked pair energy model [7] may be used instead, which favors block formation in both the alignment and the folding, and improves prediction of RNA secondary structure [7]. In principle, the dynamic programming algorithms can be changed to reflect the stacked pair energy model like in [1]. However, the main focus of this paper is on the approximability of the **basic** RNA-RNAi problem described in Section 3 (but the results can be extended to other formulations).

6 Conclusion

The RNA-RNA interaction problem is generally NP-complete. We present 1/2 and 2/3 factor approximation algorithms based on dynamic programming, but there is a need for better algorithms in terms of running time, space, and approximability. In particular, we argue that the mentioned dynamic programming algorithm do not produce entanglers (special molecular substructures), and prove that an entangler-free solution is at best a 2/3 factor approximation. However, experimental results show that variants of these approximation algorithms provide satisfactory structure prediction.

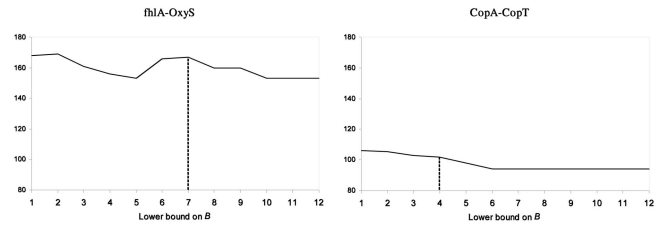


Fig. 6. Effect of block size on total weight.

Acknowledgments

The author would like to thank the CSE Department at SMU for allowing him to develop and teach a course on computational biology, William Westerman for early discussions, Nassim Sohaee from SMU for helping with initial experimentations, Steve Crozier and Skip Garner from UTSW for valuable discussions, Virginia Teller from Hunter College of CUNY for office space, Ioannis Stamos from Hunter College of CUNY for computer resources, Hisham Kassab for valuable input prior to submission, and the anonymous reviewers for their feedback to enhance this manuscript. This work is dedicated to the computing spirit of biology.

References

[1] C. Alkan, E. Karakoc, J.H. Nadeau, C. Sahinalp, and K. Zhang, "RNA-RNA Interaction Prediction and Antisense RNA Target Search," *J. Computational Biology*, vol. 13, pp. 267-282, 2006.

[2] M. Andronescu, R. Aguirre-Hernandes, A. Codon, and H. Hoos, "RNAsoft: A Suite of RNA Secondary Structure Prediction and Design Software Tool," *Nucleic Acid Research*, vol. 31, no. 13, pp. 3416-3422, 2003.

[3] L. Argaman and S. Altuvia, "fhlA Repression by OxyS RNA: Kissing Complex Formation at Two Sites Results in Stable Antisense Target RNA Complex," *J. Molecular Biology*, vol. 300, no. 5, pp. 1101-1112, July 2000.

[4] S.M. Hammond, A.A. Caudy, and G.J. Hannon, "Post Transcriptional Gene Silencing by Double Stranded RNA," *Nature Rev. Genetics*, vol. 2, pp. 110-119, 2001.

[5] D. Hirschberg, "A Linear Space Algorithm for Computing Maximal Common Subsequences," *Comm. ACM*, vol. 18, no. 6, pp. 341-343, 1975.

[6] F.A. Kolb, C. Mamgren, E. Westhof, B. Ehresmann, E.G. Wagner, and P. Romby, "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*, vol. 6, no. 3, pp. 311-324, Mar. 2000.

[7] D. Mathews, J. Sabina, M. Zuker, and D. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure," *J. Molecular Biology*, vol. 288, pp. 911-940, 1999.

[8] S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Molecular Biology*, vol. 48, pp. 443-453, 1970.

[9] R. Nussinov and A. Jacobson, "Fast Algorithm for Predicting the Secondary Structure of Single Stranded RNA," *Proc. Nat'l Academy Sciences USA*, vol. 77, pp. 6309-6313, 1980.
[10] N. Peyret and J. SantaLucia, "Hyther Version 1.0," Wayne State Univ., http://ozone2.chem.wayne.edu/Hyther/hythermenu.html, 2006.
[11] D. Pervouchine, "IRIS: Intermolecular RNA Interaction Search," *Proc. 15th Int'l Conf. Genome Informatics (GIW)*, 2004.
[12] M. Zuker, "Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406-3415, 2003.
[13] M. Zuker, "On Finding All Suboptimal Foldings of an RNA Molecule," *Science*, vol. 244, pp. 48-52, 1989.
[14] M. Zuker, "RNA Folding Lecture Notes," http://www.bioinfo.rpi.edu/~zukerm/lectures/RNAfold-html/index.html, 2008.

Saad Mneimneh's bio and photo are not available.