RNA - RNA Interaction Prediction and Antisense RNA Target Search

Can Alkan[1,2], Emre Karakoç[2], Joseph H. Nadeau[3],S. Cenk Şahinalp[2], and Kaizhong Zhang[4]

[1] Department of EECS, Case Western Reserve University, Cleveland, OH 44106, USA
[2] School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[3] Department of Genetics, Case Western Reserve University, Cleveland, OH 44106, USA
[4] Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada **

Abstract. Recent studies demonstrating the existence of special non-coding "antisense" RNAs used in post-transcriptional gene regulation have received considerable attention. These RNAs are synthesized naturally to control gene expression in C.elegans, Drosophila and other organisms; they are known to regulate plasmid copy numbers in E.coli as well. Small RNAs have also been artificially constructed to knock-out genes of interest in humans and other organisms for the purpose of finding out more about their functions.
Although there are a number of algorithms for predicting the secondary structure of a *single* RNA molecule, no such algorithm exists for reliably predicting the *joint* secondary structure of two interacting RNA molecules, or measuring the stability of such a joint structure. In this paper, we describe the RNA-RNA interaction prediction (RIP) problem between an antisense RNA and its target mRNA and develop efficient algorithms to solve it. Our algorithms minimize the joint free-energy between the two RNA molecules under a number of energy models with growing complexity. Because the computational resources needed by our most accurate approach is prohibitive for long RNA molecules, we also describe how to speed up our techniques through a number of heuristic approaches while experimentally maintaining the original accuracy. Equipped with this fast approach, we apply our method to discover targets for any given antisense RNA in the associated genome sequence.

1 Introduction

Recent studies on both prokaryotic and eukaryotic cells have demonstrated the existence of small RNAs that are used in post-transcriptional gene regulation. These small RNAs usually bind to their target mRNAs to prohibit their translation and, in effect, down-regulate the expression levels of corresponding genes [21]. Other mechanisms for down-regulating gene expression through small RNAs are described in [11].

Regulatory RNAs provide a subclass of the antisense RNA family; other antisense RNAs include snoRNAs, snRNAs, gRNAs, stRNAs, that are naturally used for rRNA modification, RNA editing, mRNA splicing, developmental regulation, and plasmid copy-number regulation. Antisense RNAs are also artificially synthesized for studying specific gene functions since they can knock out targeted genes [21].

Since the first reports on natural antisense RNAs that regulate gene expression (e.g. in C.elegans [19]), there has been substantial interest to better understand how antisense RNAs interact with target mRNAs. In this paper, we describe a new framework

** The authors are listed in alphabetical order.

and corresponding algorithms for predicting the secondary structure of two interacting RNA molecules by means of free energy minimization. We then evaluate how well our algorithms predict known secondary structures of naturally interacting RNA molecules recently observed in *E.coli* [21].

Figure 2a shows the natural joint structure of interacting RNA molecules CopA and CopT. Similarly, Figure 2b shows the natural joint structure of interacting RNA molecules OxyS and fhlA.

In Figures 3a and 4a we present the same interactions in a more illustrative manner: here blue links represent *internal bonds* whereas red links represent *external bonds* between bonded bases. Our predictions of the joint structure of these two RNA molecule pairs are also given in Figures 3 and 4.

There are a number of algorithmic tools for predicting the secondary structure of a *single* RNA molecule [23, 24, 9, 20]; there are also several algorithms to compute "similarity" or "alignment" between two *non-interacting* RNA molecules [14, 4, 22]. However, there are only a few studies related to the problem of predicting the secondary structure formed by two RNA molecules: The HyTher package by the SantaLucia Lab([18]) predicts the hybridization thermodynamics of a given duplex given the two strands; it does not aim to minimize the joint free energy or predict the secondary structure of the interacting RNA strands. The *Pairfold* program [2] aims to predict the secondary structure of two interacting RNA sequences by simply concatenating two RNA strands and performing a secondary structure prediction as if there is only one strand, using the *Mfold* algorithm (for folding a single strand [23, 9]). Because *Mfold* avoids pseudoknots, possible topologies that can be predicted by *PairFold* are very limited; e.g. it can not predict any "kissing" hairpin loops, which are essential to joint structure prediction of two RNA sequences. In principle, *PairFold* can employ the *pknots* method of Rivas and Eddy [20] which can predict certain types of pseudoknots. However the pseudoknot types allowed by *pknots* (as per the characterization in [3]) do not capture any non-trivial kissing loop complex such as the ones explored in this paper. Thus even by employing *pknots*, the *PairFold* approach would not be able to predict the joint structure of interacting RNA molecules of interest.

We recently became aware of a new paper [17] which was published after the submission of our paper. This paper descibes the *IRIS* software tool/algorithm which aims to solve the joint structure prediction problem. *IRIS* is based on a very simple energy model, almost identical to the *basepair energy model*, which we describe as a warmup exercise. As we observed in our own experiments, this approach does not provide good results: the only known natural joint RNA structure examined by [17] is the OxyS-fhlA pair; on this example, the predicted structure by *IRIS* is quite different from the naturally occuring structure.

1.1 Preliminaries and Contributions

In this paper we introduce the general RNA-RNA Interaction Prediction (RIP) Problem. Given two RNA sequences S and R (e.g. an antisense RNA and its target), RIP problem asks to predict their joint secondary structure. A joint secondary structure between S and R is a set of "pairings" where each nucleotide of S and R is paired with at most one other nucleotide, either from S or R.

Let the i^{th} nucleotide of an RNA sequence S be denoted by $S[i]$ and the substring of S extending from $S[i]$ to $S[j]$ denoted by $S[i, j]$. As a notational convenience, let $S[k, k]$ denote $S[k]$, $S[i, i-1]$ denote an empty sequence and $S[i, i-1]^r$ denote the reverse of $S[i-1, i]$. In the rest of the paper, we assume that $S[1]$ denotes the $5'$ end of S and $R[1]$ denotes the $3'$ end of R.

We compute the joint structure between S and R through minimizing their *total free energy* which is, in general, a function of (stacked) pairs of bases as well as the topology of the joint structure. In this paper we consider three models for computing the free energy of the joint structure of interacting RNA sequences.

1. We first use the sum of free energies of individual Watson-Crick base pairs as a crude approximation to the total joint free energy. This *basepair* energy model is quite similar to that used in [15] for predicting the structure of a single RNA molecule. Although the basepair energy model is known to be inaccurate, it provides a good starting point for our further explorations.
2. Our second free energy model is based mostly on *stacked pair* energies given in [10], which provide the main contribution to the energy model employed by the *Mfold* program for pseudoknot free single RNA structure prediction. Unfortunately there is very little thermodynamic information on pseudoknots or kissing loops in the literature. Thus we employ the approach used by Rivas and Eddy [20] to differentiate the thermodynamic parameters of "external" bonds from the "internal" bonds by multiplying the external parameters with a *weight* slightly smaller than 1. This *stacked pair* energy model turns out to be quite accurate, especially in predicting the joint structure of shorter ($\leq$ 150 bases) RNA molecule pairs.
3. The final energy model enriches the above models by summing up the free energies of various types of internal loops and stacked pairs as per [23, 10] as well as the weighted free energies of externally interacting ("kissing") loops. This model, which will be referred to as the *loop* energy model, appears to be more accurate especially for longer ($\geq$ 150 bases) RNA molecules.

Although we allow arbitrary loops to form kissing pairs, we impose the following constraints on the topology of a joint structure between RNA sequences. First, a joint structure can have no *internal pseudoknots*; i.e., if $S[i]$ bonds with $S[j]$ then no $S[i']$ for $i < i' < j$ can bond with any $S[j']$ for $j < j'$. The same property will be satisfied by the nucleotides of R as well. Second, a joint structure can not have any *external pseudoknots*; i.e., if $S[i]$ bonds with $R[j]$ then no $S[i']$ for $i' > i$ can bond with any $R[j']$ for $j' < j$.

These assumptions are satisfied by all examples of complex RNA-RNA interactions we have encountered in the literature. Furthermore allowing arbitrary pseudoknots in the secondary structure of even a single RNA molecule makes the energy minimization problem NP-hard [1]. In fact we prove in Section 2 that the RIP problem is NP-hard for each one of our energy models, even when no internal or external pseudoknots are allowed. This necessitates the addition of one more natural constraint on the topology of the joint secondary structure prediction, which is again satisfied by all known joint structures in the literature. Under this constraint we then show how to obtain efficient algorithms to minimize the free energy of the joint structure under all three energy models and test the accuracy of the algorithms on known joint structures. We finally apply our structure prediction techniques to compute target mRNA sequences to any given small RNA molecule.

2 RIP problem for Both Basepair and Stacked Pair Energy Models is NP-Complete

We start our discussion by showing that the RIP problem is NP-Complete under both the basepair and the stacked pair energy models.

Theorem 1. *RIP problem under the Basepair Energy Model is NP-Complete.*

Proof. The NP-Completeness of RIP is established through a reduction from the longest common subsequence of multiple binary strings (mLCS) which is a known NP-Complete problem. Our proof is an extension to the one in [1] for the single RNA secondary structure prediction problem with pseudoknots.

The decision version of the mLCS problem is as follows: Given a set of *binary* strings $L = \{S_1, S_2, \cdots, S_m\}, (|S_1| = \cdots = |S_m| = n)$ and an integer k, decide whether there exists a sequence C of length k which is a subsequence of each S_i. Here we assume that m is an odd number; if it is even, we simply add a new string $S_{m+1} = S_m$ to L.

From an instance of mLCS, we first construct two "RNA" sequences S and R, using an extended nucleotide alphabet $\Sigma^e = \{a, b, c, d, e, f, u, w, x, y, z\}$. (The NP-hardness proof for the -more interesting- stacked pair energy model below uses the standard RNA nucleotide alphabet $\{A, C, G, U\}$.)

Let v^j denote the string formed by concatenating j copies of character v and let $\overline{v}$ denote the *complementary residue* of v. In our extended alphabet we set $\overline{x} = w$, $\overline{y} = z$, $\overline{a} = b$, $\overline{c} = d$, and $\overline{e} = f$. Given a string T, we denote by $\overline{T}$ its *reverse complement*.

For $i = 1, \cdots, m$, we construct strings D_i and E_i as follows. Note that we set $s_{i,j}$ to x if the j^{th} character of string S_i is 0; if it is 1, $s_{i,j}$ is set to be y.

$$
\begin{array}{ll}
D_i = a\ s_{i,1}\ a\ s_{i,2}\ a\ \cdots\ a\ s_{i,n}\ a, & if\ i\ is\ odd; \\
D_i = a\ \overline{s_{i,n}}\ a\ \overline{s_{i,n-1}}\ a\ \cdots\ a\ \overline{s_{i,1}}\ a, & if\ i\ is\ even; \\
E_i = b\ \overline{s_{i,1}}\ b\ \overline{s_{i,2}}\ b\ \cdots\ b\ \overline{s_{i,n}}\ b, & if\ i\ is\ odd; \\
E_i = b\ s_{i,n}\ b\ s_{i,n-1}\ b\ \cdots\ b\ s_{i,1}\ b, & if\ i\ is\ even.
\end{array}
$$

We now construct the RNA sequences S and R as follows.

$$
\begin{array}{l}
S = u^k, D_1, c^1, D_2, D_3, d^1, c^2, D_4, D_5, d^2 \cdots c^{(m-1)/2}, D_{m-1}, D_m, d^{(m-1)/2} \\
R = e^1, E_1, E_2, f^1, e^2, E_3, E_4, f^2 \cdots e^{(m-1)/2}, E_{m-2}, E_{m-1}, f^{(m-1)/2}, E_m, u^k
\end{array}
$$

Note that the lengths of S and R are polynomial with the total size of all sequences $S_1 \ldots S_m$.

We now set the energy function for bonded nucleotides pairs. The bond between each nucleotide with its complement has a free energy of -1.0. The bond between u with x, y, z, w also has a free energy of -1.0. For other bonds between nucleotide pairs, the free energy is 0.0.

In the basepair energy model, the free energy of the overall structure is defined to be the sum of the free energies of all bonded pairs of nucleotides. Thus, according to the above setting, each nucleotide other than u will tend to get bonded with their complementary nucleotides, and u will tend to get bonded with any of x, y, z, w and vice versa. We call such bondings *valid* bondings. The free energy of the joint structure is minimized when the number of valid bondings between nucleotide pairs is maximized.

We now show that there exists a common subsequence of length k among $S_1, \ldots, S_m$ if and only if there exists a joint secondary structure of S and R where *every* nucleotide forms a valid bonding.

Suppose that $S_1 \ldots S_m$ have a common subsequence C of length k; we can construct a secondary structure of S and R where every nucleotide forms a valid bonding as follows.

- For each i, form a bond between the i^{th} a in S with the i^{th} b in R.
- For each i, bond the substring c^i to the substring d^i in S and bond the substring e^i to the substring f^i in R.

– For each string $S_i \in L$ there is a corresponding substring D_i in S and E_i (which is the complement of D_i) in R. Consider for each S_i the sequence that remains when the common subsequence C is deleted out; denote this sequence by C'. Bond each nucleotide in D_i that corresponds to a character in C' to its corresponding complementary nucleotide in E_i.
– All that remains in S and R are those nucleotides that correspond to the common subsequence C in each string S_i. There is also the substring u^k at the left end of S and another substring of the form u^k at the right end of R. Bond the u^k block in S to the unbonded nucleotides (that correspond to C) in D_1. For all $1 \leq i \leq (m-1)/2$, bond the unbonded nucleotides in E_{2i-1} to those in E_{2i}. Similarly bond the unbonded nucleotides in D_{2i} to those in D_{2i+1}. Finally bond the unbonded nucleotides in E_m to the u^k block in R.

The reader can easily verify that our construction establishes a valid bonding for all nucleotides in S and R. The process of constructing S and R and establishing the bonds described above is demonstrated in Figure 1. Here $L = \{s_1 = xyxx,\ s_2 = xxyx,\ s_3 = xyyx\}$.

Now we show that if there is a joint secondary structure between S and R where every nucleotide forms a valid bonding, then there is a common subsequence of strings $S_1, S_2, \cdots, S_m$ of length k.

– Nucleotides a and b are complementary and do not form bonds with u. S only has as and R only has bs. If all as and bs form valid bonds, the i^{th} a *must* form a bond with the i^{th} b.
– Nucleotides c, d only occur in S and only form valid bonds with each other. Because we do not allow internal pseudoknots, each c^i block will be bonded with the d^i block. Similarly, nucleotides e, f only occur in R and only form valid bonds with each other. Again, because there are no internal pseudoknots, each e^i block will be bonded with the f^i block.
– The above bondings necessitate that nucleotides of the u^k block in S must bond with those in D_1 and nucleotides of the u^k block in R must bond with those in E_m. The remaining nucleotides of D_1 must bond with corresponding nucleotides in E_1 and the remaining nucleotides of E_m must bond with corresponding nucleotides in D_m.
– The nucleotides that are left in E_1 are the nucleotides that correspond to those in D_1 which have been bonded to u^k block - they must be bonded to complementary nucleotides in E_2. The bonds between E_1 and E_2 corresponds to a common subsequence of S_1 and S_2 of size k.
– Inductively, for $i = 1 \ldots (m-1)/2$, the nucleotides left out in E_{2i} must form bonds with corresponding nucleotides in D_{2i}. The ones that are left out in D_{2i} must form bonds with complementary nucleotides in D_{2i+1}. The bonds between D_{2i} and D_{2i+1} corresponds to a common subsequence of S_{2i} and S_{2i+1}.
– Similarly, the nucleotides left out in D_{2i+1} must form bonds with corresponding nucleotides in E_{2i+1}. The ones that are left out in E_{2i+1} must form bonds with complementary nucleotides in E_{2i+2}. The bonds between E_{2i+1} and E_{2i+2} corresponds to a common subsequence of S_{2i+1} and S_{2i+2}.
– Finally, the nucleotides that are left out in E_m must be bonded to nucleotides in u^k block in R.

The bonds between consecutive D_i, D_{i+1} pairs and E_i, E_{i+1} pairs correspond to common subsequences between S_i and S_{i+1}. Thus the strings $S_1, \ldots, S_m$ must have a common subsequence of length k.

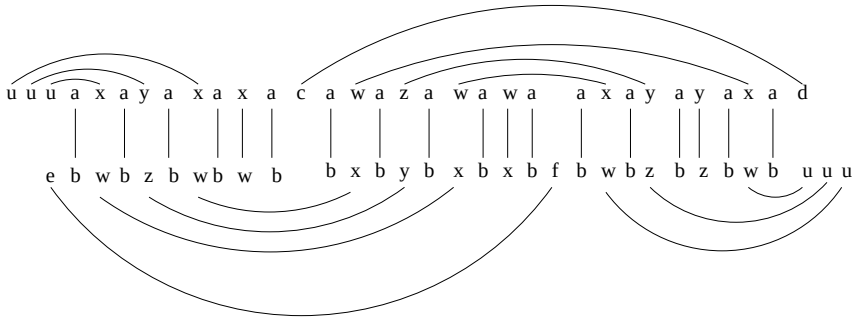


Fig. 1. Sample RIP solution for mLCS problem on $S_1 = \{xyxx\}, S_2 = \{xxyx\}, S_3 = \{xyyx\}$. The mLCS is determined with the internal bondings, here it is xyx.

We now establish the NP-hardness of the RIP problem under the stacked pair energy model.

Theorem 2. *RIP problem under the Stacked Pair Energy Model is NP-Complete.*

Proof. The proof is through an indirect reduction from the mLCS problem as per Theorem 1. Consider the reduction of the mLCS problem to the RIP problem under the basepair energy model. Given sequences S and R that were obtained as a result of this reduction, we construct two new RNA sequences S' and R' from the standard nucleotide alphabet by replacing each character in S and R with quadruplets of nucleotides as follows: $a \leftarrow CCGU$, $b \leftarrow GGCU$, $c \leftarrow GCCU$, $d \leftarrow CGGU$, $e \leftarrow CGCU$, $f \leftarrow GCGU$, $u \leftarrow AAAU$, $x \leftarrow ACAU$, $z \leftarrow CACU$, $y \leftarrow AGAU$, $w \leftarrow GAGU$.

We now determine the energy function for stacked pairs of nucleotides. The free energy of the following stacked pairs are all set to -0.5:
$(A-A, A-C), (A-A, C-A), (A-A, A-G), (A-A, G-A), (A-C, A-A), (A-C, C-A), (A-G, A-A), (A-G, G-A), (C-A, A-A), (C-A, A-C), (C-G, C-G), (C-G, G-C), (G-A, A-A), (G-A, A-G), (G-C, C-G), (G-C, G-C)$.
For other bondings between nucleotides, the free energy is set to 0.0. Thus bonding U with any nucleotide will not reduce the free energy of the joint structure.

In the stacked pair energy model, the free energy of the overall structure is defined to be the sum of the free energies of all stacked pairs of bonded nucleotides. The reader can verify that above setting of stacked pair energies ensure that the bonds between the characters of S and R presented in Theorem 1 will be preserved between S' and R'. (e.g. a bond between a and b has free energy -1.0. Because a corresponds to $CCGU$ and b corresponds to $GGCU$, the stacked pairs obtained will be $(C-G, C-G)$ and $(C-G, G-C)$ each with free energy -0.5. The total free energy will thus be -1.0.)

2.1 Additional topological constraints on joint structures

The hardness of the RIP problem under both basepair and stacked pair energy models necessitate one more constraint on the topology of the interaction between two RNA molecules. Based on our observations of known joint structures of RNA molecule pairs in Figure 2, we impose the following constraint (which is satisfied by all known structures in the literature). Let $S[i]$ be bonded with $S[j]$ and $R[i']$ be bonded with $R[j']$. Then exactly one of the following must be satisfied:

1. There are no $i < k < j$ and $i' < k' < j'$ such that $S[k]$ bonds with $R[k']$.
2. For all $i < k < j$, if $S[k]$ bonds with some $R[k']$ then $i' < k' < j'$.
3. For all $i' < k' < j'$, if $R[k']$ bonds with some $S[k]$ then $i < k < j$.

The condition simply states that if two "substructures" $S[i, j]$ and $R[i', j']$ interact, then one must "subsume" the other. A joint structure of two RNA sequences S and R is considered to be *valid* if all above conditions are satisfied.

3 Structure prediction in the Basepair Energy Model

The basepair energy model approximates the free energy of the joint structure between interacting RNA molecules as the sum of the free energies of bonded nucleotide pairs. We denote the Watson-Crick free energy of a bond between nucleotides x and y by $e(x, y)$ if they are on the same RNA strand (this is called an *internal bond*) and by $e'(x, y)$ if they are on different strands (this is called an *external bond*). Although in our experiments we set $e' = e$, our formulation also allows to differentiate these two energy functions. Below, we obtain a valid pairing between the nucleotides of S and R that minimizes the free energy of their joint structure through the computation of $E(S[i, j], R[i', j'])$ the free energy between interacting RNA strands $S[i, j]$ and $R[i', j']$ for all $i < j$ and $i' < k'$. Clearly E gives the overall free energy between S and R when $i = i' = 1$ and $j = |S|$ and $j' = |R|$. We set $E(S[i, i], R[i', i'])$ to $e'(S[i], R[i'])$ and compute the value of $E(S[i, j], R[i', j'])$ inductively as the minimum of the following:

1. $\min_{i-1 \le k \le j; i'-1 \le k' \le j'' : (k \ne i-1 \text{ or } k' \ne i'-1), (k \ne j \text{ or } k' \ne j')} E(S[i, k], R[i', k']) + E(S[k + 1, j]), R[k' + 1, j'])$.
2. $E(S[i + 1, j - 1], R[i', j']) + e(S[i], S[j])$.
3. $E(S[i, j], R[i' + 1, j' - 1]) + e(R[i'], R[j'])$.

Lemma 1. *The above dynamic programming formulation is correct.*

Proof. There are two cases to deal with:

1. Consider the case that either $S[i]$ or $S[j]$ or $R[i']$ or $R[j']$ bonds with a nucleotide on the other RNA strand. Wlog, let $S[i]$ bond with $R[h']$; then either (i) $R[i']$ bonds with $R[j']$ for which condition (3) will be satisfied, or (ii) $i' = h'$ so that $R[i']$ bonds with $S[i]$ for which condition (1) will be satisfied for $k = i$ and $k' = i'$, or (iii) $R[i']$ bonds with some $R[\ell']$ for which condition (1) will be satisfied for some "breakpoint" $S[k]$, $R[k']$, for $i \le k \le j$ and $i' \le k' \le j'$ such that $S[i, k]$ interacts only with $R[i', k']$ and $S[k + 1, j]$ interacts only with $R[k' + 1, j']$.
2. If the above condition is not satisfied then wlog we can assume that $S[i]$ bonds with $S[h]$ and $R[i']$ bonds with $R[h']$. If for no $\ell > h$, $S[\ell]$ interacts with any $R[\ell']$ for $\ell' > h'$ then condition (1) will be satisfied with $k = h$ and either $k' = i' - 1$ or $k' = j' + 1$. If for no $\ell < h$, $S[\ell]$ interacts with any $R[\ell']$ for $\ell' < h'$ then condition (1) will be satisfied again with $k = h$ and either $k' = i' - 1$ or $k' = j' + 1$. The possibility of none of these two cases hold is excluded by our topological constraints described earlier.

Lemma 2. *The table E can be computed in time $O(|S|^3 \cdot |R|^3)$ and in space $O(|S|^2 \cdot |R|^2)$.*

3.1 Testing the Basepair Energy Model

We tested the basepair energy model on naturally occurring joint structures of interacting RNA molecule pairs CopA-CopT and OxyS-fhlA. Our results are given in Figures 3b and 4b. Perhaps not surprisingly, the predicted joint structures by the Basepair Energy Model is quite different from the natural secondary structures (Figures 3a and 4a). Observe that in natural joint structures, internal or external bonds usually form stacked pairs; i.e., a bond $S[i]-S[j]$ usually implies bonds $S[i+1]-S[j-1]$ and $S[i-1]-S[j+1]$. Similarly a bond $S[i]-R[i']$ usually implies bonds $S[i+1]-R[i'+1]$ and $S[i-1]-R[i'-1]$. Furthermore, in natural joint structures unbonded nucleotides seem to form uninterrupted sequences rather than being scattered around.

4 Structure Prediction Based on Stacked Pair Energy Model

The limitations of the Basepair Energy Model promotes the use of a Stacked Pair Energy Model where the bonds between nucleotide pairs form uninterrupted sequences. We denote by $ee(X[i,i+1],X[j-1,j])$ the energy of the internal stacked pair $(X[i]-X[j],X[i+1]-X[j-1])$ and by $ee'(X[i,i+1],Y[j,j+1])$ the energy of the external stacked pair $(X[i]-Y[j],X[i+1]-Y[j+1])$. As per the *pknots* approach [20] we set $ee'=\sigma\cdot ee$ for a user defi ned weight parameter $0<\sigma\leq 1$ (externally kissing pairs are similar in nature to pseudoknots). The thermodynamic free energy parameters we used in our tests are taken from [10]. Note that the energy functions $ee(.,.)$ and $ee'(.,.)$ are not symmetric; they can differ according to the relative directions of the stacked pairs ($3'-5'$ or $5'-3'$) involved.

To compute the joint structure between S and R under the Stacked Pair Energy Model we will need to introduce four energy functions.

1. $E_S(S[i,j],R[i',j'])$ denotes the free energy between S and R such that $S[i]$ bonds with $S[j]$.
2. $E_R(S[i,j],R[i',j'])$ denotes the free energy between S and R such that $R[i']$ bonds with $R[j']$.
3. $E_l(S[i,j],R[i',j'])$ denotes the free energy between S and R such that $S[i]$ bonds with $R[i']$.
4. $E_r(S[i,j],R[i',j'])$ denotes the free energy between S and R such that $S[j]$ bonds with $R[j']$.

The overall energy is then defi ned to be:

$$E(S[i,j],R[i',j'])=\min\left\{\begin{array}{l}E_S(S[i,j],R[i',j']),\ \ E_R(S[i,j],R[i',j']),\\ E_r(S[i,j],R[i',j']),\ \ E_l(S[i,j],R[i',j']),\\ \min_{i\leq k\leq j-1;i'\leq k'\leq j'-1}\left\{\begin{array}{l}E(S[i,k],R[i',k'])+\\ E(S[k+1,j]),R[k'+1,j'])\end{array}\right\}\\ \min_{i\leq k\leq j-1}\left\{\begin{array}{l}E(S[i,k],-)+\\ E(S[k+1,j]),R[i',j'])\end{array}\right\}\\ \min_{i\leq k\leq j-1}\left\{\begin{array}{l}E(S[i,k],R[i',j'])+\\ E(S[k+1,j]),-)\end{array}\right\}\\ \min_{i'\leq k'\leq j'-1}\left\{\begin{array}{l}E(S[i,j],R[i',k'])+\\ E(-,R[k'+1,j'])\end{array}\right\}\\ \min_{i'\leq k'\leq j'-1}\left\{\begin{array}{l}E(-,R[i',k'])+\\ E(S[i,j]),R[k'+1,j'])\end{array}\right\}\end{array}\right\}$$

Now we show how to compute E_S,E_R,E_r,E_l via dynamic programming. The initial settings of the energy functions are determined as follows.

$$
\begin{array}{ll}
E_l(S[i,j],-) = \infty & E_r(S[i,j],-) = \infty \\
E_l(-,R[i',j']) = \infty & E_r(-,R[i',j']) = \infty \\
E_l(S[i,i],R[i',i']) = 0 & E_r(S[i,i],R[i',i']) = 0
\end{array}
$$

$$
E_S(S[i,i],-) = \infty \qquad E_R(-,R[i',i']) = \infty
$$

Now we give the complete description of the dynamic programming formulation. Note that because sequence R is assumed to be in $3'-5'$ direction, reversing the stacked pairs involved is required for the correct use of ee function in E_R.

$$
E_l(S[i,j],R[i',j']) = \min \left\{ \begin{array}{c} E_l(S[i+1,j],R[i'+1,j']) + ee'(S[i,i+1],R[i',i'+1]), \\ E(S[i+1,j],R[i'+1,j']) \end{array} \right\}
$$

$$
E_r(S[i,j],R[i',j']) = \min \left\{ \begin{array}{c} E_r(S[i,j-1],R[i',j'-1]) + ee'(S[j-1,j],R[j'-1,j']), \\ E(S[i,j-1],R[i',j'-1]) \end{array} \right\}
$$

$$
E_S(S[i,j],R[i',j']) = \min \left\{ \begin{array}{c} E_S(S[i+1,j-1],R[i',j']) + ee(S[i,i+1],S[j-1,j]), \\ E(S[i+1,j-1],R[i',j']) \end{array} \right\}
$$

$$
E_R(S[i,j],R[i',j']) = \min \left\{ \begin{array}{c} E_R(S[i,j],R[i'+1,j'-1]) + ee(R[j'-1,j']^r,R[i',i'+1]^r), \\ E(S[i,j],R[i'+1,j'-1]) \end{array} \right\}
$$

The following lemmas follow from the DP formulation above and their proofs are left for the full version for the paper.

Lemma 3. *The above dynamic programming formulation is correct.*

Lemma 4. *The tables E_S, E_R, E_l, E_r and the overall energy table E can be computed in time $O(|S|^3 \cdot |R|^3)$ and in space $O(|S|^2 \cdot |R|^2)$.*

As will be discussed below, the Stacked Pair Energy Model formulation works very well with the joint structure prediction problems considered in this paper. However this formulation does not necessarily aim to cluster gaps in uninterrupted sequences, as observed in natural joint structures. Thus, we also provide a more general formulation for the Stacked Pair Energy Model, that employs an "affine" cost model for the gaps involved. Also considered in this formulation are penalties for switching from internal to external bonds (and vice versa). This general formulation does not necessarily improve our predictions for the joint structures considered in this paper; however it could be useful for other examples and thus is provided in the Appendix.

4.1 Testing Stacked Pair Energy Model

The Stacked Pair Energy Model as defined above has only one user defined parameter (as per [20]), σ, which is the ratio between the free energies of internal and external stacked pairs. Unfortunately no miracle prescription for determining the right value of σ is available (see for example [20]). It is possible to approximately determine the value for σ by closely inspecting the natural joint structure of CopA-CopT pair (Figure 3a). CopA and CopT sequences are perfectly complimentary to each other, thus they can, in principle, form a stable duplex structure that would prevent any internal bonding pairs, leaving out just *one* nucleotide unbonded. However, as one can observe from Figure 3a this does not happen. The ratio between the length of the external bonding sequences in the joint structure and that of the internal bonding sequences implies that $\sigma \in [0.7, 0.8]$.

Under these observations we tested our algorithm that implements the Stacked Pair Energy Model with $\sigma \in [0.7, 0.8]$. The secondary structures predicted by our algorithm

on CopA-CopT and OxyS-fhlA pairs are given in Figures 3c and 4c. As one can observe, there are only very slight differences between the natural joint structure and the predicted joint structure of the RNA pairs. For example, the predicted joint structure of OxyS-fhlA pair (Figure 4c) has 53 internal-bonds, 14 external-bonds, and 23 unbonded nucleotides. In *all* aspects, these figures are superior to the natural joint structure of the pair (Figure 4a), which has 50 internal-bonds, 16 external-bonds, and 25 unbonded nucleotides. Because the external bond scores are smaller than internal ones, under *any selection of* $\sigma < 1$ the prediction of our algorithm results in a higher score/lower free energy than that implied by the natural joint structure of OxyS-fhlA pair. Nevertheless, the differences between the natural structures and the predicted ones are very small implying that the Stacked Pair Energy Model can be used as the central tool of our RNA target prediction algorithm.

5 Structure Prediction Based on Loop Energy Model

The structure prediction algorithm to find the optimal joint structure between two RNA molecules based on Stacked Pair Energy Model requires substantial resources in terms of running time and memory. On a Sun Fire v20z server with 16GB RAM and AMD Opteron 2.2GHz processor, the running time for predicting the joint secondary structure of OxyS-fhlA pair is 15 minutes; this could be prohibitive for predicting the targets of sufficiently long RNA molecules. In this section we make a number of observations on the natural joint structures of RNA molecule pairs for speeding up our approach through heuristic shortcuts - without losing its (experimental) predictive power.

An interesting observation is that the (predicted) self structures are mostly preserved in the joint secondary structures. In fact, external interactions only occur between pairs of predicted hairpins. Thus it may be be sufficient to compute the joint structure of two RNA sequences by simply computing the set of loop pairs that form bonds to minimize the total joint free energy.

The above observation prompts an alternative, simpler approach which is described below. This new approach maintains that each RNA sequence will tend to preserve much of its original secondary structure after interacting with the other RNA sequence, which is achieved by means of preserving what we call "independent subsequences" that form hairpins. More formally:

Definition 1. *Independent Subsequences:*

Given an RNA sequence R and its secondary structure, the substring $R(i, j)$ is an independent subsequence of R if it satisfies the following conditions.

- $R[i]$ is bonded with $R[j]$.
- $j - i \leq \kappa$ for some user specified length κ.
- There exists no $i' < i$ and $j' > j$ such that $R[i']$ is bonded with $R[j']$ and $j' - i' \leq \kappa$. (This condition prohibits overlaps between independent subsequences).

It is possible to compute the (locations of) independent sequences of a given RNA molecule, from its secondary structure predicted by *Mfold*, through a simple greedy algorithm as follows.

1. Let IS be the set of independent subsequences in R; initially we set $IS = \emptyset$.
2. Starting from the first nucleotide of R find the first nucleotide $R[i]$ which bonds with another nucleotide $R[j]$, ($j > i$).

3. If $j - i \leq \kappa$ then update $IS = IS \cup R[i, j]$ and move to $R[j + 1]$.
   Else move to $R[i + 1]$.
4. Repeat Step 2.

The proofs of the following are quite easy to obtain and hence are not given.

Lemma 5. *The above algorithm finds the correct independent subsequences.*

Lemma 6. *Given the secondary structure of an RNA sequence R, its independent subsequences can be computed in $O(|R|)$ time via the above algorithm.*

5.1 Computing the Interactions between Independent Subsequences

In our new model, the external bondings between nucleotide pairs will be permitted among the independent subsequences of the two RNA sequences S and R, predicted by *Mfold*. Below we show how to compute the external bonds between such nucleotides which minimize the total free energy in the interacting RNA sequences.

From this point on we will treat each RNA molecules as an (ordered) set of independent subsequences (IS), where each IS is indeed a string of nucleotides. The i^{th} IS of an RNA molecule S is denoted by $S_{IS}[i]$. The sequence of ISs between $S_{IS}[i]$ and $S_{IS}[j]$ are thus denoted as $S_{IS}[i, j]$.

We calculate the joint structure between R and S by minimizing the total free energy of their ISs via means of establishing bonds between their nucleotides as follows. Let the minimum free energy of the joint secondary structure of the two ISs $S_{IS}[i]$ and $R_{IS}[j]$ be $e_{IS}(i, j)$. The value of $e_{IS}(i, j)$ can be computed via the algorithm we described in Section 4.

The minimum joint free energy between the consecutive sets of ISs of R and S is calculated once $e_{IS}(i, j)$ is computed for all i, j. Let n and m denote the number of ISs in S and R respectively. Now let $E(S_{IS}[i], R_{IS}[j]) = E[i, j]$ be the smallest free energy of the interacting independent subsequence lists $S_{IS}[1, i]$ and $R_{IS}[1, j]$ (which satisfy the distance constraint) provided that $S_{IS}[i]$ and $R_{IS}[j]$ interact with each other.

Before we show how to compute the values of $E[i, j]$, we make one final observation on the OxyS-fhlA pair that the "distance" between two interacting subsequences in OxyS appears to be be very close to that in fhlA. This may be due to the limited flexibility of "root stems" that support the independent subsequences when they interact with each other. In order to ensure that the predictions made by our algorithm satisfy such limitations we impose restrictions on the "distances" between interacting independent subsequences as follows.

Definition 2. *Let $S_{IS}[i]$ and $S_{IS}[j]$ be two independent subsequences in a given RNA sequence S. The distance between $S_{IS}[i]$ and $S_{IS}[j]$, denoted $d(S_{IS}[i], S_{IS}[j])$ is defined as the number of nucleotides $S[k]$ that do not lie between a bonded pair of nucleotides $S[h]$ and $S[h']$ that are both located between $S_{IS}[i]$ and $S_{IS}[j]$.*

The above definition simply ignores all nucleotides that lie in the independent subsequences between $S_{IS}[i]$ and $S_{IS}[i']$ regardless of their lengths. Our algorithm ensures that if $S_{IS}[i] - R_{IS}[j]$ and $S_{IS}[i'] - R_{IS}[j']$ are pairs of consecutive independent subsequences that interact with each other and if $d(S_{IS}[i], S_{IS}[i']) \geq d(R_{IS}[j], R_{IS}[j'])$ then $d(S_{IS}[i], S_{IS}[i']) \leq (1 + \epsilon) \cdot d(R_{IS}[j], R_{IS}[j']) + \delta$; here $\epsilon < 1$ and $\delta > 0$ are user defined constants.

The value of $E[i,j]$ can be computed through dynamic programming as follows.

$$E[i,j] \;=\; \min_{i'<i, j'<j \;\mid\; d(S_{IS}[i'],S_{IS}[i]) \le (1+\epsilon)\cdot d(R_{IS}[j'],R_{IS}[j]) + \delta} \left(\begin{array}{l} E[i',j'] + e_{IS}(i,j) \;+ \\ \sum_{i'<i''<i} e_{IS}(i'',0) \;+ \\ \sum_{j'<j''<j} e_{IS}(0,j''). \end{array} \right)$$

Here $e_{IS}(i'',0)$ and $e_{IS}(0,j'')$ denote the free energy of independent subsequences $S_{IS}[i'']$ and $R_{IS}[j'']$ respectively. The overall free energy of the interacting independent subsequence sets of R and S is thus:

$$\min_{\forall i,j} E[i,j] \;+\; \sum_{i<i'} e_{IS}(i',0) \;+\; \sum_{j<j'} e_{IS}(0,j').$$

The following lemmas are easy to verify and are left to the full version of the paper.

Lemma 7. *The above dynamic programming formulation correctly computes $E[i,j]$.*

Lemma 8. *Given $e_{IS}(i,j)$ for all i,j, the values of E can be computed in time $O(n^3 \cdot m^3)$. As the computation of $e_{IS}(i,j)$ takes $O(\kappa^6)$ time the overall running time of the algorithm is $O(n \cdot m \cdot \kappa^6 + n^3 \cdot m^3)$.*

Because $n \le |S|/\kappa$ and $m \le |R|/\kappa$ the worst case running time of this algorithm is $O(|S| \cdot |R| \cdot \kappa^4 + |S|^3 \cdot |R|^3/\kappa^6)$. This is substantially faster than our earlier approach requiring $O(|S|^3 \cdot |R|^3)$ time. In fact this version can predict the joint structure of the OxyS-fhlA pair in 5 seconds, improving our earlier approach by a factor of 180.

5.2 Testing the Loop Energy Model

We tested our third model on the interacting RNA pairs CopA-CopT and OxyS-fhlA, with the same σ values we used in Stacked Pair Energy Model: $\sigma \in [0.7, 0.8]$. Joint structure predictions obtained by Loop Energy Model are given in Figures 3d for CopA-CopT pair, and 4d for OxyS-fhlA pair. Although there is a slight loss in the prediction quality in CopA-CopT pair with respect to the Stacked Pair Energy Model prediction (Figure 3c), the "kissing" hairpin sequence is predicted correctly. In our OxyS-fhlA test, notice that the predictions obtained by the Loop Energy Model and the Stacked Pair Energy Model are even more similar. Furthermore, careful observation shows that the total free energy in the predicted structure is still better than the natural joint structure (Figure 4a).

6 Target Prediction for Antisense RNAs

An important byproduct of our algorithms for the RIP problem is the ability to search for target sequences for specific antisense RNA molecules in whole genomic and plasmid sequences. Because of the time and space constraints, the Stacked Pair Energy Model is not efficient when searching through large sequences. Therefore, in our target prediction approach is based on Loop Energy Model. Our search strategy employs the following steps:

1. First, we need to find the "candidate" target sequences from a given genome sequence (or plasmid) that is known to include the target. This is achieved via using the cDNA annotation available for genomic sequences. To compute the potential mRNA each such cDNA is extended towards $5'$ and $3'$ UTR ends as follows.

(a) Each cDNA is extended up to l_1 nucleotides at its $5'$ UTR, and by l_2 nucleotides at its $3'$ UTR, where l_1 and l_2 are user defined parameters. (In our experiments we set $l_1 = 250$ and $l_2 = 25$).
(b) Then each "extended" cDNA sequence is trimmed from both ends via a dynamic programming routine in order to compute its subsequence which has the lowest "energy density" (this will be the subsequence of the extended cDNA sequence whose secondary structure is most stable.) We predict the resulting mRNA of each such cDNA as its trimmed extension. The details of this mRNA prediction approach (from a given cDNA) is of independent interest and is left to the full version of the paper.

2. We then run our joint secondary structure prediction algorithm based on Loop Energy Model to determine if there are any external bonds formed between each candidate target sequence and the antisense RNA sequence under the following constraints. (1) At least one IS in the candidate target sequence which lies before the start codon (i.e. AUG) should interact with an independent subsequence in the query sequence. We impose this constraint in order to capture the ribosome binding site interactions. (2) All predicted interactions between pairs of ISs should include at least ξ uninterrupted bonds for some user specified constant ξ. We impose this constraint to favor long uninterrupted external bonds, since ribosomes are capable of breaking shorter interactions. (3) At least two pairs of independent sequences must be interacting with each other.

We tested the above approach on both RNA-RNA interactions that we considered in the paper. (1) We first searched the target mRNA sequences for CopA in the R1 plasmid sequence in *E.coli*. It is known that CopA regulates the copy number of R1 plasmid by binding to the CopT sequence which is a part of the 125Kb long plasmid [6, 21]. Our program needed about 40 hours on a PC equipped with 2 Ghz Pentium IV processor and 1 GB of main memory to detect all targets of the CopA sequence on the complete R1 plasmid. Out of the 141 potential mRNA segments obtained from the annotated cDNA sequences it returned only the correct target CopT as a potential target.
(2) We then used our program to detect the target mRNA sequences of the OxyS antisense RNA on a 130Kb long portion of E.coli genome that included the known target fhlA [16]. Out of the 100 potential mRNA segments obtained from the annotated cDNA sequences, our program returned 9 hits including the known target fhlA.

Notice that the joint structure between CopA and CopT are much more stable than that between OxyS and fhlA (the former one has a half-life of about an hour where as the latter one has a half-life of only a couple of minutes). It is possible that OxyS may have other targets in the E.coli genome with which it may establish unstable joint structures, not strong enough to make impact. We are aiming to test whether such *in silico* interactions between OxyS and its 8 unknown targets (i.e. those predicted by our program) actually take place *in vitro* or *in vivo* in the future.

Acknowledgments. We would like to thank the anonymous referees for pointing out the *IRIS* software package to our attention.

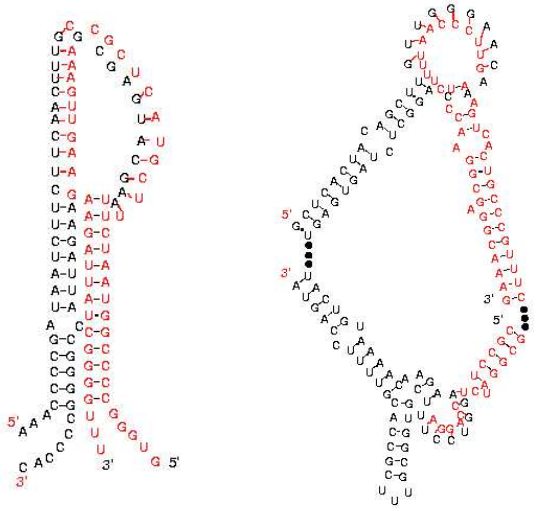
References

1. Akutsu T., Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, *Discrete Applied Mathematics*, 104:45-62, 2000.
2. Andronescu M., Aguirre-Hernandes R., Condon A., and Hoos H., RNAsoft: a suite of RNA secondary structure prediction and design software tools, *Nucleic Acids Research*, 31(13):3416-3422, 2003.

3. Condon A., Davy B., Rastegari B., Zhao S., Tarrant F., Classifying RNA pseudoknotted structures. *Theoretical Computer Science*, 320(1): 35-50, 2004.
4. Collins G., Le S., Zhang, K., A new algorithm for computing similarity between RNA structures, *Proc. 5th Joint Conf. on Information Science*, Atlantic City, NJ, vol. 2, pp. 761-765, March, 2000.
5. Kim, C.-H., and Tinoco Jr., I. A Retroviral RNA Kissing Complex Containing Only Two G-C Base Pairs, *Proc.Nat.Acad.Sci*. USA 97 pp. 9396, 2000.
6. Kolb, F.A., Engdahl, H.M, Slagter-Jager, J.G., Ehresmann, B., Ehresmann, C., Westhof, E., Wagner, E.G.H., and Romby, P., Progression of a loop-loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA, *EMBO Journal*, 19(21):5905-5915, 2000.
7. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T., Identification of novel genes coding for small expressed RNAs, *Science*, 294:853-857, 2001.
8. Lau, N.C., Lim, L.P., Weinstein, E.G., Bartel, D.P., An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, *Science*, 294:858-862, 2001.
9. Lyngso, R.B., Zuker, M., and Pedersen, C.N.S. Fast evaluation of internal loops in RNA secondary structure prediction, *Bioinformatics*, 15:440-445, 1999.
10. Mathews, D., Sabina, J., Zuker, M. and Turner, D., Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911-940, 1999.
11. McManus, Michael T., and Sharp, Phillip A., Gene silencing in mammals by small interfering RNAs, *Nature Reviews Genetics*, 3(10):737-747, 2002.
12. Moss, Eric G., RNA interference: It's a small RNA world, *Current Biology*, 11:R772-R775, 2001.
13. Moss, Eric G., MicroRNAs: Hidden in the Genome, *Current Biology*, 12:R138-R140, 2002.
14. Notredame, C., O'Brien, E.A., and Higgins, D.G., RAGA: RNA sequence alignment by genetic algorithm, *Nucleic Acids Research*, 25(22):4570-4580, 1997.
15. Nussinov, R. and Jacobson, A. Fast algorithm for predicting the secondary structure of single stranded RNA, *PNAS*, 77:6309-6313, 1980.
16. NCBI web site, *http://www.ncbi.nlm.nih.gov*
17. Pervouchine, Dmitri D. IRIS: Intermolecular RNA Interaction Search, *15th Int. Conf. Genome Informatics*, 2004.
18. Peyret N. and SantaLucia, J., HYTHER^{TM} version 1.0, *http://ozone2.chem.wayne.edu/Hyther/hythermenu.html*, Wayne State University.
19. Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G., The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans, *Nature*, 403:901-906, 2000.
20. Rivas, E., and Eddy, S.R., A dynamic programming algorithm for RNA structure prediction including pseudoknots, *J Mol Biol.*, 285(5):2053-68, 1999.
21. Wagner, E.G.H., and Flardh, K., Antisense RNAs everywhere?, *TRENDS in Genetics*, 18(5):223-226, 2002.
22. Zhang, K., Wang, L., and Ma, B., Computing similarity between RNA structures, *Theoretical Computer Sciences*, 276(1-2):111-132, 2002. March, 2000.
23. Zuker,M. and Stiegler,P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133148, 1981.
24. Zuker, Michael, On finding all suboptimal foldings of an RNA molecule, *Science*, 244:48-52, 1989.

A A More General Stacked Pair Energy Formulation.

Our more general formulation of the Stacked Pair Energy Model adds two more energy functions e and e', and two penalty parameters g and G. This necessitates the use of four additional energy tables $E_{S,l}, E_{S,r}, E_{R,l}, E_{R,r}$ to the set (E_S, E_R, E_l, E_r) already used in Section 4:



(a) (b)

Fig. 2. (a) Natural joint structure between small RNA molecules CopA (antisense) and CopT (its target) in *E.Coli.* (b) Natural joint structure between small RNA molecules fhlA (target[black]) and OxyS (antisense[red]) in *E.Coli*.

1. $E_{S,l}(S[i,j], R[i',j'])$ denotes the free energy between S and R such that $S[i]$ remains unbonded.
2. $E_{S,r}(S[i,j], R[i',j'])$ denotes the free energy between S and R such that $S[j]$ remains unbonded.
3. $E_{R,l}(S[i,j], R[i',j'])$ denotes the free energy between S and R such that $R[i']$ remains unbonded.
4. $E_{R,r}(S[i,j], R[i',j'])$ denotes the free energy between S and R such that $R[j']$ remains unbonded.

The addition of four more parameters (and four new degrees of freedom) makes this approach more adjustable to specifi c properties of the input RNA strands.

In addition to the stacked pair energies, this formulation also considers the free energies of an internally and externally bonded individual nucleotide pairs denoted $e(X[i], Y[j])$ and $e'(X[i], X[j])$ respectively. For further generality, this formulation induces an additive penalty for switching between the two types of bonds. More specifically, the energy function has an additive penalty g to any nucleotide $X[k]$ (X could be S or R), if (i) $X[k]$ is bonded with $X[j]$ however $X[k+1]$ is not bonded with $X[j-1]$, (ii) $X[k]$ is bonded with $X[j]$ however $X[k-1]$ is not bonded with $X[j+1]$, (iii) $X[k]$ is bonded with $Y[k']$ however $X[k+1]$ is not bonded with $Y[k'+1]$, (iv) $X[k]$ is bonded with $Y[k']$ however $X[k-1]$ is not bonded with $Y[k'-1]$. For unbonded nucleotides $X[k]$ another additive penalty G is charged if (i) $X[k+1]$ is bonded, (ii) $X[k-1]$ is

bonded. The gap penalties are also added to the first and last nucleotides of X - this is only for avoiding further complexity in the dynamic programming formulation and does not affect the energy minimization process or the resulting structure prediction.

Our new energy formulation is as follows.

$$E(S[i,j],R[i',j']) = \min \left\{ \begin{array}{l} E_S(S[i,j],R[i',j'])+2g, \quad E_R(S[i,j],R[i',j'])+2g, \\ E_r(S[i,j],R[i',j'])+2g, \quad E_l(S[i,j],R[i',j'])+2g, \\ E_{S,l}(S[i,j],R[i',j'])+G, \quad E_{S,r}(S[i,j],R[i',j'])+G, \\ E_{R,l}(S[i,j],R[i',j'])+G, \quad E_{R,r}(S[i,j],R[i',j'])+G, \\ \min_{i\le k\le j-1;i'\le k'\le j'-1} \left\{ \begin{array}{l} E(S[i,k],R[i',k'])+ \\ E(S[k+1,j]),R[k'+1,j']) \end{array} \right\} \\ \min_{i\le k\le j-1} \left\{ \begin{array}{l} E(S[i,k],-)+ \\ E(S[k+1,j]),R[i',j']) \end{array} \right\} \\ \min_{i\le k\le j-1} \left\{ \begin{array}{l} E(S[i,k],R[i',j'])+ \\ E(S[k+1,j]),-) \end{array} \right\} \\ \min_{i'\le k'\le j'-1} \left\{ \begin{array}{l} E(S[i,j],R[i',k'])+ \\ E(-,R[k'+1,j']) \end{array} \right\} \\ \min_{i'\le k'\le j'-1} \left\{ \begin{array}{l} E(-,R[i',k'])+ \\ E(S[i,j]),R[k'+1,j']) \end{array} \right\} \end{array} \right\}$$

These tables need to be initialized as follows.

$$\begin{array}{ll} E_l(S[i,j],-)=\infty & E_r(S[i,j],-)=\infty \\ E_l(-,R[i',j'])=\infty & E_r(-,R[i',j'])=\infty \\ E_l(S[i,i],R[i',i'])=e'(S[i],R[i'])+2g & E_r(S[i,i],R[i',i'])=e'(S[i],R[i'])+2g \end{array}$$

$$\begin{array}{ll} E_S(S[i,i],-)=\infty & E_R(-,R[i',i'])=\infty \\ E_{S,l}(S[i,i],-)=e(S[i],-)+G & E_{R,l}(-,R[i',i'])=e(-,R[i'])+G \\ E_{S,r}(S[i,i],-)=e(S[i],-)+G & E_{R,r}(-,R[i',i'])=e(-,R[i'])+G \end{array}$$

Here is the complete description of the dynamic programming formulation.

$$E_l(S[i,j],R[i',j']) = \begin{array}{l} e'(S[i],R[i'])+ \\ \min \left\{ \begin{array}{c} E_l(S[i+1,j],R[i'+1,j'])+ee'(S[i,i+1],R[i',i'+1]), \\ E(S[i+1,j],R[i'+1,j'])+2g \end{array} \right\} \end{array}$$

$$E_r(S[i,j],R[i',j']) = \begin{array}{l} e'(S[j],R[j'])+ \\ \min \left\{ \begin{array}{c} E_r(S[i,j-1],R[i',j'-1])+ee'(S[j-1,j],R[j'-1,j']), \\ E(S[i,j-1],R[i',j'-1])+2g \end{array} \right\} \end{array}$$

$$E_S(S[i,j],R[i',j']) = \begin{array}{l} e(S[i],S[j])+ \\ \min \left\{ \begin{array}{c} E_S(S[i+1,j-1],R[i',j'])+ee(S[i,i+1],S[j-1,j]), \\ E(S[i+1,j-1],R[i',j'])+2g \end{array} \right\} \end{array}$$

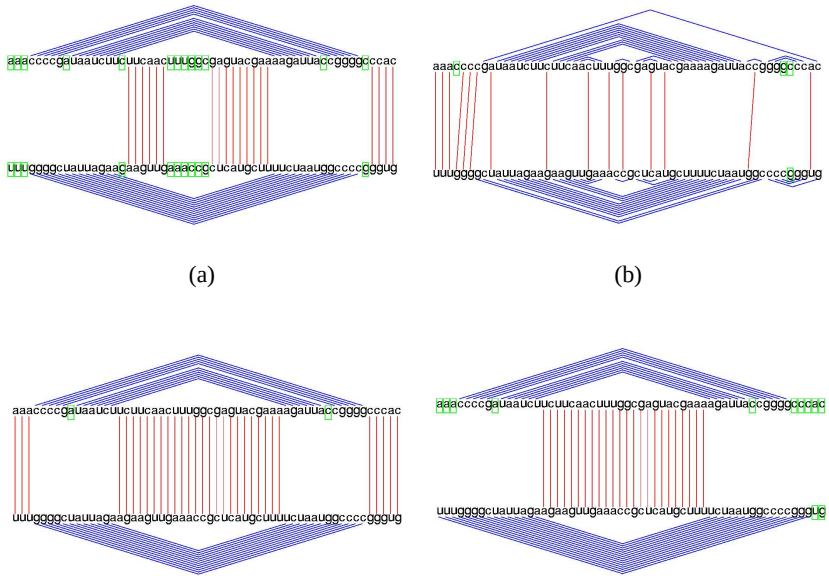
$$E_R(S[i,j],R[i',j']) = \begin{array}{l} e(R[i'],R[j'])+ \\ \min \left\{ \begin{array}{c} E_R(S[i,j],R[i'+1,j'-1])+ee(R[j'-1,j']^r,R[i',i'+1]^r), \\ E(S[i,j],R[i'+1,j'-1])+2g \end{array} \right\} \end{array}$$

$$E_{S,l}(S[i,j],R[i',j']) = \min \left\{ \begin{array}{l} E_{S,l}(S[i+1,j],R[i',j'])+e(S[i],-), \\ E(S[i+1,j],R[i',j'])+e(S[i],-)+G \end{array} \right\}$$

$$E_{S,r}(S[i,j],R[i',j']) = \min \left\{ \begin{array}{l} E_{S,r}(S[i,j-1],R[i',j'])+e(-,S[j]), \\ E(S[i,j-1],R[i',j'])+e(-,S[j])+G \end{array} \right\}$$

$$E_{R,l}(S[i,j],R[i',j']) = \min \left\{ \begin{array}{l} E_{R,l}(S[i,j],R[i'+1,j'])+e(R[i'],-), \\ E(S[i,j],R[i'+1,j'])+e(R[i'],-)+G \end{array} \right\}$$

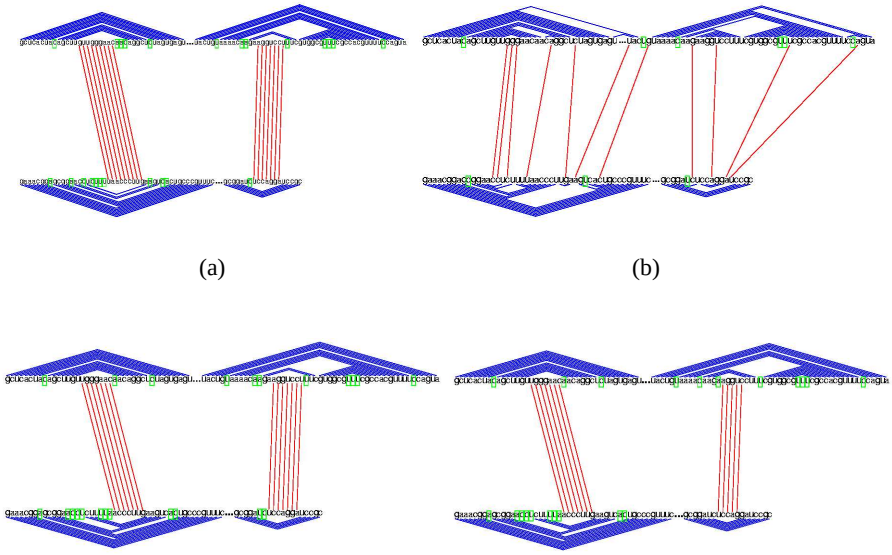
$$E_{R,r}(S[i,j],R[i',j']) = \min \left\{ \begin{array}{l} E_{S,r}(S[i,j],R[i',j'-1])+e(-,R[j']), \\ E(S[i,j],R[i',j'-1])+e(-,R[j'])+G \end{array} \right\}$$



(a) (b)

(c) (d)

Fig. 3. (a) Known joint structure of CopA and CopT, (b) as predicted by Basepair Energy Model, (c) as predicted by Stacked Pair Energy Model, (d) as predicted by Loop Energy Model,



(a) (b)

(c) (d)

Fig. 4. (a) Known joint structure of OxyS and fhlA, (b) as predicted by Basepair Energy Model, (c) as predicted by Stacked Pair Energy Model, (d) as predicted by Loop Energy Model.