A partition function algorithm for interacting nucleic acid strands

Hamidreza Chitsaz1,†, Raheleh Salari1,†, S. Cenk Sahinalp1,* and Rolf Backofen2,*

1Lab for Computational Biology, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada and 2Department of Bioinformatics, Institute of Computer Science, Albert-Ludwigs-Universität, Freiburg, Germany

ABSTRACT
Recent interests, such as RNA interference and antisense RNA regulation, strongly motivate the problem of predicting whether two nucleic acid strands interact.

Motivation: Regulatory non-coding RNAs (ncRNAs) such as microRNAs play an important role in gene regulation. Studies on both prokaryotic and eukaryotic cells show that such ncRNAs usually bind to their target mRNA to regulate the translation of corresponding genes. The specificity of these interactions depends on the stability of intermolecular and intramolecular base pairing. While methods like deep sequencing allow to discover an ever increasing set of ncRNAs, there are no high-throughput methods available to detect their associated targets. Hence, there is an increasing need for precise computational target prediction. In order to predict base-pairing probability of any two bases in interacting nucleic acids, it is necessary to compute the interaction partition function over the whole ensemble. The partition function is a scalar value from which various thermodynamic quantities can be derived. For example, the equilibrium concentration of each complex nucleic acid species and also the melting temperature of interacting nucleic acids can be calculated based on the partition function of the complex.

Results: We present a model for analyzing the thermodynamics of two interacting nucleic acid strands considering the most general type of interactions studied in the literature. We also present a corresponding dynamic programming algorithm that computes the partition function over (almost) all physically possible joint secondary structures formed by two interacting nucleic acids in \(O(n^6)\) time. We verify the predictive power of our algorithm by computing (i) the melting temperature for interacting RNA pairs studied in the literature and (ii) the equilibrium concentration for several variants of the OxyS–fhlA complex. In both experiments, our algorithm shows high accuracy and outperforms competitors.

Availability: Software and web server is available at http://compbio.cs.sfu.ca/taverna/pirna/

Contact: cenk@cs.sfu.ca; backofen@informatik.uni-freiburg.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Starting with the discovery of microRNAs (miRNAs) and the advent of genome-wide transcriptomics, it has become clear that RNA plays a large variety of important roles in living organisms that extend far beyond being a mere intermediate in protein biosynthesis (Storz, 2002). Several of these non-coding RNAs (ncRNAs) regulate gene expression post-transcriptionally through base pairing (and establishing a joint structure) with a target mRNA, as per the eukaryotic miRNAs and small interfering RNAs (siRNAs) (Bartel, 2004; Hannon, 2002; Zamore and Haley, 2005), antisense RNAs (Brantl, 2002; Wagner and Fldrhd, 2002) or bacterial small regulatory RNAs (sRNAs) (Gottesman, 2005). In addition to such endogenous regulatory ncRNAs, antisense oligonucleotides have been used as exogenous inhibitors of gene expression; antisense technology is now commonly used as a research tool as well as for therapeutic purposes. Furthermore, synthetic nucleic acids systems have been engineered to self-assemble into complex structures performing various dynamic mechanical motions (Seeman, 2005; Seeman and Lukeman, 2005; Simmel and Dittmer, 2005; Venkataraman et al., 2007; Yin et al., 2008).

Despite all the above advances, computational methods for predicting ncRNA–target mRNA interactions suffer from a low specificity (see below for an overview); this is possibly due to two technical reasons. First, several of these methods consider restricted versions of the problem (e.g. simplified energy functions or restricted types of interactions)—this is mostly for computational reasons. Second, a quantitative analysis of binding thermodynamics between oligonucleotides and target RNAs is often lacking. To determine the binding effectiveness, an accurate analysis of the thermodynamics of two interacting nucleic acid strands is necessary.

In this article, we aim to compute how likely two RNA or DNA strands are to interact, and to predict base-pairing probability of any two bases—which we then use to quantitatively measure the strength, probability and stability of the joint structure established by the interacting strands (Mathews, 2004). To correctly calculate those probabilities, it is necessary to compute the partition function over the whole ensemble of possible individual and joint secondary structures. The partition function is a scalar value from which various thermodynamic quantities can be derived (Landau and Lifshitz, 1969). For instance, one can compute the equilibrium concentration of each complex nucleic acid species from their partition functions. Also, the partition function can be used to predict the melting temperature of interacting nucleic acids.

Although algorithms for predicting the most likely (the lowest total free energy) joint structure that can be formed by two interacting RNA strands are available [see, for example, Alkan et al. (2006)], previous little work has been done for computing the partition function of interacting RNA strands. It is important to note that designing an algorithm to compute the partition function is more challenging than that to predict the minimum free energy...
secondary structure: a partition function algorithm should guarantee that every joint structure is considered exactly once.

In this article, we present an \(O(n^3)\) time algorithm to compute the partition function over the type of interactions that Alkan et al. (2006) considered.\(^1\) We extend the standard energy model for a single RNA model to an energy model for the joint secondary structure of interacting strands by considering new types of (joint) structural components. We verify our algorithm (and the associated software we developed) by computing the melting temperature for RNA pairs available (Diamond et al., 2001; Mathews and Turner, 2002; Xia et al., 1998) and the equilibrium concentration for OxyS–IHa complexes for wild-type IHa and four other mutants reported in the literature (Argaman and Altuvia, 2000). In both experiments, our algorithm shows high accuracy and outperforms reported in the literature (Argaman and Altuvia, 2000). In both approaches, our algorithm shows high accuracy and outperforms reported in the literature (Argaman and Altuvia, 2000).

## 1 METHODS

### 2.1 Preliminaries

Throughout this article, we denote the two nucleic acid strands by \(R\) and \(S\). Strand \(R\) is indexed from 1 to \(L_R\), and \(S\) is indexed from 1 to \(L_S\) both in 5’ to 3’ direction. Note that the two strands interact in opposite directions, e.g., \(R\) in 5’ to 3’ with \(S\) in 3’ to 5’ direction. Each nucleotide is paired at most one nucleotide in the same or the other strand. We refer to the \(i\)-th nucleotide in \(R\) and \(S\) by \(i_R\) and \(i_S\), respectively. The subsequence from the \(i\)-th nucleotide to the \(j\)-th nucleotide in a strand is denoted by \([i,j]\).

An intramolecular base pair between the nucleotides \(i\) and \(j\) in a strand is called an arc and denoted by a bullet (\(\bullet\)). An intramolecular base pair between the nucleotides \(i_R\) and \(i_S\) is called a bond and denoted by a circle (\(\bigcirc\)). An arc \(i_R\bullet j_R\) covers a bond \(i_R\bigcirc j_R\) if \(i_R = i_S < j_R\). We call \(i_R\bullet j_R\) an interaction arc if there is a bond \(i_R\bigcirc j_R\) covered by \(i_R\bullet j_R\). A killing arc is an interaction arc that directly covers a bond. More precisely, we call \(i_R\bullet j_R\) a killing arc if it covers a bond \(i_R\bigcirc j_R\) such that if \(i_R'\bullet j_R'\) covers the same bond \(i_R\bigcirc j_R\), then \(i_R' < i_R\) and \(i_R' < j_R\). A subsequence \([i_R,j_R]\) contains a direct bond \(i_R\bigcirc j_R\) if \(i_R < j_R\). A subsequence \([i_R,j_R]\) covers a direct bond \(i_R\bigcirc j_R\) if \(i_R < j_R\leq j_S\) and no arc within \([i_R,j_R]\) covers \(i_R\bigcirc j_R\). Assuming \(i_R < j_R\), two bonds \(i_R\bigcirc j_R\) and \(i_S\bigcirc j_S\) are called crossing bonds if \(i_R < j_S\). An interaction arc \(i_R\bullet j_R\) in a strand \(s\) subsumes a subsequence \([i_S,j_S]\) in the other strand if for all bonds \(i_R\bigcirc j_R\) and \(i_S\bigcirc j_S\) in the subsequence \([i_S,j_S]\) it subsumes \([i_R,j_R]\).

In this article, we assume there are no pseudoknots in individual secondary structures of \(R\) and \(S\), and also there are no crossing bonds and zigzags between \(R\) and \(S\).

### 2.2 Interaction energy model

An unpseudoknotted secondary structure \(s\) of a single nucleic acid, in the standard energy model (Mathews et al., 1999), is decomposed into loops,
These components and their free energy contributions are the standard case, an interaction secondary structure model by defining those new kinds of interaction components. Similar to work, new kinds of components can appear. We extend the standard energy (remember we do not allow pseudoknots, crossing bonds and zigzags in this total free energy structure into the interaction components (Argaman and Altuvia, 2000).

Interaction components of OxyS–fhlA pair presented in Argaman Fig. 1.

A hybrid component between the two strands whose free energy is $G^\text{hybrid} = p_1 + \sigma (G^\text{pivot} + G^\text{pivot} - G^\text{intra}) = G^\text{hybrid}$ and a free energy is associated with every loop in $s$. The total free energy $G_s$ is the sum of loop free energies. The standard model consists of the following loop types: (i) hairpin, (ii) interior, which can be stack, bulge or internal loop and (iii) multiloop whose energies are denoted by $G^\text{intra}$, $G^\text{pivot}$ and $G^\text{hybrid}$, respectively.

In an interaction, secondary structure of two strands under our assumptions (remember we do not allow pseudoknots, crossing bonds and zigzags in this work), new kinds of components can appear. We extend the standard energy model by defining those new kinds of interaction components. Similar to the standard case, an interaction secondary structure $s$ can be decomposed into intramolecular loops and the new interaction components such that the total free energy $G_s$ is sum of the free energies of loops and interaction components. Figure 1 shows the decomposition of OxyS–fhlA pair secondary structure into the interaction components (Argaman and Altuvia, 2000). These components and their free energy contributions are:

- **Hybrid**: $G^\text{hybrid}(G_{k1}^{b1}\ldots G_{kn}^{bl})$ is the free energy of a joint secondary structure consisting of a series of bonds, $G_{k1}^{b1}\ldots G_{kn}^{bl}$, with no intramolecular base pairing or branching. We call such a component hybrid. We define the energy associated with a hybrid component by

  \[
  G^\text{hybrid}(G_{k1}^{b1}\ldots G_{kn}^{bl}) = p_1 + \sigma \sum_{i=1}^{n} G^\text{intra}(G_{k1}^{b1}\ldots G_{ki}^{bi}\ldots G_{kn}^{bl}),
  \]

  where $p_1$ is the penalty for the formation of the hybrid, and $\sigma \leq 1$ is the ratio of the free energy of intramolecular to that of intramolecular interior loops (as suggested by Alkan et al. (2006)) (Fig. 2). Note that with $p_1 = 0, \sigma = 1$, $G^\text{hybrid}$ is identical to the energy proposed by Böhm et al. (2007), first introduced by Rehmsmeier et al. (2004), which considers only one hybrid component for mRNA/target duplexes and does not allow any intramolecular structure.

- **Kissing**: $G^\text{kissing}(G_{k1}^{b1}\ldots G_{kn}^{bl})$ is the energy of an intramolecular loop (hairpin, interior or multiloop) that makes interaction with the other strand. Such a hybrid component is called a kissing loop. The energy associated with a kissing loop is given by

  \[
  G^\text{kissing}(G_{k1}^{b1}\ldots G_{kn}^{bl}) = G^\text{stem} + G^\text{bulge} + G^\text{intra}(G_{k1}^{b1}\ldots G_{kn}^{bl}),
  \]

  in which $R^\text{b}$ is the number of loops and $R^\text{u}$ the number of unpaired bases in the kissing loop (Fig. 3). Note that in our model we use different $p_1$ and $\sigma$ values for a hybrid component covered by a kissing loop.

- **Inter-hybrid**: $G^\text{inter-hybrid}$ is the energy of an intermolecular loop bounded by two bonds belonging to two consecutive hybrid components. Bases in either sequence facing this kind of loop might be the end points of only arcs and not bonds. We call such a component inter-hybrid loop. In this work, the energy contribution of an inter-hybrid loop is assumed to be 0.

2.3 Interaction partition function

The partition function is a weighted sum over the set of all possible secondary structures $\mathcal{S}$

\[
Q(T) = \sum_{\mathcal{S}} e^{-\beta G_{\mathcal{S}}/RT}
\]

where $R$ is the universal gas constant and $T$ is the temperature. Efficient algorithms for computing the partition function for a single strand have been given. McCaskill (1990) gave the first partition function algorithm for a single unpseudoknotted nucleic acid strand and Dirks and Pierce (2003) gave a partition function algorithm for a single strand allowing pseudoknots. However, computing the partition function for multiple interacting strands has not been properly addressed. In previous attempts, multiple strands are concatenated in some order and partition function for the resulting single strand is computed (Bernhart et al., 2006; Dimitrov and Zuker, 2004; Dirks et al., 2007). That approach is not accurate as even if pseudoknots are considered, some useful interactions are excluded while many physically impossible interactions are included (e.g. physically impossible crossing interactions). On the other hand, considering all possible secondary structures makes the problem NP-hard (Alkan et al., 2006). Therefore, we only consider those secondary structures that do not contain pseudoknots, crossing bonds and zigzags.

**Interaction partition function (IPF) problem**

Given a pair of nucleic acid strands $R$ and $S$, and a temperature $T$, compute the partition function, $Q^\dagger(T)$, over $\mathcal{S}^\dagger$ the set of all possible single or duplex secondary structures that do not contain pseudoknots, crossing bonds and zigzags.

**Input**: nucleic acid strands $R$ and $S$.

**Output**: $Q^\dagger(T) = \sum_{\mathcal{S}} e^{-\beta G_{\mathcal{S}}/RT}$

We give a recursive algorithm, called Partition function for InteRacting Nucleic Acids (piRNA), for the IPF problem. In all of our recursions, the considered cases are disjoint. This fact shows that every possible secondary structure is reached by exactly one trajectory in the recursion process. Our algorithm guarantees to consider all possible secondary structures exactly...
in a multiloop and the penalty of multiloop should be applied to it. Explicit applied.

cases as the corresponding white regions, except that for the green regions defined by the energy model. Green and red regions have the same recursion region specify a recursive quantity. White regions are recursed over and blue its two terminal bases that may be paired or unpaired. Letter(s) within a line indicates an arc and a dashed curved line encloses a region and denotes

and (ii) the joint partition function of subsequences \([i_k,j_k]\) and \([i_l,j_l]\). A region is the domain over which a partition function is computed. Terminal bases are the boundaries of a region. For the first type, region is \([i,j]\) with i and j terminal bases. For the second type, region is \([i_k,j_k]\times[i_l,j_l]\) with \(i_k,j_k\) and \(i_l,j_l\) terminal bases. The length pair of region \([i_k,j_k]\times[i_l,j_l]\) is \(l_k=(p_i-p_j+1,l_p-j_l+1)\). Our algorithm starts with \(l_p=1,l_r=1\) and considers all length pairs incrementally up to \((l_k=L_R,l_r=L_L)\). For a fixed length pair \((l_k,l_r)\), recursive quantities for all the regions \([i_k,i_k+1],[i_k,i_k+2]\) are computed.

For computing the partition function of a region in one strand we use McCaskill’s (1990) algorithm. McCaskill’s algorithm is shown in Figure 4, in which \(Q_{ij}\) is the partition function for the subsequence \([i,j]\). Throughout this article, a horizontal line indicates the phosphate backbone, a solid curved line indicates an arc and a dashed curved line encloses a region and denotes its two terminal bases that may be paired or unpaired. Letter(s) within a region specify a recursive quantity. White regions are recursed over and blue regions indicate those portions of the secondary structure that are fixed at the current recursion level and contribute their energy to the partition function as defined by the energy model. Green and red regions have the same recursion cases as the corresponding white regions, except that for the green regions multiloop energy and for red regions kissing loop energy is applied. i.e. the corresponding penalties for each unpaired base and base pair should be applied.

In Figure 4, the first case of \(Q_{ij}\) corresponds to an empty structure (that constitutes no base pairs) whose free energy is assumed to be 0, thus its contribution to the partition function is \(e^{-\gamma \gamma_{corr}}\), \(\gamma_{corr}=1\). In the other case, there exists at least one arc and the leftmost one is \(k_1\). It contributes \(Q_{i_1,j_1}Q_{i_1,j_1}\) to the partition function, therefore,

\[
Q_{ij}=1+\sum_{(i_k-1,j_k-1)} Q_{i_k,j_k}Q_{i_k,j_k+1}.
\]

The second line shows the cases of \(Q_{ij}\) which is the partition function for the subsequence \([i,j]\) assuming i and j are base paired. The arc \(k_1\) can close different substructures: hairpin, interior or multiloop. The energy contribution of each substructure is calculated based on the standard thermodynamics energy model. The third line shows cases of \(Q_{ij}\) which is the partition function for the subsequence \([i,j]\) assuming the region constitutes at least one arc. A region tagged by \(k_2\) and colored by green is contained in a multiloop and the penalty of multiloop should be applied to it. Explicit equations for \(Q_{ij}\) and \(Q_{ij}\) are given in the Supplementary Material.

In the following, we present all cases of \(Q_{ij}\) which is the interaction partition function for the region \([i_k,j_k]\) \times \([i_l,j_l]\). A solid vertical line indicates a bond, a dashed vertical line denotes two terminal bases of a region which may be base paired or unpaired and a dotted vertical line denotes two terminal bases of a region which are assumed to be unpaired. For the interaction partition functions, gray regions indicate a reference to the partition functions for the single sequences. Figure 5 shows the cases of \(Q_{ij}\): (i) there is no bond between the two subsequences, (ii) the leftmost bond is a direct bond in both subsequences and (iii) the leftmost bond is covered by an arc in at least one subsequence. Therefore,

\[
Q_{i_k,j_k,i_l,j_l} = Q_{i_k,j_k} Q_{i_l,j_l}
\]

\[
+ \sum_{(k_1 \neq k_2) \times [i_k,j_k]} Q_{i_k,j_k-1} Q_{i_k,j_k+1} Q_{i_l,j_l} Q_{i_l,j_l-1} Q_{i_l,j_l+1} + \sum_{(k_1 \neq k_2) \times [i_k,j_k]} Q_{i_k,j_k-1} Q_{i_k,j_k+1} Q_{i_l,j_l} Q_{i_l,j_l-1} Q_{i_l,j_l+1},
\]

\[
(5)
\]

in which \(Q_{i_k,j_k,i_l,j_l}\) is the interaction partition function for the region \([k_1,k_2]\times[i_l,j_l]\) assuming \(k_1,k_2\) is a bond, and \(Q_{i_k,j_k,\text{empty}}\) is the interaction partition function for the region \([k_1,k_2]\times[i_l,j_l]\) assuming that the leftmost bond in the region is covered by an arc in at least one subsequence. Figures 6 and 7 show the recursion for \(Q_{ij}\) and \(Q_{ij}\) where b stands for bond and a stands for arc.
two hybrid components. The subsequent intermolecular bond starts a new hybrid component. In the other cases, we have always at least one hybrid component. In order to determine whether the region contains one or several hybrid components, we use the symmetric counterpart of $QIh$ defined in (6) and (7). Here, we report our implementation of the algorithm and two types of experiments.

Figure 6 shows the recursion for $QIh_{d,j}$, the interaction partition function for the region $[i_j, j_r]$ assuming $i_j = j_r$ is a bond. Since we have a $\beta$ penalty for each hybrid component, the recursion for $QIh$ has to determine whether the region contains one or several hybrid components. In all cases, $QIh$ contains the full hybrid component containing the bond $i_j = j_r$ (see Fig. 7 for $QIh$ recursion). The first possibility reflects the case where we have only one hybrid component. In the other cases, we have always at least two hybrid components. The subsequent intermolecular bond starts a new hybrid component iff (i) it is not directed in at least one subsequence, i.e. it is covered by an arc in the associated regions (Case 2 of the $QIh$ recursion), or (ii) there is at least one arc between the two subsequent intermolecular bonds (Cases 3 and 4 of the $QIh$ recursion). Using the above matrices $QIhm$ and $QIh$, we get

$$ QIh_{d,j} = QIh_{d,j} + \sum_{i < k < j} QIh_{d,i} QIh_{k,j} $$

$$ + \sum_{i < k < j} QIh_{i} QIh_{k,j} $$

Figure 7 shows the cases of $QIh$: (i) there is no bond other than $i_j = j_r$ and $i_r = j_r$ in the region, and (ii) there exist more bonds between $i_r = j_r$ and $i_r = j_r$, the leftmost of which is $k_1 = k_2$.

Figure 8 shows the cases of $QIh_{d,j}$, for which at least one of $i_r$ and $j_r$ is the end point of an interaction arc: (i) $i_r = j_r$, subsumes $k_2$, $j_r$ and $k_1$ is not base paired with $j_r$, (ii) $k_1 = k_2$ subsumes $i_r = j_r$ and $j_r$ is not base paired with $k_1$, and (iii) $i_r = j_r$, and $k_1$ is base paired with $k_2$. If just one of $i_r$ and $j_r$ is the end point of an interaction arc while the other one is the end point of a bond, then the interaction arc subsumes the other subsequence. If both $i_r$ and $j_r$ are end points of interaction arcs, then one of the arcs subsumes the other one or they are equivalent. Therefore

$$ QIh_{d,j} = QIh_{d,j} + \sum_{k_1 = k_2} QIh_{d,i_r} QIh_{k_1,j_r} $$

$$ + \sum_{i_r < j_r < k_1} QIh_{d,i_r} QIh_{k_1,j_r} $$

$$ + \sum_{i_r < j_r < k_1} QIh_{d,i_r} QIh_{k_1,j_r} $$

in which $QIh_{d,i_r,j_r}$ is the interaction partition function of $[i_r,k_1] \times [k_1,j_r]$ assuming $i_r = j_r = k_1$ is an interaction arc that subsumes $[k_1,j_r]$. $QIh_{d,i_r,j_r}$ is the symmetric counterpart of $QIh$ and $QIh_{d,i_r,j_r}$ is the interaction partition function of $[i_r,k_1] \times [k_1,j_r]$, for which $i_r \bullet j_r$ and $i_r \bullet j_r$ are equivalent interaction arcs.

Figure 9 shows the cases of $QIh_{d,j}$, for which we assume at least one of $i_r$ and $j_r$ is the end point of an interaction arc.

For $QIh$, it does not make any difference which one of the covering arcs, $i_r \bullet j_r$ and $i_r \bullet j_r$, is extracted first. We first extract the covering arc from $S$ (Fig. 9). Extracting the covering arc, the remaining subsequence of $S$ contains either at least one direct bond, in which case kissing loop penalty should be applied, or multiple interaction arcs, in which case multiloop penalty should be applied. Hence, Figure 9 is appropriately colored by green and red to remind the type of penalty.

Note that $QIh_{d,i_r,j_r}$ and $QIh_{d,j_r}$ are the partition functions for $[i,j]$ excluding the gap $[d,e]$ assuming $i$ and $j$ are base paired. For $QIh$, the gap is assumed to contain a direct bond, and for $QIh$ the gap is assumed to contain multiple interaction arcs. In both cases, $i < k_1 \leq d$ and $e < k_2 < j$.

The union of the cases of $QIh$ and $QIh$ comprises the cases of $QIh$. Similar to the cases of $QIh$, we extract the covering arc from $QIh$ and $QIh$ to obtain $QIh^{\text{mm}}$, $QIh^{\text{mm}}$, $QIh^{\text{mm}}$, and $QIh^{\text{mm}}$, where $k$ stands for kissing (or equivalently containing a direct bond) and $m$ for multiple interaction arcs. Note that all four terminal bases of the region of these four quantities are paired, i.e. each terminal base is either the end point of a bond or of an interaction arc. These four quantities have complicated cases. Due to lack of space, we explain them in our Supplementary Material.

## 3 EXPERIMENTS

Here, we report our implementation of the algorithm and two types of experiments we performed to test the predictive power of our algorithm.

1. Predicting the melting temperature of RNA duplexes is an important application of the partition function for interacting nucleic acid pairs (Dimitrov and Zuker, 2004): our first
experiment tests how accurately our algorithm predicts the melting temperature of RNA pairs collected from several sources in the literature with respect to the accuracy of available alternatives. RNAcofold from Vienna package v1.7.2 (Bernhart et al., 2006) and UNAfold v3.6 which is a new version of former mfold (Markham and Zuker, 2008). We remind the reader that RNAcofold concatenates the two RNA strands and computes the partition function for the resulting single strand. Therefore, it does not consider many cases that our algorithm considers. UNAfold v3.6, on the other hand, simplifies the problem by forbidding intramolecular base pairing. It computes the partition function of the two strands over just hybridization structures. As can be expected, our algorithm consistently outperforms the alternatives in all three datasets.

(2) A novel experiment (which, to our knowledge has not been performed successfully by any other program to date) uses our algorithm to predict the equilibrium concentration of an RNA-RNA complex, in particular the OxyS-fhA interaction (Argaman and Altuvia, 2000). We successfully predicted the equilibrium concentrations for OxyS with wild-type fhA and four other fhA mutants.

Note that the parameters used by our program in the above experiments have been manually optimized as computational learning methods for fine tuning the parameters require prohibitive computational resources. It may be possible to improve the accuracy of our program through a better selection of parameters.

3.1 Implementation

We remind the reader that the time and space complexity of our algorithm are $O(n^3)$ and $O(n^4)$, respectively, here $n = \max(L_L, L_L)$ is the maximum length of the two input strands. We implemented our algorithm in C++, and used the energy functions and energy parameters of UNAfold v3.6 for a single strand (Markham and Zuker, 2008). The parameters used by our program for our own interaction energy model are given in the next section. As per Section 2.2, we use a different $\beta_1$ penalty and $\sigma$ for a hybrid component that is covered by a kissing loop. The parameters for a hybrid component that is not covered by a kissing loop is denoted by $\beta'_1$ and $\sigma'$. We add an Adenine-Uracil base pair (AU) penalty to the energy of a hybrid component per each terminal AU base pair; this penalty is motivated by Xia et al. (1998). Similar to RNAhybrid, the interior loops in a hybrid component are restricted to a constant maximum length, in either sequence, which is set to 15 in this work.

Since our algorithm considers many more possible secondary structures in comparison to alternative methods, our program has a higher running time. Fortunately, our algorithm can be easily parallelized as the dynamic programming tables computed by our program on subsequence pairs depend only on their (proper) subregions. We parallelized our program using OpenMP 3.0. Our experiments were performed on a large-scale shared memory parallel platform with 64 PPC 1.9 GHz processors with 256 GB RAM. We ran our program for strands of length between 5 nt to 120 nt. The running time of our program for short strands (~20 nt) was < 1 m—for longer strands (~120 nt) it was ~10 h.

3.2 Datasets

The first dataset that we used for predicting melting temperature contains all nine different RNA pairs reported in Table 3 of Xia et al. (1998). It contains almost complementary 5- to 7- nt RNA pairs that were designed to optimize the thermodynamic parameters for terminal base pairs. Their melting temperatures vary from 29.8°C to 53.7°C.

The second dataset that we used for computing melting temperature contains all 12 different RNA pairs reported in Table 1 of Diamond et al. (2001). These RNA pairs are designed to optimize the thermodynamic parameters for three-way multi loops. In each pair of this dataset, the first RNA has ~20 nt and the second one has ~10 nt. The experimental melting temperatures were determined from heat absorbance measurements by two different methods that are explained as ‘Method 3’ and ‘Method 4’ in Puglisi and Tinoco (1989). Although these pairs are very similar, the average difference of the two methods for this dataset is 2.49°C. This suggests that there may exist RNA pairs with exceptional features in this set.

The third dataset that we used for computing melting temperature contains all 62 different RNA pairs reported in Tables 3 and 4 of Mathews and Turner (2002). These pairs are designed to optimize the thermodynamic parameters for three- and four-way multi loops. In each pair of this dataset, the first RNA has 22–40 nt and the second one has 10–14 nt. Again, the experimental melting temperatures were determined by two different methods. This dataset is large enough with longer sequences, and the average difference of the two methods for this dataset is 0.7°C smaller than that for the second dataset. Moreover, the variance and maximum of the difference is smaller than those of the second dataset. Overall, this dataset is more reliable than the previous one. These three datasets are all we were able to collect from the literature.

3.3 Melting temperature

As mentioned before, the melting temperature of RNA duplexes is one of the most important applications of the partition function for interacting nucleic acid pairs (Dimitrov and Zuker, 2004). Table 1 shows the melting temperatures computed by our program, RNAcofold, and UNAfold v3.6 for the first dataset. In this set, the strands are short, and as we expected, our algorithm is highly accurate with only 1.48°C absolute difference from experimental values on average. It can be seen that RNAcofold and UNAfold perform relatively poorly, and their predicted melting temperatures differ from the experimental values by about 9°C on average.

Table 2 shows the melting temperatures predicted by the three programs for the second dataset. Each pair is referred to by an identifier (A, B, ..., L). Please refer to our Supplementary Material or Diamond et al. (2001) to see the exact sequences of each pair. As mentioned before, the experimental melting temperatures were determined from heat absorbance measurements by two different
### Table 1. Experimental and predicted melting temperatures for the first dataset [see Section 3.2 and Xia et al. (1998)]

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Experiment</th>
<th>pRNA</th>
<th>RNAcofold</th>
<th>UNAFold</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOCA/ACCOU</td>
<td>29.8</td>
<td>29.41</td>
<td>42.64</td>
<td>46.14</td>
</tr>
<tr>
<td>GCA/CAGAGCGC</td>
<td>37.45</td>
<td>36.07</td>
<td>46.61</td>
<td>43.91</td>
</tr>
<tr>
<td>AGGAC/TGCAGC</td>
<td>30.2</td>
<td>30.38</td>
<td>42.68</td>
<td>45.15</td>
</tr>
<tr>
<td>GC/CUGCGACGC</td>
<td>37.2</td>
<td>36.88</td>
<td>47.75</td>
<td>44.71</td>
</tr>
<tr>
<td>AGCUC/GACAGAGU</td>
<td>48.2</td>
<td>44.91</td>
<td>56.8</td>
<td>57.59</td>
</tr>
<tr>
<td>GUCUC/GAGUGAGC</td>
<td>51.1</td>
<td>49.4</td>
<td>58.44</td>
<td>55.91</td>
</tr>
<tr>
<td>AGCAD/GACAGGAC</td>
<td>45.7</td>
<td>45.47</td>
<td>56.4</td>
<td>56.68</td>
</tr>
<tr>
<td>GACUC/GACAGAGC</td>
<td>52.4</td>
<td>49.96</td>
<td>59.11</td>
<td>56.25</td>
</tr>
<tr>
<td>GAC/GAGCA/GACUC</td>
<td>53.7</td>
<td>49.07</td>
<td>58.07</td>
<td>56.00</td>
</tr>
</tbody>
</table>

**Avg. error**

1.48 0.97 0.97 0.57

**Spearman rank corr.**

1.48 0.35 8.55 0.97

All values, except Spearman’s rank correlation, are in degree centigrade. Bold numbers are the most accurate predictions. In other words, they have the least difference from experimental measurements.

### Table 2. Experimental and predicted melting temperatures for the set of RNA pairs reported in Diamond et al. (2001)

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Experiment</th>
<th>pRNA</th>
<th>RNAcofold</th>
<th>UNAFold</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.7</td>
<td>30.18</td>
<td>31.5</td>
<td>33.22</td>
</tr>
<tr>
<td>B</td>
<td>19.0</td>
<td>20.51</td>
<td>31.5</td>
<td>32.22</td>
</tr>
<tr>
<td>C</td>
<td>33.6</td>
<td>33.6</td>
<td>32.94</td>
<td>39.77</td>
</tr>
<tr>
<td>D</td>
<td>33.9</td>
<td>39.6</td>
<td>32.43</td>
<td>26.85</td>
</tr>
<tr>
<td>E</td>
<td>23.2</td>
<td>24.4</td>
<td>31.66</td>
<td>32.22</td>
</tr>
<tr>
<td>F</td>
<td>34.9</td>
<td>36.9</td>
<td>33.28</td>
<td>39.91</td>
</tr>
<tr>
<td>G</td>
<td>32.4</td>
<td>35.6</td>
<td>32.76</td>
<td>44.27</td>
</tr>
<tr>
<td>H</td>
<td>16.1</td>
<td>18.9</td>
<td>36.41</td>
<td>29.76</td>
</tr>
<tr>
<td>J</td>
<td>19.0</td>
<td>23.2</td>
<td>32.32</td>
<td>29.18</td>
</tr>
<tr>
<td>K</td>
<td>32.3</td>
<td>37.1</td>
<td>37.01</td>
<td>28.8</td>
</tr>
<tr>
<td>L</td>
<td>23.4</td>
<td>30.7</td>
<td>31.45</td>
<td>49.36</td>
</tr>
<tr>
<td>Pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Avg. difference</strong></td>
<td>0.87</td>
<td>0.36</td>
<td>0.45</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.16</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

| Spearman rank corr.    | 0.07       | 1.07     | 1.95      | 14.27   |
|                        | 14.41      | 26.3     | 26.56     |         |

Each pair is referred to by an identifier (A, B, . . . , L). Please refer to our Supplementary Material or Diamond et al. (2001) to see the exact sequences of each pair. All values, except Spearman’s rank correlation, are in degree centigrade. Bold entries are the most accurate predictions. In other words, they have the least difference from experimental measurements.

### Table 3. Experimental and predicted melting temperatures for the set of RNA pairs reported in Mathews and Turner (2002)

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Experiment</th>
<th>pRNA</th>
<th>RNAcofold</th>
<th>UNAFold</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45.4</td>
<td>46.1</td>
<td>56.81</td>
<td>37</td>
</tr>
<tr>
<td>B</td>
<td>51.8</td>
<td>52.2</td>
<td>56.84</td>
<td>37</td>
</tr>
<tr>
<td>C</td>
<td>55.6</td>
<td>56.8</td>
<td>56.85</td>
<td>37</td>
</tr>
<tr>
<td>D</td>
<td>58.4</td>
<td>57.3</td>
<td>56.85</td>
<td>37</td>
</tr>
<tr>
<td>E</td>
<td>57.5</td>
<td>56.9</td>
<td>56.84</td>
<td>37</td>
</tr>
<tr>
<td>F</td>
<td>56.7</td>
<td>57.1</td>
<td>56.85</td>
<td>37</td>
</tr>
<tr>
<td>G</td>
<td>51.2</td>
<td>49.3</td>
<td>56.94</td>
<td>37</td>
</tr>
<tr>
<td>H</td>
<td>55.2</td>
<td>54.7</td>
<td>56.96</td>
<td>41</td>
</tr>
<tr>
<td>I</td>
<td>56.1</td>
<td>55.6</td>
<td>56.98</td>
<td>41</td>
</tr>
<tr>
<td>J</td>
<td>54.8</td>
<td>54.3</td>
<td>56.98</td>
<td>37</td>
</tr>
<tr>
<td>K</td>
<td>55.3</td>
<td>54.8</td>
<td>56.98</td>
<td>37</td>
</tr>
<tr>
<td>L</td>
<td>55.1</td>
<td>52.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>M</td>
<td>56.3</td>
<td>57.5</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>N</td>
<td>57.2</td>
<td>56.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>O</td>
<td>57.9</td>
<td>57.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>P</td>
<td>56.5</td>
<td>57.5</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>Q</td>
<td>57.2</td>
<td>56.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>R</td>
<td>57.9</td>
<td>57.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>S</td>
<td>56.5</td>
<td>57.5</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>T</td>
<td>57.2</td>
<td>56.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>U</td>
<td>57.9</td>
<td>57.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>V</td>
<td>56.5</td>
<td>57.5</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>W</td>
<td>57.2</td>
<td>56.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>X</td>
<td>57.9</td>
<td>57.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>Y</td>
<td>56.5</td>
<td>57.5</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>Z</td>
<td>57.2</td>
<td>56.9</td>
<td>57.18</td>
<td>41.64</td>
</tr>
<tr>
<td>Pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Avg. difference</strong></td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
</tbody>
</table>

| Spearman rank corr.    | 0.02       | 0.02     | 0.02      | 0.02    |
|                        | 0.02       | 0.02     | 0.02      | 0.02    |

Each pair is referred to by an identifier (A, B, . . . , L). Please refer to our Supplementary Material or Mathews and Turner (2002) to see the exact sequences of each pair. All values, except Spearman’s rank correlation, are in degree centigrade. Bold entries are the most accurate predictions. In other words, they have the least difference from experimental measurements.
Altuvia (2000). Those plots can be predicted from the partition functions. To validate our algorithm, we computed these partition functions for OxyS, fhlA, OxyS–OxyS, fhlA–fhlA and OxyS-fhlA. Their plots are reported in Figure 8 and Table 2 of Argaman and Altuvia (2000). Both RNAcofold and UNAFold predict that the percentage of OxyS in complex is approximately 0 for the considered fhlA concentrations.

The running time of our program for the first dataset was about a few seconds, for the second dataset about 10 min and for the third dataset ~72 h on a Linux PC with Pentium-D 3.6 GHz CPU and 4 GB of RAM. Note that we did not use any learning methods for tuning our six interaction energy parameters because of the running time of our program. Our interaction energy parameters in melting experiments are $\beta_1 = 5.1, \beta_2 = 0.1, \sigma = 0.92, \beta'_1 = 4.1$ and $\sigma' = 0.95$, which were manually optimized using only the first data set. The second and the third datasets were used as test sets.

### 3.4 Equilibrium concentration

Our second set of experiments, to the best of our knowledge, have not been successfully performed by the use of any available program to date. Here we predict the equilibrium concentrations for OxyS with wild-type fhlA and four other fhlA mutants. OxyS is a small untranslated RNA (109 nt) that is induced in response to oxidative stress in *Escherichia coli*. It acts as a regulator affecting the expression of multiple genes. In particular, OxyS represses the translation of fhlA, a transcriptional activator for formate metabolism, by binding to it. Argaman and Altuvia (2000) carried out a series of experiments to measure equilibrium dissociation constants for OxyS with wild-type fhlA and its mutants. To measure the equilibrium dissociation constants, they measured the concentration of OxyS–fhlA complex for a fixed initial OxyS concentration (2 nM) and various initial concentrations of fhlA. Their plots are reported in Figure 8 and Table 2 of Argaman and Altuvia (2000). Those plots can be predicted from the partition functions for OxyS, fhlA, OxyS–OxyS, fhlA–fhlA and OxyS-fhlA. To validate our algorithm, we computed these partition functions using our program, and predicted the equilibrium concentrations of OxyS–fhlA complex. Our results are compatible with experimental measurements, as we had expected.

Figure 11 shows the experimental measurements and our results. Interestingly, our algorithm predicted the equilibrium concentration of OxyS–fhlA complex quite accurately for the wild-type fhlA and all of its mutants. We also experimented with RNAcofold and UNAFold in this case. Both RNAcofold and UNAFold predict that the percentage of OxyS in complex is approximately 0 in all five cases for the considered fhlA concentrations. This is probably not very surprising as correctly predicting the equilibrium concentrations is a very difficult task and is highly sensitive to the accuracy of the partition functions. We describe below how to compute the concentrations from partition functions, and why it is difficult to correctly predict those equilibrium concentrations.

Given two nucleic acid strands $R$ and $S$, we can compute the equilibrium concentrations of $R$, $S$, $RR$, $SS$ and $RS$ species, denoted by $N_R$, $N_S$, $N_{RR}$, $N_{SS}$ and $N_{RS}$, respectively, from their partition functions (Dimitrov and Zuker, 2004). In the equilibrium, the free energy of a closed system at constant temperature, volume and pressure tends toward a minimum (Landau and Lifshitz, 1969). Equilibrium concentrations are computed from the chemical equilibrium constants

$$
K_R = \frac{Q_{RR}}{Q_R} \frac{N_{RR}}{N_R^2},
$$

$$
K_S = \frac{Q_{SS}}{Q_S} \frac{N_{SS}}{N_S^2},
$$

$$
K_{RS} = \frac{Q_{RS}}{Q_R Q_S} \frac{N_{RS}}{N_R N_S},
$$

under the constraint $N_{RS} = N_{RS} = 0$. $N_R$, $N_S$, $N_{RR}$, $N_{SS}$ and $N_{RS}$, respectively, from their partition functions (Dimitrov and Zuker, 2004). In the equilibrium, the free energy of a closed system at constant temperature, volume and pressure tends toward a minimum (Landau and Lifshitz, 1969). Equilibrium concentrations are computed from the chemical equilibrium constants

$$
K_R = \frac{Q_{RR}}{Q_R} \frac{N_{RR}}{N_R^2},
$$

$$
K_S = \frac{Q_{SS}}{Q_S} \frac{N_{SS}}{N_S^2},
$$

$$
K_{RS} = \frac{Q_{RS}}{Q_R Q_S} \frac{N_{RS}}{N_R N_S},
$$

under the constraint $N_{RS} = N_{RS} = 0$. $N_R$, $N_S$, $N_{RR}$, $N_{SS}$ and $N_{RS}$, in which $N_R$ and $N_S$ are the initial concentrations of single strands. We noticed that $Q_R$ and $Q_S$ computed by the three programs are very close because they use the same algorithm for a single strand (i.e. McCaskill’s algorithm). Therefore based on (8), a method can compute equilibrium concentrations correctly only if it computes each individual $Q_i$ accurately. As one can observe in Figure 11, our program has been able to predict OxyS–fhlA complex concentrations accurately, thus we can conclude that our program computes all $Q_i$ accurately.

As mentioned above, the parameters used by our program on this dataset have been manually optimized. Our energy parameters in this experiment are $\beta_1 = 6.6, \beta_2 = 0.1, \sigma = 0.9, \beta'_1 = 4.5$ and $\sigma' = 0.9$. 

![Fig. 11. Experimental and computational determination of equilibrium constants for pairs of OxyS with wild-type and mutated fhlA. Horizontal axis denotes the initial concentration of fhlA, and the vertical axis denotes the percentage of OxyS in OxyS–fhlA complex. Initial concentration of OxyS was 2 × 10⁻⁸ M (Argaman and Altuvia, 2000). Both RNAcofold and UNAFold predict that the percentage of OxyS in complex is approximately 0 for the considered fhlA concentrations.](image-url)
4 CONCLUSION AND FUTURE WORK

In this article, we present piRNA, an efficient algorithm to compute a partition function of two interacting nucleic acid strands. Our algorithm considers almost all physically possible secondary structures that do not contain pseudoknots, crossing interactions and zigzag’s. In order to specify the free energy of a joint structure established by interacting strands, we extend the standard nearest neighbor single-strand thermodynamic energy model to an energy model for two interacting strands by introducing three new nearest neighbor single-strand thermodynamic energy models. We verified our algorithm by computing the melting temperature for RNA pairs available in the literature and the equilibrium concentration for OxyS-fhlA complex. In both experiments our algorithm provides high accuracy and identifies available alternatives.

We computed the melting temperature for RNA pairs in (Diamond et al., 2001; Mathews and Turner, 2002; Xia et al., 1998) (Tables 1–3). On average, the predicted melting temperature by our Bioch. Biophys. Acta algorithms are more accurate than the alternatives, RNAcofold and UNAFold, on average. It is important to note that RNAcofold and UNAFold both perform poorly in at least one of the three datasets, while our program is consistently accurate across all three datasets. Therefore, neither RNAcofold nor UNAFold is as reliable as our program for melting temperature prediction. In addition, our algorithm is able to compute the OxyS-fhlA complex equilibrium concentrations for wild-type and mutated fhlA accurately. Both RNAcofold and UNAFold predict those equilibrium concentrations to be approximately 0, which does not even roughly follow the experimental measurements.

Although our algorithm is fairly efficient, improving the generality and complexity of our algorithm will be one of our priorities in the near future. In particular, we aim to explore whether it is possible to cover more general interactions without increasing the computational complexity of the algorithm.

**Funding:** Bioinformatics for Combating Infectious Diseases (BCID) initiative (to H.C.), MITACS Research Grant (to R.S.), Michael Smith Foundation for Health Research Career Award (to S.C.S.). German Research Foundation (DFG grant BA 2168/2-1 SPP 1258), and from the German Federal Ministry of Education and Research (BMBF grant O31392 FRISYS) (to R.B.).

**Conflict of Interest:** none declared.

**REFERENCES**


