Supplementary Materials: A Partition Function Algorithm for Interacting Nucleic Acid Strands

Hamidreza Chitsaz [1,*] Raheleh Salari[1,*] S. Cenk Sahinalp [1,†] and Rolf Backofen [2,†]

[1]School of Computing Science, Simon Fraser University,
Burnaby, BC, Canada

[2]Bioinformatics, Institute of Computer Science,
Albert-Ludwigs-Universität, Freiburg, Germany

In this supplementary material, we describe in full the algorithm for computing the partition function of two interacting nucleic acid strands. We also present the sequence pairs in the data sets used for verification of our algorithm.

1 Preliminaries

Throughout this paper, 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 $5' \rightarrow 3'$ with **S** in $3' \leftarrow 5'$ direction. Each nucleotide is paired with at most one nucleotide in the same or the other strand. The subsequence from the i^{th} nucleotide to the j^{th} nucleotide in a strand is denoted by $[i, j]$. We refer to the i^{th} nucleotide in **R** and **S** by i_R and i_S respectively.

An intramolecular base pair between the nucleotides i and j in a strand is called an *arc* and denoted by a bullet $i \bullet j$. An intermolecular base pair between the nucleotides i_R and i_S is called a *bond* and denoted by a circle $i_R \circ i_S$. An arc $i_R \bullet j_R$ *covers* a bond $l_R \circ k_S$ if $i_R < l_R < j_R$. We call $i_R \bullet j_R$ an *interaction arc* if there is a bond $l_R \circ k_S$ covered by $i_R \bullet j_R$. A *kissing arc* is an interaction arc that directly covers a bond. More precisely, we call $i_R \bullet j_R$ a kissing arc if it covers a bond $l_R \circ k_S$ such that if $i'_R \bullet j'_R$ covers the same bond $l_R \circ k_S$, then $i'_R \leq i_R$ and $j_R \leq j'_R$. A subsequence $[i_R, j_R]$ contains a *direct bond*, $l_R \circ k_S$, if $i_R \leq l_R \leq j_R$ and no arc within $[i_R, j_R]$ covers $l_R \circ k_S$. Assuming $i_R < j_R$, two bonds $i_R \circ i_S$ and $j_R \circ j_S$ are called *crossing bonds* if $i_S < j_S$. An interaction arc $i_R \bullet j_R$ in a strand *subsumes* a subsequence $[i_S, j_S]$ in the other strand if for all bonds $l_R \circ k_S$, if $i_S \leq k_S \leq j_S$ then $i_R < l_R < j_R$. Two interaction arcs are *equivalent* if they subsume one another. Two interaction arcs $i_R \bullet j_R$ and $i_S \bullet j_S$ are part of a *zigzag*, if neither $i_R \bullet j_R$ subsumes $[i_S, j_S]$ nor $i_S \bullet j_S$ subsumes $[i_R, j_R]$.

In this work, 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**.

*Joint first authors

†Corresponding authors

2 Interaction Energy Model

An unpseudoknotted secondary structure s of a single nucleic acid, in the standard energy model [4], is decomposed into loops, 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 energy model consists of the following loop types: 1) Hairpin, 2) Interior, and 3) Multiloop. In an interaction secondary structure of two strands under our assumptions [1], 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. Our extended energy model consists of the following components:

- Empty: $G_{i,j}^{\text{empty}}$ is the free energy of a subsequence $[i, j]$ that contains no base pairs and is external to all loops. Its energy contribution is assumed to be zero.
- Hairpin: $G_{i,j}^{\text{hairpin}}$ is the free energy of a hairpin closed by the arc $i \bullet j$. That depends on the sequence and loop size.
- Interior: $G_{i,k_1,k_2,j}^{\text{interior}}$ is the free energy of the interior loop enclosed by the closing arc $i \bullet j$ and the interior arc $k_1 \bullet k_2$. That free energy depends on the closing base pairs and the loop size. An interior loop is called bulge iff one side of the loop has zero length. Stacked pairs are a special case of bulge loops in which case the size of the loop is zero. A stem is a series of stacked pairs.
- Multi: $G_{U,B}^{\text{multi}}$ is the energy of a multiloop with B base pairs and U unpaired bases. It is approximated by
$$G_{U,B}^{\text{multi}} = \alpha_1 + \alpha_2 U + \alpha_3 B, \tag{1}$$
in which α_1 is the penalty for the formation of the multiloop, α_2 is the penalty for each unpaired base in the multiloop, and α_3 is the penalty for each loop in the multiloop.
- Hybrid: $G_{\{k_R^i \circ k_S^i\}}^{\text{hybrid}}$ is the free energy of a joint secondary structure consisting of a series of bonds, $k_R^i \circ k_S^i, i = 1, \ldots, m$, with no intramolecular base pairing or branching. We call such a component *hybrid*. We define the energy associated with a hybrid component by
$$G_{\{k_R^i \circ k_S^i\}}^{\text{hybrid}} = \beta_1 + \sigma \sum_{i=1}^{m-1} G_{k_R^i, k_R^{i+1}, k_S^{i+1}, k_S^i}^{\text{interior}}, \tag{2}$$
in which β_1 is the penalty for the formation of the hybrid, and $\sigma \leq 1$ is the ratio of the free energy of intermolecular to that of intramolecular interior loops (as suggested by [1]). Note that with $\beta_1 = 0, \sigma = 1$, G^{hybrid} is identical to the energy proposed by `RNAhybrid`, first introduced by Rehmsmeier et al. which considers only one hybrid component for mRNA/target duplexes and does not allow any intramolecular structure [7],
- Kissing: $G_{U^k,B^k}^{\text{kissing}}$ is the energy of an intramolecular loop (hairpin, interior, or multiloop) that makes interaction with the other strand and has B^k base pairs and U^k unpaired bases. Such component is called a *kissing loop*. The energy associated with a kissing loop is given by
$$G_{U^k,B^k}^{\text{kissing}} = \beta_2 U^k + \beta_3 B^k, \tag{3}$$

[1] Remember we do not allow pseudoknots, crossing bonds, and zigzags in this work.

in which B^k is the number of loops and U^k the number of unpaired bases in the kissing loop. Note in our model we use different β_1 and σ 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 zero.

3 Interaction Partition Function

Here, we describe a recursive algorithm for computing the partition function of two interacting nucleic acid strands, called `piRNA`. Our algorithm guarantees to consider all possible secondary structures exactly once. We prove every possible secondary structure is reached by exactly one trajectory in the recursion process.

We present our algorithm using recursion diagrams [3, 8]. Our algorithm computes two types of recursive quantities: 1) the partition function of a subsequence $[i, j]$ in one strand, and 2) the joint partition function of subsequences $[i_R, j_R]$ and $[i_S, j_S]$. 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 $[i_R, j_R] \times [i_S, j_S]$ with i_R, j_R, i_S, and j_S terminal bases. The *length pair* of region $[i_R, j_R] \times [i_S, j_S]$ is $(l_R = j_R - i_R + 1, l_S = j_S - i_S + 1)$. Our algorithm starts with $(l_R = 1, l_S = 1)$ and considers all length pairs incrementally up to $(l_R = L_R, l_S = L_S)$. For a fixed length pair (l_R, l_S), recursive quantities for all the regions $[i_R, i_R + l_R - 1] \times [i_S, i_S + l_S - 1]$ are computed.

3.1 Partition Function for Non-Interacting Subsequences

For computing the partition function of a subsequence in one strand we use McCaskill's algorithm [6]. McCaskill's algorithm is shown in Figure 1, in which $Q_{i,j}$ is the partition function for the subsequence $[i, j]$. Throughout this paper, 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 which 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 1, the first case of $Q_{i,j}$ corresponds to an empty structure (that constitutes no base pairs) whose free energy is assumed to be zero, thus its contribution to the partition function is $e^{-G^{empty}_{i,j}/RT} = 1$. In the other case, there exists at least one arc and the leftmost one is $k_1 \bullet k_2$. It contributes $Q^b_{k_1,k_2} Q_{k_2+1,j}$ to the partition function, therefore,

$$Q_{i,j} = 1 + \sum_{i \le k_1 < k_2 \le j} Q^b_{k_1,k_2} Q_{k_2+1,j}. \tag{4}$$

The second line shows the cases of $Q^b_{i,j}$ which is the partition function for the subsequence $[i, j]$ assuming i and j are base paired. The arc $i \bullet j$ can close different substructures: hairpin, interior,

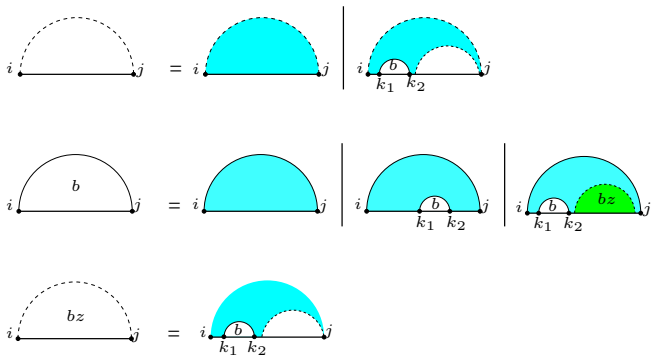

Figure 1: McCaskill's algorithm: recursion for $Q_{i,j}$, the partition function for the subsequence $[i,j]$. Above, $Q^b_{i,j}$ is the partition function for the subsequence $[i,j]$ assuming i and j are base paired, and $Q^{bz}_{i,j}$ is the partition function for the subsequence $[i,j]$ assuming there is at least one arc in the region.

or multiloop. The energy contribution of each substructure is calculated based on the standard thermodynamics energy model.

$$\begin{aligned} Q^b_{i,j} =& e^{-G^{\text{hairpin}}_{i,j}/RT} + \sum_{i\le k_1<k_2\le j} e^{-G^{\text{interior}}_{i,k_1,k_2,j}/RT} + \\ & \sum_{i\le k_1<k_2\le j} Q^b_{k_1,k_2} Q^{bz.green}_{k_2+1,j-1}\ e^{-(\alpha_1+\alpha_2(k_1-i-1)+\alpha_3)/RT}. \end{aligned} \tag{5}$$

The third line shows cases of $Q^{bz}_{i,j}$ which is the partition function for the subsequence $[i,j]$ assuming the region constitutes at least one arc. Therefore,

$$Q^{bz}_{i,j} = \sum_{i\le k_1<k_2\le j} Q^b_{k_1,k_2} Q_{k_2+1,j}. \tag{6}$$

As mentioned before, a green region is contained in a multiloop. The region has the same recursion as if it was white, however the base pair and unpaired penalties of multiloop should be applied to it. Explicitly,

$$Q^{bz}_{i,j} = \sum_{i\le k_1<k_2\le j} Q^b_{k_1,k_2} Q^{\cdot green}_{k_2+1,j}\ e^{-(\alpha_2(k_1-i-1)+\alpha_3)/RT}, \tag{7}$$

$$Q^{\cdot green}_{i,j} = e^{-\alpha_2(j-i-1)/RT} + \sum_{i\le k_1<k_2\le j} Q^b_{k_1,k_2} Q^{\cdot green}_{k_2+1,j}\ e^{-(\alpha_2(k_1-i-1)+\alpha_3)/RT}. \tag{8}$$

3.2 Partition Function for Interacting Subsequences

In the following, we present all cases of $Q^I_{i_R,j_R,i_S,j_S}$ which is the interaction partition function for the region $[i_R,j_R]\times[i_S,j_S]$. 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

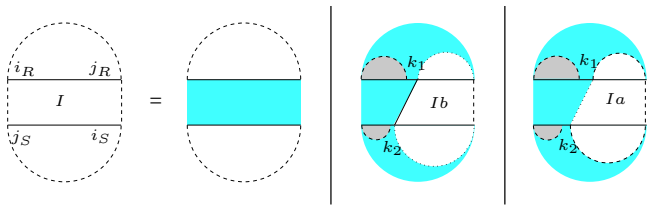

Figure 2: Cases of the interaction partition function $Q^I_{i_R,j_R,i_S,j_S}$. The first case constitutes no bonds. In the second case, the leftmost bond is a direct bond on both subsequences. In the third case, the leftmost bond is covered by an interaction arc in at least one subsequence.

denotes two terminal bases of a region which are assumed to be unpaired. Figure 2 shows the cases of Q^I: 1) there is no bond between the two subsequences, 2) the leftmost bond is a direct bond in both subsequences, and 3) the leftmost bond is covered by an arc in at least one subsequence. Therefore,

$$\begin{aligned} Q^I_{i_R,j_R,i_S,j_S} =& Q_{i_R,j_R} Q_{i_S,j_S} + \sum_{\substack{i_R \le k_1 < j_R \\ i_S < k_2 \le j_S}} Q_{i_R,k_1-1} Q_{k_2+1,j_S} Q^{Ib}_{k_1,j_R,i_S,k_2} + \\ & \sum_{\substack{i_R \le k_1 < j_R \\ i_S < k_2 \le j_S}} Q_{i_R,k_1-1} Q_{k_2+1,j_S} Q^{Ia}_{k_1,j_R,i_S,k_2}, \end{aligned} \tag{9}$$

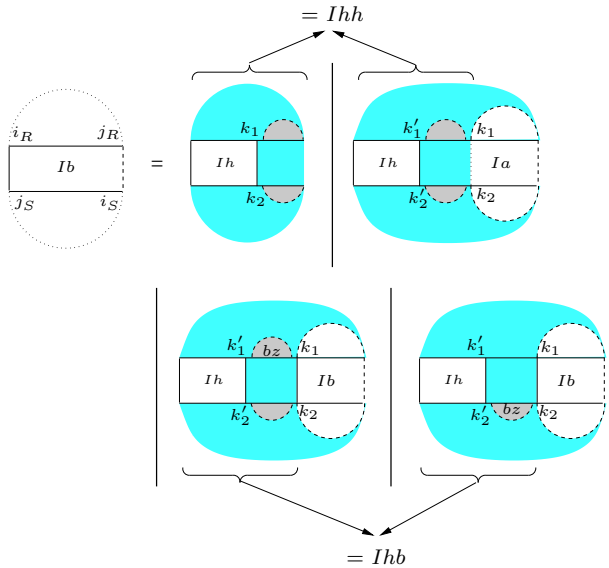

Figure 3: Recursion for $Q^{Ib}_{i_R,j_R,i_S,j_S}$ assuming $i_R \circ j_S$ is a bond. We show a version of the recursion that contains two split points in each sequence for simplicity reasons. However, this would increase the complexity and can easily be resolved by introducing two additional matrices Q^{Ihh} and Q^{Ihb} for the region $[i_R, k_1] \times [k_2, j_S]$ as indicated by the arrows.

Figure 3 shows the recursion for $Q^{Ib}_{i_R,j_R,i_S,j_S}$, the interaction partition function for the region $[i_R, j_R] \times [i_S, j_S]$ assuming $i_R \circ j_S$ is a bond. Since we have penalties for opening and closing

a hybrid component, the recursion for Q^{Ib} has to determine whether the region contains one or several hybrid components. In all cases, Q^{Ih} contains the full hybrid component containing the bond $i_R \circ j_S$ (see Figure 5 for Q^{Ih} 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 1) it is not direct in at least one subsequence, i.e. it is covered by an arc in the associated regions (case 2 of the Q^{Ib} recursion), or 2) there is at least one arc between the two successive intermolecular bonds (case 3 and 4 of the Q^{Ib} recursion). Using the additional matrices Q^{Ihh} and Q^{Ihb}, we get

$$\begin{aligned} Q^{Ib}_{i_R,j_R,i_S,j_S} =& Q^{Ihh}_{i_R,j_R,i_S,j_S} + \sum_{\substack{i_R<k_1<j_R \\ i_S<k_2<j_S}} Q^{Ihb}_{i_R,k_1,k_2,j_S} Q^{Ib}_{k_1,j_R,i_S,k_2} + \\ & \sum_{\substack{i_R<k_1<j_R \\ i_S<k_2<j_S}} Q^{Ihh}_{i_R,k_1,k_2,j_S} Q^{Ia}_{k_1,j_R,i_S,k_2}. \end{aligned} \tag{10}$$

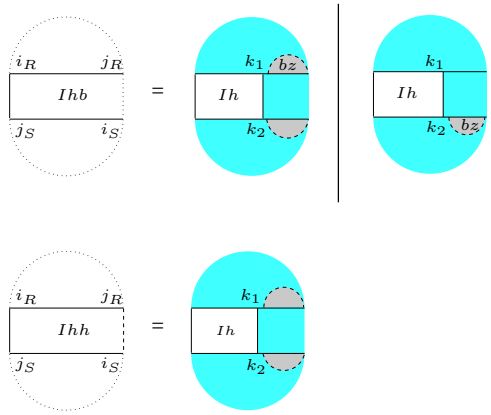

Figure 4: Cases of $Q^{Ihb}_{i_R,j_R,i_S,j_S}$ and $Q^{Ihh}_{i_R,j_R,i_S,j_S}$ whose region contains one hybrid component on the left. Here, region $[i_R, k_1] \times [k_2, j_S]$ represents a hybrid component. Figure 5 shows the recursion for Q^{Ih}.

The quantities Q^{Ihh} and Q^{Ihb} are defined by the recursion diagrams in Figure 4 and equivalently by the following equations:

$$Q^{Ihb}_{i_R,j_R,i_S,j_S} = \sum_{\substack{i_R\le k_1\le j_R \\ i_S\le k_2\le j_S}} e^{-\beta_1/RT} Q^{Ih}_{i_R,k_1,k_2,j_S} (Q^{bz}_{k_1+1,j_R} Q_{i_S,k_2-1} + Q^{bz}_{i_S,k_2-1}) \tag{11}$$

and

$$Q^{Ihh}_{i_R,j_R,i_S,j_S} = \sum_{\substack{i_R\le k_1\le j_R \\ i_S\le k_2\le j_S}} e^{-\beta_1/RT} Q^{Ih}_{i_R,k_1,k_2,j_S} Q_{k_1+1,j_R} Q_{i_S,k_2-1}, \tag{12}$$

in which Q^{Ih} is the interaction partition function for a hybridization region (Figure 5).

Figure 5 shows the recursion for Q^{Ih}. Since we do not allow isolated bond the base case of Q^{Ih} is an interior loop, otherwise it can be an isolated bond. Two cases are possible: 1) there is no bond other than $i_R \circ j_S$ and $i_S \circ j_R$ in the region, and 2) there exist more bonds between $i_R \circ j_S$ and $i_S \circ j_R$, the leftmost of which is $k_1 \circ k_2$. Precisely,

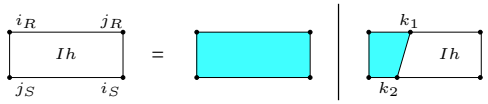

Figure 5: Cases of $Q^{Ih}_{i_R,j_R,i_S,j_S}$ the interaction partition function for a single hybrid component.

$$Q^{Ih}_{i_R,j_R,i_S,j_S} = e^{-\sigma G^{\text{interior}}_{i_R,j_R,i_S,j_S}/RT} + \sum_{\substack{i_R \le k_1 \le j_R \\ i_S \le k_2 \le j_S}} e^{-\sigma G^{\text{interior}}_{i_R,k_1,k_2,j_S}/RT} Q^{Ih}_{k_1,j_R,i_S,k_2}. \tag{13}$$

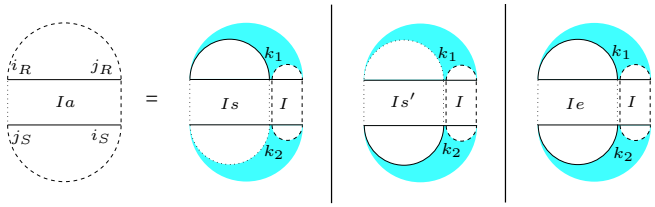

Figure 6: Cases of $Q^{Ia}_{i_R,j_R,i_S,j_S}$, for which we assume at least one of i_R and j_S is the end point of an interaction arc.

Figure 6 shows the cases of $Q^{Ia}_{i_R,j_R,i_S,j_S}$ for which at least one of i_R and j_S is the end point of interaction arc: 1) $i_R \bullet k_1$ subsumes $[k_2, j_S]$ and k_2 is not base paired with j_S, 2) $k_2 \bullet j_S$ subsumes $[i_R, k_1]$ and i_R is not base paired with k_1, and 3) $i_R \bullet k_1$ and $k_2 \bullet j_S$ are equivalent. If only one of i_R and j_S 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_S are end points of interaction arcs, then one of the arcs subsumes the other one or they are equivalent. Therefore,

$$\begin{aligned} Q^{Ia}_{i_R,j_R,i_S,j_S} = & \sum_{\substack{i_R < k_1 \le j_R \\ i_S \le k_2 \le j_S}} Q^{Is}_{i_R,k_1,k_2,j_S} Q^{I}_{k_1+1,j_R,i_S,k_2-1} + \\ & \sum_{\substack{i_R \le k_1 \le j_R \\ i_S < k_2 \le j_S}} Q^{Is'}_{i_R,k_1,k_2,j_S} Q^{I}_{k_1+1,j_R,i_S,k_2-1} + \\ & \sum_{\substack{i_R < k_1 \le j_R \\ i_S < k_2 \le j_S}} Q^{Ie}_{i_R,k_1,k_2,j_S} Q^{I}_{k_1+1,j_R,i_S,k_2-1}, \end{aligned} \tag{14}$$

in which $Q^{Is}_{i_R,k_1,k_2,j_S}$ is the interaction partition function of $[i_R, k_1] \times [k_2, j_S]$ assuming $i_R \bullet k_1$ is an interaction arc that subsumes $[k_2, j_S]$, $Q^{Is'}_{i_R,k_1,k_2,j_S}$ is the symmetric counterpart of Q^{Is}, and $Q^{Ie}_{i_R,k_1,k_2,j_S}$ is the interaction partition function of $[i_R, k_1] \times [k_2, j_S]$ assuming $i_R \bullet k_1$ and $k_2 \bullet j_S$ are equivalent interaction arcs.

For Q^{Ie}, it does not make any difference which one of the covering arcs $i_R \bullet j_R$ and $i_S \bullet j_S$ is extracted first. We first extract the covering arc from $\mathbf{S}$ (see Figure 7). Extracting the covering arc,

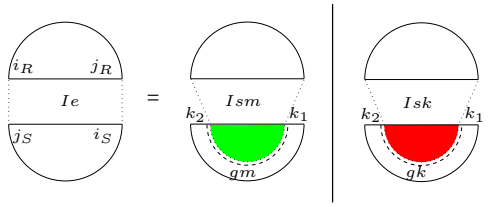

Figure 7: Cases of $Q^{Ie}_{i_R,j_R,i_S,j_S}$, for which $i_R \bullet j_R$ and $i_S \bullet j_S$ are equivalent interaction arcs.

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 7 is appropriately colored by green and red to remind the type of penalty. So, we have

$$Q^{Ie}_{i_R,j_R,i_S,j_S} = \sum_{i_S<k_1<k_2<j_S} Q^{Ism.green}_{i_R,k_1,k_2,j_S} Q^{gm}_{i_S,k_1-1,k_2+1,j_S} + Q^{Isk.red}_{i_R,k_1,k_2,j_S} Q^{gk}_{i_S,k_1-1,k_2+1,j_S}. \tag{15}$$

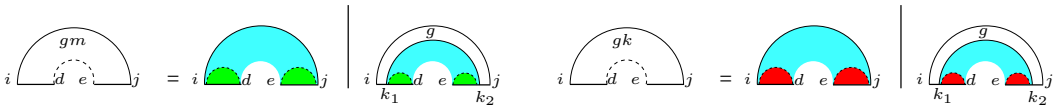

Figure 8: Recursion for $Q^{gk}_{i,d,e,j}$ and $Q^{gm}_{i,d,e,j}$ the partition functions for $[i,j]$ excluding the gap $[d,e]$, assuming i and j are base paired. For Q^{gk}, the gap contains a direct bond, and for Q^{gm} the gap contains multiple interaction arcs.

Note that $Q^{gm}_{i,d,e,j}$ and $Q^{gk}_{i,d,e,j}$ are the partition functions for $[i,j]$ excluding the gap $[d,e]$, assuming i and j are base paired. For Q^{gm} the gap contains multiple interaction arcs, and for Q^{gk}, the gap contains a direct bond (see Figure 8). Therefore,

$$Q^{gm}_{i,d,e,j} = \sum_{\substack{i\le k_1\le d \\ e\le k_2\le j}} Q^{g}_{i,k_1,k_2,j} Q^{.green}_{k_1+1,d} Q^{.green}_{k_2-1,j}, \tag{16}$$

and

$$Q^{gk}_{i,d,e,j} = \sum_{\substack{i\le k_1\le d \\ e\le k_2\le j}} Q^{g}_{i,k_1,k_2,j} Q^{.red}_{k_1+1,d} Q^{.red}_{k_2-1,j}. \tag{17}$$

The gap partition function Q^g is defined by the recursion in Figure 9. This quantity is similar to the g in Dirks-Pierce's algorithm [3]. For $Q^{g}_{i,d,e,j}$, we assume $i \bullet j$ and $d \bullet e$. Note that $i = d, j = e$ is a single arc case. There are two groups of cases: 1) there is no more spanning arc in the region, and 2) there is at least another outermost spanning arc $k_1 \bullet k_2$. In both groups, there could be some additional structure in the region. If there is no additional structure in the region, then the spanning region is an interior loop. If there is at least one arc in either side of the gap, then the spanning region forms a multiloop and penalty of multiloop should be applied.

Let $Q^{Is}_{i_R,j_R,i_S,j_S}$ be the partition function for $[i_R,j_R] \times [i_S,j_S]$ assuming $i_R \bullet j_R$ is an interaction arc that subsumes $[i_S,j_S]$. Since the union of the cases of Q^{Isk} and Q^{Ism} comprise the cases of

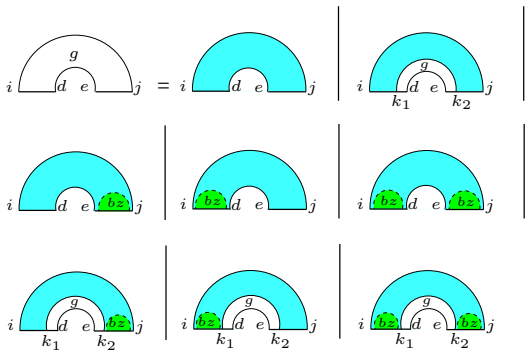

Figure 9: Recursion for $Q^g_{i,d,e,j}$ the partition function for the subsequence $[i, j]$ excluding the gap $[d, e]$ assuming $i \bullet j$ and $d \bullet e$.

Q^{Is},

$$Q^{Is}_{i_R,j_R,i_S,j_S} = Q^{Isk}_{i_R,j_R,i_S,j_S} + Q^{Ism}_{i_R,j_R,i_S,j_S}. \tag{18}$$

In particular, Q^{Isk} contains all cases of Q^{Is} in which $[i_S, j_S]$ has at least one direct bond, and Q^{Ism} contains all cases of Q^{Is} in which $[i_S, j_S]$ includes multiple interaction arcs. Similarly, we extract the covering arc from Q^{Isk} and Q^{Ism} to obtain Q^{Imm}, Q^{Imk}, Q^{Ikm}, and Q^{Ikk}, where k stands for kissing (or equivalently containing a direct bond) and m for multiple interaction arcs. The quantities $Q^{Imm}_{i_R,j_R,i_S,j_S}$, $Q^{Ikm}_{i_R,j_R,i_S,j_S}$, $Q^{Imk}_{i_R,j_R,i_S,j_S}$, and $Q^{Ikk}_{i_R,j_R,i_S,j_S}$ are defined by recursions in Figs. 11, 12, 13, and 14. Note that all four terminal bases of the region have to be the end points of a bond or of an interaction arc. In summary:

- Q^{Imm} includes all cases that have multiple interaction arcs in both $[i_R, j_R]$ and $[i_S, j_S]$.
- Q^{Imk} includes all cases where $[i_R, j_R]$ has multiple interaction arcs and $[i_S, j_S]$ has at least one direct bond.
- Q^{Ikm} is symmetric to Q^{Imk} with respect to **R** and **S**.
- Q^{Ikk} includes all cases where both $[i_R, j_R]$ and $[i_S, j_S]$ have at least one direct bond.

In Q^{Imm}, both subsequences $[i_R, j_R]$ and $[i_S, j_S]$ include multiple interaction arcs and have no direct bond (Figure 11). All four terminal bases are endpoints of interaction arcs. Since i_R and j_S are endpoints of interaction arc, there must exist a Q^{Ia} on the left side of the region. This Q^{Ia} has no direct bond on both subsequences from **R** and **S**, which we call $Q^{Ia_{nn}}$. The bases j_R and i_S are also end points of interaction arc, so there are interaction arcs on the right side of the Q^{Imm} in both subsequences. These arcs can have three types: 1) arc in subsequence $[i_R, j_R]$ subsumes the arc in subsequence $[i_S, j_S]$, 2) arc in subsequence $[i_S, j_S]$ subsumes the arc in subsequence $[i_R, j_R]$, or 3) two arcs are equivalent. Note that for multiple interaction arcs there are an Q^{Ie}, Q^{Is} or $Q^{Is'}$ on both left and right side of the region. The left one is contained in an extracted Q^{Ia}, and the right one is extracted separately. This scheme will continue for the other cases with multiple interaction arcs.

In Q^{Imk}, subsequence $[i_R, j_R]$ has multiple interaction arcs and subsequence $[i_S, j_S]$ has at least one direct bond (Figure 12). Here, i_R and j_R are the end points of an interaction arc and i_S and j_S are the end points of a bond or interaction arc. Since i_R is the end point of an interaction arc, there must exist a Q^{Ia} on the left side of the region. The Q^{Ia} has no direct bond in the subsequence

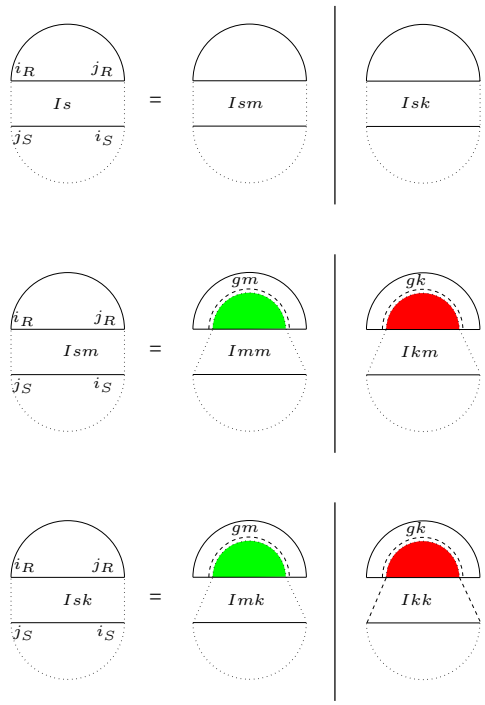

Figure 10: Recursion for $Q^{Is}_{i_R,j_R,i_S,j_S}$, interaction partition function assuming $i_R \bullet j_R$ is an interaction arc subsuming $[i_S, j_S]$. In Q^{Ism}, $[i_S, j_S]$ contains multiple interaction arcs and in Q^{Isk}, $[i_S, j_S]$ contains at least one direct bond.

of **R**, but it can have two cases with direct bond in subsequence of **S**. We denote the special Q^{Ia} that has at least one direct bond in the subsequence of **S** by $Q^{Ia_{nd}}$. In this case, the arc on the right side of the subsequence of **R** can have three types: 1) it subsumes an interacting region in $[i_S, j_S]$, 2) it is subsumed by the interaction arc on the right side of $[i_S, j_S]$, and 3) it is equivalent to the interaction arc on the right side of $[i_S, j_S]$. Note that the arc on $[i_S, j_S]$ can only subsume subsequences with multiple interaction arcs. If the Q^{Ia} has no direct bond in **S** subsequence, in which case it is denoted by $Q^{Ia_{nn}}$, the arc on the right side of $[i_R, j_R]$ should subsume a subsequence on the right side of $[i_S, j_S]$ that has at least one direct bond. The quantity Q^{Ikm} is symmetric to Q^{Imk} with respect to **R** and **S** (Figure 13).

In Q^{Ikk}, both subsequences of **R** and **S** have at least one direct bond, and all four terminal bases of the region can be end points of bond or interaction arc (Figure 14). We go through the cases based on different possibilities of terminal bases. If two terminal bases at the same side of the region are end points of a bond, then obviously they are base paired, otherwise at least one of them is the end point of an interaction arc.

In the first case of Figure 14, all four terminal bases are end points of bond, i.e. $i_R \circ j_S$ and $j_R \circ i_S$. This case is similar to Q^{Ib} with a bond on its right. We denote this special Q^{Ib} by Q^{Ib_r} which is shown in Figure 15. If just $i_R \circ j_S$, then there is a Q^{Ib} on left side of the region. In that case, the right side has three cases: 1) the right side of $[i_R, j_R]$ contains an interaction arc that subsumes a subsequence on the right side of $[i_S, j_S]$, 2) the right side of $[i_S, j_S]$ contains an interaction arc that subsumes a subsequence on the right side of $[i_R, j_R]$, and 3) there are equivalent interaction arcs on the right sides of $[i_R, j_R]$ and $[i_S, j_S]$. If just $j_R \circ i_S$, then the case is similar to a Q^{Ia} with a bond on its right. We denote this special Q^{Ia} by Q^{Ia_r} (Figure 15).

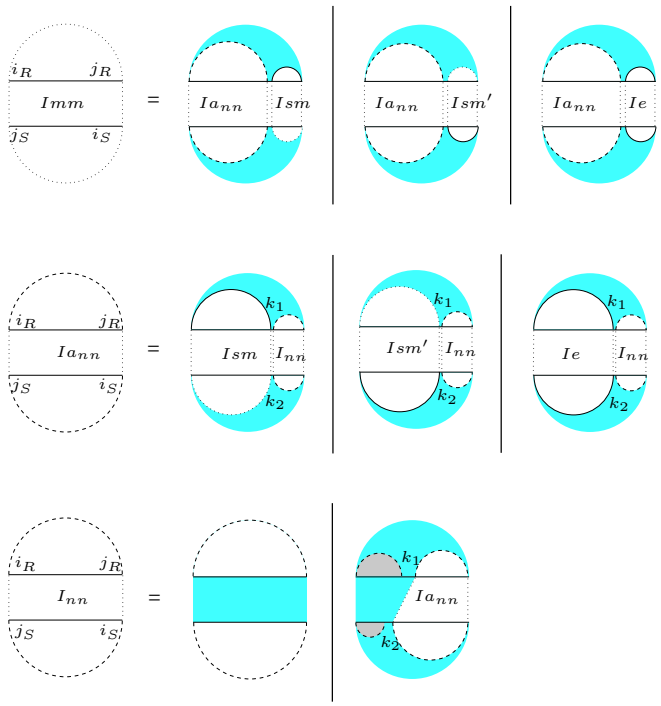

Figure 11: Recursions for $Q^{Imm}_{i_R,j_R,i_S,j_S}$ assuming both $[i_R, j_R]$ and $[i_S, j_S]$ have multiple interaction arcs.

Now consider cases in which terminal bases neither on the left nor on the right make bond with one another. In this type of cases, there must exist a Q^{Ia} on the left side of the region. This Q^{Ia} may contain direct bonds on either subsequence. Denote the special Q^{Ia} that has at least one direct bond in both subsequences by $Q^{Ia_{dd}}$. The right side of the region has three cases: 1) there is an interaction arc on the right side of the remaining subsequence of **R** that subsumes a subsequence on the right side of **S**, 2) there is an interaction arc on the right side of the subsequence of **S**, that subsumes a subsequence on the right side of **R**, and 3) there are equivalent interaction arcs on the right sides of the subsequences of **R** and **S**. Denote the special Q^{Ia} that has at least one direct bond in the subsequence of **R** by $Q^{Ia_{dn}}$. There must exist an interaction arc on the right side of the subsequence of **R** that subsumes a subsequence on the right side of **S**. Note that the subsequence on the right side of **S** should have at least one direct bond. We denote the special Q^{Ia} that has at least one direct bond in the subsequence of **S** by $Q^{Ia_{nd}}$. In that case, there must exist an interaction arc on the right side of the subsequence of **S** that subsumes a subsequence on the right side of **R**. Note that the subsequence on the right side of **R** should have at least one direct bond.

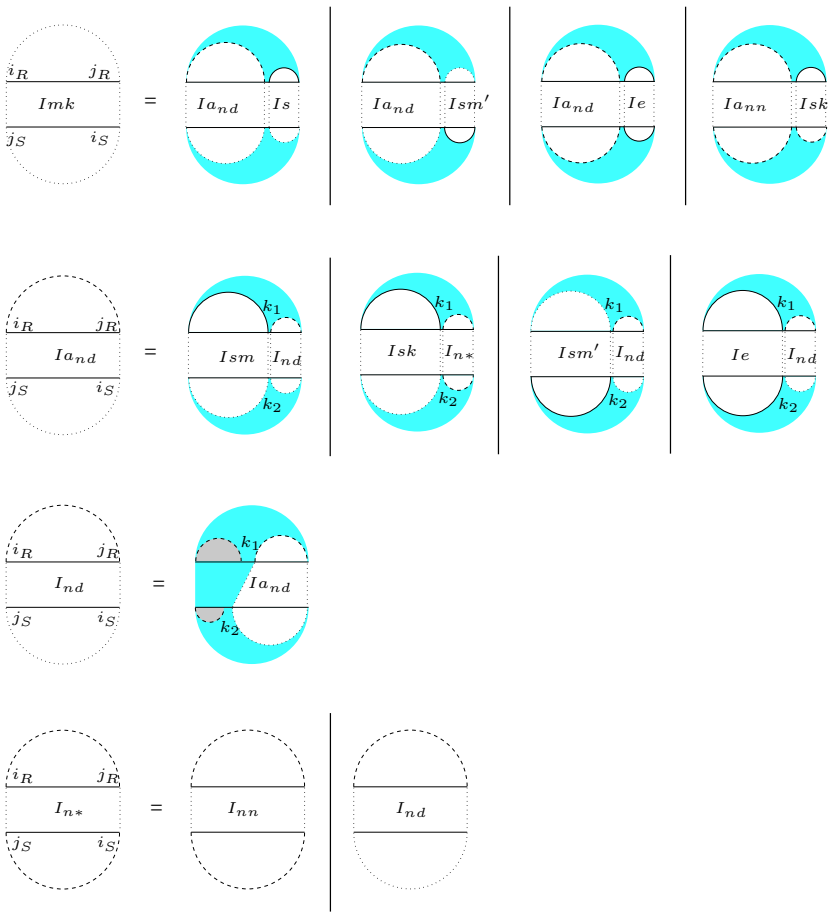

Figure 12: Recursions for $Q^{Imk}_{i_R,j_R,i_S,j_S}$ assuming $[i_R, j_R]$ has multiple interaction arcs and $[i_S, j_S]$ has at least one direct bond.

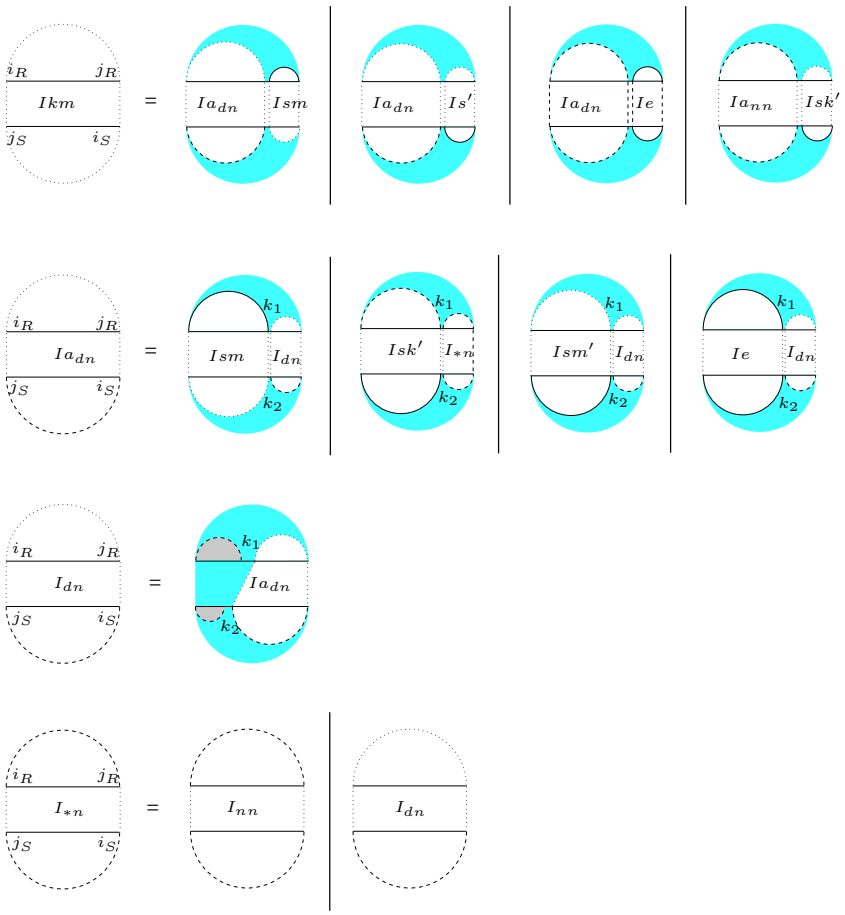

Figure 13: Recursions for $Q^{Ikm}_{i_R,j_R,i_S,j_S}$ assuming $[i_R, j_R]$ has at least one direct bond and $[i_S, j_S]$ has multiple interaction arcs.

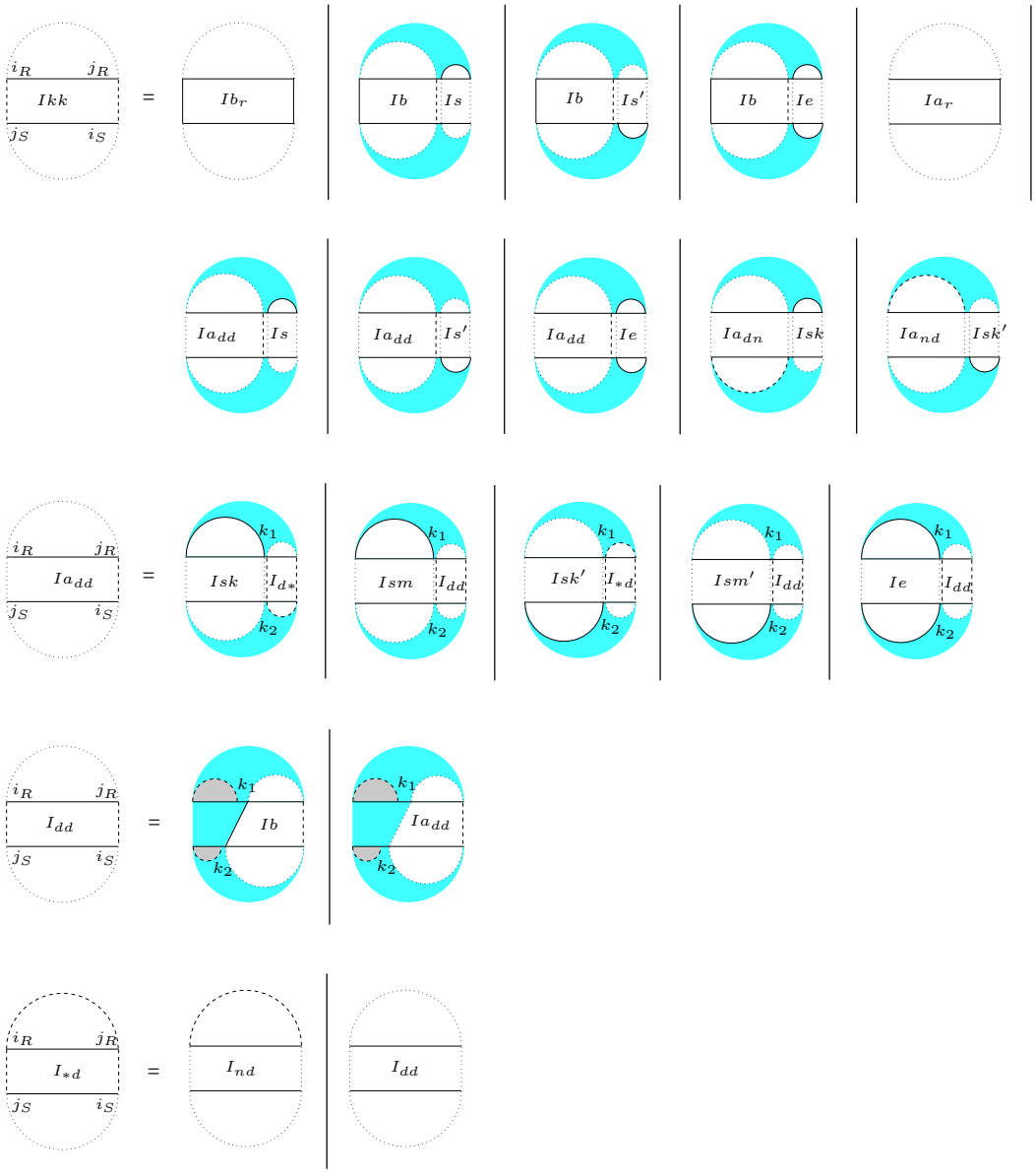

Figure 14: Recursions for $Q^{Ikk}_{i_R,j_R,i_S,j_S}$ assuming both $[i_R, j_R]$ and $[i_S, j_S]$ have at least one direct bond.

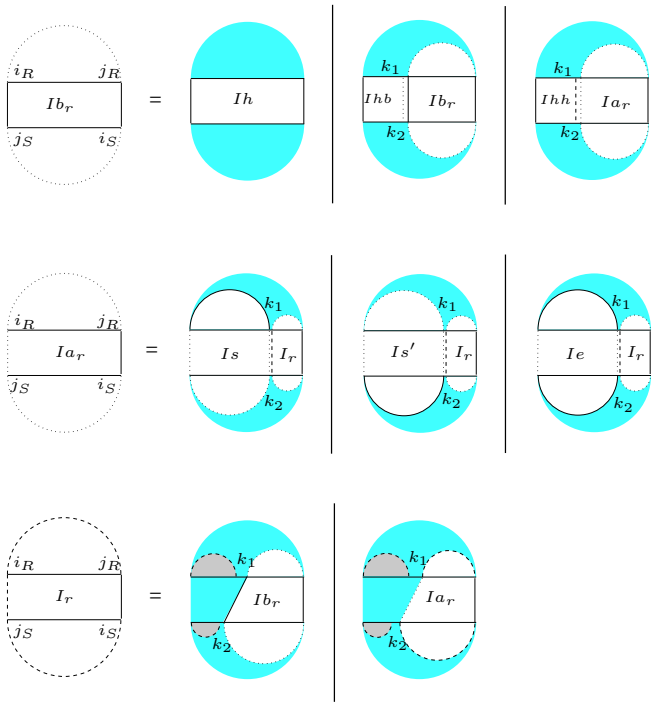

Figure 15: The quantities Q^{I_r}, Q^{Ib_r} and Q^{Ia_r} are some auxiliary quantities similar to Q^I, Q^{Ib} and Q^{Ia} except that there is a bond on their right side.

4 Data Sets

For verification of our algorithm in predicting the melting temperature, we used three data sets available in the literature. Sequences of the RNA pairs from the first data set, which were originally reported in Table 3 of [9], have been explicitly mentioned in the paper. Here, we present sequence of the RNA pairs from the second (originally reported in Table 1 of [2]) and third (originally reported in Tables 3 and 4 of [5]) data sets that have been used in Tables 2 and 3 of the paper.

Table 1: Sequences of the set of RNA pairs reported in Table 2 of the paper.

Pairs	**Sequences**
A	GGAGCGGCUUCGGCCGGACG /CGUCaaCUCC
B	GGAGaCGGCUUCGGCCGGACG /CGUCauaCUCC
C	GGAGaCGGCUUCGGCCGGCAG /CUGCauaCUCC
D	GGAGgCGGCUUCGGCCGuGACG /CGUCcauaCUCC
E	GGAGaCGGCUUCGGCCGcGACG /CGUCauaCUCC
F	GGAGgCGGCUUCGGCCGuGACG /CGUCauaCUCC
G	GGAGCGGCUUCGGCCGGACG /CGUCCUCC
H	GGAGaCGGCUUCGGCCGGACG /CGUCcauaCUCC
I	GGAGCGGCUUCGGCCGGACG /CGUCauaCUCC
J	GGAGCGGCUUCGGCCGGACG /CGUCcauaCUCC
K	GGAGaCGGCUUCGGCCGcGACG /CGUCcauaCUCC
L	GGAGaCGGCUUCGGCCGaGACG /CGUCcauaCUCC

Table 2: Sequences of the set of RNA pairs reported in Table 3 of the paper.

Pairs	**Sequences**
G-GC-G/C-C	GGCAGGCGCUUCGGCGCGGAGG /CCUCCCUGCC
G-GC-G/CaC	GGCAGGCGCUUCGGCGCGGAGG /CCUCCaCUGCC
G-GC-G/Ca_2C	GGCAGGCGCUUCGGCGCGGAGG /CCUCCaaCUGCC
G-GC-G/Ca_3C	GGCAGGCGCUUCGGCGCGGAGG /CCUCCaaaCUGCC
G-GC-G/CauaC	GGCAGGCGCUUCGGCGCGGAGG

Continued on next page

Table 2 – continued from previous page

Pairs	Sequences
	/CCUCCauaCUGCC
G-GC-G/Ca_4C	GGCAGGCGCUUCGGCGCGGAGG
	/CCUCCaaaaCUGCC
GaGC-G/C-C	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCCUGCC
GaGC-G/CaC	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCaCUGCC
GaGC-G/Ca_2C	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCaaCUGCC
GaGC-G/Ca_3C	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCaaaCUGCC
GaGC-G/CauaC	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCauaCUGCC
GaGC-G/Ca_4C	GGCAGaGCGCUUCGGCGCGGAGG
	/CCUCCaaaaCUGCC
Ga_2GC-G/C-C	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCCUGCC
Ga_2GC-G/CaC	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCaCUGCC
Ga_2GC-G/Ca_2C	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCaaCUGCC
Ga_2GC-G/Ca_3C	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCaaaCUGCC
Ga_2GC-G/CauaC	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCauaCUGCC
Ga_2GC-G/Ca_4C	GGCAGaaGCGCUUCGGCGCGGAGG
	/CCUCCaaaaCUGCC
Ga_2GCaG/C-C	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCCUGCC
Ga_2GCaG/CaC	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCaCUGCC
Ga_2GCaG/Ca_2C	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCaaCUGCC
Ga_2GCaG/Ca_3C	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCaaaCUGCC
Ga_2GCaG/CauaC	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCauaCUGCC
Ga_2GCaG/Ca_4C	GGCAGaaGCGCUUCGGCGCaGGAGG
	/CCUCCaaaaCUGCC
Ga_2GCa_2G/C-C	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCCUGCC
Ga_2GCa_2G/CaC	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCaCUGCC
Ga_2GCa_2G/Ca_2C	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCaaCUGCC
Ga_2GCa_2G/Ca_3C	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCaaaCUGCC
Ga_2GCa_2G/CauaC	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCauaCUGCC
Ga_2GCa_2G/Ca_4C	GGCAGaaGCGCUUCGGCGCaaGGAGG
	/CCUCCaaaaCUGCC

Continued on next page

Table 2 – continued from previous page

Pairs	Sequences
G-UA-G/C-C	GGCAGUCGCUUCGGCGAGGAGG /CCUCCCUGCC
G-UA-G/CaC	GGCAGUCGCUUCGGCGAGGAGG /CCUCCaCUGCC
G-UA-G/Ca_2C	GGCAGUCGCUUCGGCGAGGAGG /CCUCCaaCUGCC
G-UA-G/Ca_3C	GGCAGUCGCUUCGGCGAGGAGG /CCUCCaaaCUGCC
G-UA-G/CauaC	GGCAGUCGCUUCGGCGAGGAGG /CCUCCauaCUGCC
G-UA-G/Ca_4C	GGCAGUCGCUUCGGCGAGGAGG /CCUCCaaaaCUGCC
GaUA-G/C-C	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCCUGCC
GaUA-G/CaC	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCaCUGCC
GaUA-G/Ca_2C	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCaaCUGCC
GaUA-G/Ca_3C	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCaaaCUGCC
GaUA-G/CauaC	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCauaCUGCC
GaUA-G/Ca_4C	GGCAGaUCGCUUCGGCGAGGAGG /CCUCCaaaaCUGCC
G-CG-GC-G/C-C	GGCAGCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCCUGCC
G-CG-GC-G/CaC	GGCAGCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaCUGCC
G-CG-GC-G/Ca_2C	GGCAGCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaCUGCC
G-CG-GC-G/Ca_3C	GGCAGCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaaCUGCC
G-CG-GC-G/Ca_4C	GGCAGCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaaaCUGCC
GaCG-GC-G/C-C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCCUGCC
GaCG-GC-G/CaC	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaCUGCC
GaCG-GC-G/Ca_2C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaCUGCC
GaCG-GC-G/Ca_3C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaaCUGCC
GaCG-GC-G/CauaC	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCauaCUGCC
GaCG-GC-G/Ca_4C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCGGAGG /CCUCCaaaaCUGCC
GaCG-GCaG/C-C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG /CCUCCCUGCC
GaCG-GCaG/CaC	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG /CCUCCaCUGCC
GaCG-GCaG/Ca_2C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG

Continued on next page

Table 2 – continued from previous page

Pairs	Sequences
	/CCUCCaaCUGCC
GaCG-GCaG/Ca_3C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG /CCUCCaaaCUGCC
GaCG-GCaG/CauaC	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG /CCUCCauaCUGCC
GaCG-GCaG/Ca_4C	GGCAGaCGGCUUCGGCCGGCGCGCAAGCGCaGGAGG /CCUCCaaaaCUGCC
$Ga_2CGa_2GCa_2G$/C-C	GGCAGaaCGGCUUCGGCCGaaGCGCGCAAGCGCaaGGAGG /CCUCCCUGCC
$Ga_2CGa_2GCa_2G$/CaC	GGCAGaaCGGCUUCGGCCGaaGCGCGCAAGCGCaaGGAGG /CCUCCaCUGCC
$Ga_2CGa_2GCa_2G$/Ca_2C	GGCAGaaCGGCUUCGGCCGaaGCGCGCAAGCGCaaGGAGG /CCUCCaaCUGCC

References

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

[2] J. Diamond, D. Turner, and D. Mathews. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, 40:6971–6981, Jun 2001.

[3] R. M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.

[4] D. Mathews, J. Sabina, M. Zuker, and D. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, May 1999.

[5] D. Mathews and D. Turner. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, 41:869–880, Jan 2002.

[6] J. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[7] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, Oct 2004.

[8] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, Feb 1999.

[9] T. Xia, J. SantaLucia, M. Burkard, R. Kierzek, S. Schroeder, X. Jiao, C. Cox, and D. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, Oct 1998.