BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–10

Partition Function and Base Pairing Probabilities for RNA-RNA Interaction Prediction

Fenix W.D. Huang [1], Jing Qin [1], Christian M. Reidys [1,2]*, and Peter F. Stadler[3−7]

[1]Center for Combinatorics, LPMC-TJKLC, Nankai University Tianjin 300071, P.R. China
[2]College of Life Science, Nankai University Tianjin 300071, P.R. China
[3]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany.
[4]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
[5]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1,D-04103 Leipzig, Germany
[6]Inst. f. Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria
[7]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico, USA

Received on *****; revised on *****; accepted on *****

Associate Editor: *****

ABSTRACT

The RNA-RNA interaction problems (RIP) deals with the energetically optimal structure of two RNA molecules that bind to each other. The standard model introduced by Alkan *et al.* (J. Comput. Biol. **13**: 267-282, 2006) allows secondary structures in both partners as well as additional base pairs between the two RNAs subject to certain restrictions that allow a polynomial time dynamic programming solution. We derive the partition function for RIP based on a notion of "tight structures" as an alternative to the approach of Chitsaz *et al.* (Bioinformatics, **25**, i365-i373, 2009). This dynamic programming approach is extended here by a full-fledged computation of the base pairing probabilities. The $O(N^6)$ time and $O(N^4)$ space algorithm is implemented in C (available from `http://www.combinatorics.cn/cbpc/rip.html`) and is efficient enough to investigate for instance the interactions of small bacterial RNAs and their target mRNAs.

1 INTRODUCTION

RNA-RNA interactions constitute one of the fundamental mechanisms of cellular regulation. In an important subclass, small RNAs specifically bind a larger (m)RNA target. Examples include the regulation of translation in both prokaryotes (Narberhaus and Vogel, 2007) and eukaryotes (McManus and Sharp, 2002; Banerjee and Slack, 2002), the targeting of chemical modifications (Bachellerie *et al.*, 2002), and insertion editing (Benne, 1992), transcriptional control (Kugel and Goodrich, 2007). The common theme in many RNA classes, including miRNAs, siRNAs, snRNAs, gRNAs, and snoRNAs is the formation of RNA-RNA interaction structures that are more complex than simple sense-antisense interactions. The ability to predict the details of RNA-RNA interactions both in terms of the thermodynamics of binding in its structural consequences is a necessary prerequisite to understanding RNA based regulation mechanisms. The exact location of binding and the subsequent impact of the interaction on the structure of the target molecule can have profound biological consequences. In the case of sRNA-mRNA interactions, these details decide whether the sRNA is a positive or negative regulator of transcription depending on whether binding exposes or covers the Shine-Dalgarno sequence (Sharma *et al.*, 2007; Majdalani *et al.*, 2002). Similar effects have been observed using artificially designed opener and closer RNAs that regulate the binding of the *HuR* protein to human mRNAs (Meisner *et al.*, 2004; Hackermüller *et al.*, 2005).

In its most general form, the RNA-RNA interaction problem (RIP) is NP-complete (Alkan *et al.*, 2006; Mneimneh, 2007). The argument for this statement is based on an extension of the work of Akutsu (2000) for RNA folding with pseudoknots. Polynomial-time algorithms can be derived, however, by restricting the space of allowed configurations in ways that are similar to pseudoknot folding algorithms (Rivas and Eddy, 1999). The second major problem concerns the energy parameters since the standard loop types (hairpins, internal and multiloops) are insufficient; for the additional types, such as kissing hairpins, experimental data are virtually absent. Tertiary interactions, furthermore, are likely to have a significant impact.

Several restricted versions of RNA-RNA interaction have been considered in the literature. The simplest approach concatenates the two interacting sequences, essentially employing a slightly modified secondary structure folding algorithm. The algorithms `RNAcofold` (Hofacker *et al.*, 1994; Bernhart *et al.*, 2006), `pairfold` (Andronescu *et al.*, 2005), and `NUPACK` (Ren *et al.*, 2005) belong to this class. One major shortcoming of this approach is that it cannot predict important motifs such as kissing-hairpin loops. The paradigm of concatenation has also been generalized to the pseudoknot folding algorithm of Rivas and Eddy (1999).

*to whom correspondence should be addressed. Phone: *86-22-2350-6800; Fax: *86-22-2350-9272; `duck@santafe.edu`

The resulting model, however, still does not generate all relevant interaction structures (Chitsaz *et al.*, 2009; Qin and Reidys, 2008). An alternative approach is to neglect all internal base pairings in either strand and to compute the minimum free energy (mfe) secondary structure for their hybridization under this constraint. For instance, `RNAduplex` and `RNAhybrid` (Rehmsmeier *et al.*, 2004) follow this paradigm. `RNAup` (Mückstein *et al.*, 2006, 2008) and `intaRNA` (Busch *et al.*, 2008) restrict interactions to a single interval that remains unpaired in the secondary structure for each partner. These models have proved particularly useful for bacterial sRNA-mRNA interactions. Due to the highly conserved interaction motif, snoRNA-target interaction structures can be dealt with efficiently using specialized tools (Tafer *et al.*, 2009).

Pervouchine (2004) and Alkan *et al.* (2006) independently derived and implemented mfe folding algorithms for predicting the joint secondary structure of two interacting RNA molecules with polynomial time complexity. In their model, a "joint structure" means that the intramolecular structures of each molecule are pseudoknot-free, the intermolecular binding pairs are noncrossing and there exist no so-called "zigzags", see Fig. 1**(A)** and 2**(A)** for examples of the "joint structures". The optimal "joint structure" can be computed in $O(N^6)$ time and $O(N^4)$ space by means of dynamic programming.

Recently, Chitsaz *et al.* (2009) presented `piRNA`, a tool that uses dynamic programming algorithm to compute the partition function of "joint structures", also in $O(N^6)$ time. The algorithmic cores of the forward recursions of `piRNA` and `rip` were developed independently. Albeit differing in design details, they are equivalent. In addition, we identified here a basic data structure that forms the basis for computing additional important quantities such as the base pairing probability matrix, and probabilities of hybrid formations (see (Huang *et al.*, 2009) for the latter). Further differences between the two approaches will be discussed in Section 5.

The key innovation for passing from the mfe folding of Alkan *et al.* (2006) to the partition function is a unique grammar by which each interaction structure can be generated. Then, the computation of the partition function follows McCaskill's approach for RNA secondary structures (McCaskill, 1990). The key idea is to identify a certain subclass of interaction structures that serve as building blocks in a recursive decomposition generalizing the loop decomposition of secondary structures. These are the "tight structures", a generalization of the subsecondary structures enclosed by a unique closing pair.

In the following two sections we first derive a grammar that allows the unambiguous parsing of zigzag-free interaction structures, thus forming the basis for the computation of the partition function in $O(N^6)$ time and $O(N^4)$ memory, corresponding the mfe algorithm of Alkan *et al.* (2006). Then we proceed by deriving the recursions for the base pairing probabilities, which are based on a conceptual reversing of the production rules. Indeed, one has to compute the pairing probabilities by explicitly "tracing back" all contributing joint structures. The output of `rip` consists of the partition function, the base pairing probability matrix and the joint structure predicted by the maximal weighted (in terms of the base pair probabilities) matching (MWM) algorithm (Cary and Stormo, 1995; Gabow, 1973) and the most likely hybrid loops.

The *sodB-RhyB* interaction structure (Geissmann and Touati, 2004) is a well-known paradigmatic example with a unique interaction region, Fig. 1. The `rip` software predicts this interaction region correctly. Results obtained with other algorithms deviate noticeably from the known structure, see Supplemental Fig. S2.

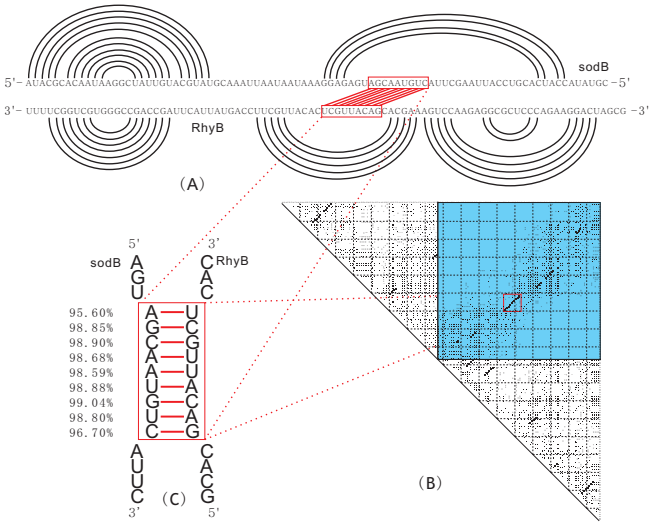

Fig. 1. (A) The natural structure of *sodB-RhyB* (Geissmann and Touati, 2004). (B) The base pairing probability matrix (McCaskill, 1990) generated via `rip`. This matrix represents all potential base pairs of the *sodB-RhyB* structure as squares, whose area is proportional to their respective probability. Intermolecular and intramolecular base pairs are depicted in the blue upper right rectangle and the two white triangles, respectively. (C) "Zoom" into the most likely interaction region as predicted by `rip`. All base pairs of the hybrid are labeled by their probabilities.

To-date, only a handful of interaction structures are known that are more complex than those covered by `intaRNA`/`RNAup`. The best-known example is the repression of *fhlA* by *OxyS* RNA, which involves two widely separated kissing-hairpin loops (Argaman and Altuvia, 2000). In Fig. 2, we display the natural interaction structure as well as the output of `rip`, which predicts two distinct interaction regions. The left (red) one coincides exactly with the published structure, while the right (blue) one differs by only two base pairs. We shall return to the *fhlA-OxyS* prediction in more detail in the Discussion section.

2 JOINT STRUCTURES

Given two RNA sequences R and S (e.g. an antisense RNA and its target) with N and M vertices, we index the vertices such that R_1 is the $5'$ end of R and S_1 denotes the $3'$ end of S. The edges of R and S represent the intramolecular base pairs. A *pre-structure*, $G(R, S, I)$, is a graph with the following properties:

1. R, S are secondary structures (each nucleotide being paired with at most one other nucleotide via hydrogen bonds, without internal pseudoknots);
2. I is a set of arcs of the form $R_i S_j$ without pseudoknots, i.e., if $R_{i_1} S_{j_1}$, $R_{i_2} S_{j_2} \in I$ where $i_1 < i_2$, then $j_1 < j_2$ holds.

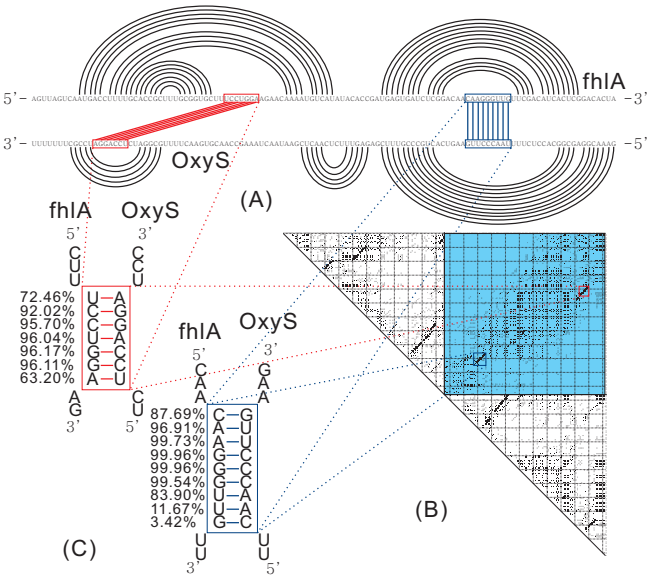

Fig. 2. (A) The natural structure of *f*hlA-*OxyS* (Chitsaz *et al.*, 2009). (B) The base pairing probability matrix see the caption of Fig. 1 for notation. (C) "Zoom" into the two distinct, most likely interaction regions, as predicted by `rip`.

An arc is called *exterior* if it is of the form R_iS_j and *interior*, otherwise. Let G be a graph and V be a subset of G-vertices. The *(induced) subgraph* of G induced by V has vertex set V and contains all G-edges having both incident vertices in V. In particular, we use $S[i,j]$ to denote the subgraph of the pre-structure $G(R,S,I)$ induced by $\{S_i, S_{i+1}, \ldots, S_j\}$, where $S[i,i] = S_i$ and $S[i,i-1] = \varnothing$. In absence of interactions a pre-structure is a pair of induced secondary structures on R and S, which we will refer to as a pair of *segments*. A segment $S[i_1,j_1]$ is called maximal if there is no segment, $S[i,j]$ strictly containing $S[i_1,j_1]$.

An interior arc $R_{i_1}R_{j_1}$ is an *R-ancestor* of the exterior arc R_iS_j if $i_1 < i < j_1$. Analogously, $S_{i_2}S_{j_2}$ is an S-ancestor of R_iS_j if $i_2 < j < j_2$. The sets of R-ancestors and S-ancestors of R_iS_j are denoted by $A_R(R_iS_j)$ and $A_S(R_iS_j)$, respectively. We will also refer to R_iS_j as a descendant of $R_{i_1}R_{j_1}$ and $S_{i_2}S_{j_2}$ in this situation. The R- and S-ancestors of R_iS_j with minimum arc-length are referred to as R- and S-parents, see Fig. 3, (A). Finally, we call $R_{i_1}R_{j_1}$ and $S_{i_2}S_{j_2}$ dependent if they have a common descendant and independent, otherwise.

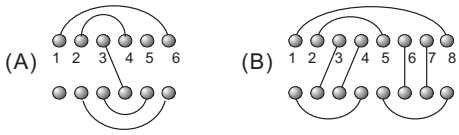

Fig. 3. (A) Ancestors and parents: for the exterior arc R_3S_4, we have the following ancestor sets $A_R(R_3S_4) = \{R_1R_6, R_2R_4\}$ and $A_S(R_3S_4) = \{S_2S_6, S_3S_5\}$. In particular, R_2R_4 and S_3S_5 are the R-parent and S-parent respectively. (B) Subsumed and equivalent arcs: R_1R_8 subsumes S_1S_4 and S_5S_8. Furthermore, R_2R_5 is equivalent to S_1S_4.

Suppose there is an exterior arc R_aS_b with ancestors R_iR_j and $S_{i'}S_{j'}$. Then R_iR_j is *subsumed* in $S_{i'}S_{j'}$, if for any $R_kS_{k'} \in I'$, $i < k < j$ implies $i' < k' < j'$, see Fig. 3, (B). If $R_{i_1}R_{j_1}$ is subsumed in $S_{i_2}S_{j_2}$ and *vice versa*, we call these arcs *equivalent*. A *zigzag*, is a subgraph containing two dependent interior arcs $R_{i_1}R_{j_1}$ and $S_{i_2}S_{j_2}$ neither one subsuming the other, see Fig. 4, **(A)**.

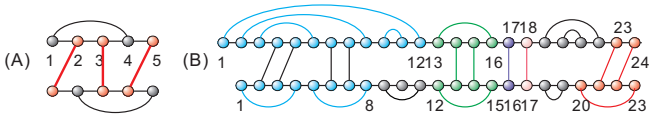

Fig. 4. (A): A zigzag, generated by R_2S_1, R_3S_3 and R_5S_4. (B): the joint structure $J_{1,24;1,23}$, we display the different segments and tight structures in which $J_{1,24;1,23}$ decomposes.

A *joint structure*, $J(R,S,I)$, is a zigzag-free pre-structure, see Fig. 4, **(B)**. Joint structures are exactly the configurations that are considered in the maximum matching approach of Pervouchine (2004), in the energy minimization algorithm of Alkan *et al.* (2006), and in the partition function approach of Chitsaz *et al.* (2009). The subgraph of a joint structure $J(R,S,I)$ induced by a pair of subsequences $\{R_i, R_{i+1}, \ldots, R_j\}$ and $\{S_h, S_{h+1}, \ldots, S_\ell\}$ is denoted by $J_{i,j;h,\ell}$. In particular, $J(R,S,I) = J_{1,N;1,M}$. We say $R_aR_b(S_aS_b, R_aS_b) \in J_{i,j;h,\ell}$ if and only if $R_aR_b(S_aS_b, R_aS_b)$ is an edge of the graph $J_{i,j;h,\ell}$. Furthermore, $J_{i,j;h,\ell} \subset J_{a,b;c,d}$ if and only if $J_{i,j;h,\ell}$ is a subgraph of $J_{a,b;c,d}$ induced by $\{R_i, \ldots, R_j\}$ and $\{S_h, \ldots, S_\ell\}$.

We next define a *tight* structure (ts). Given a joint structure, $J_{a,b;c,d}$, its tight $J_{a',b';c',d'}$ is either a single exterior arc $R_{a'}S_{c'}$ (in the case $a' = b'$ and $c' = d'$), or the minimal block centered around the leftmost and rightmost exterior arcs α_l, α_r, (possibly being equal) and an interior arc subsuming both, i.e., $J_{a',b';c',d'}$ is tight in $J_{a,b;c,d}$ if it has either an arc $R_{a'}R_{b'}$ or $S_{c'}S_{d'}$ if $a' \neq b'$ or $c' \neq d'$.

More formally, let $J_{a',b';c',d'}$ be contained in $J_{a,b;c,d}$ with rightmost and leftmost exterior arc R_iS_j and $R_{i_0}S_{j_0}$ and let M be the set of R_iS_j-ancestors in $J_{a,b;c,d}$ with maximal length. Then $J_{a',b';c',d'}$ is tight in $J_{a,b;c,d}$ if

1. for $M = \varnothing$: $J_{a',b';c',d'} = \{R_iS_j\}$;
2. for $M = \{R_{i_1}R_{j_1}\}$: $J_{a',b';c',d'} = J_{i_1,j_1;c',j}$, where c' is the origin (left) of the S-ancestor of $R_{i_0}S_{j_0}$ with maximal length (or i_0 if there is none). The case $M = \{S_{r_1}S_{s_1}\}$ is analogous;
3. for $M = \{R_{i_1}R_{j_1}, S_{r_1}S_{s_1}\}$, suppose $R_{i_1}R_{j_1}$ subsumes $S_{r_1}S_{s_1}$. Then $J_{a',b';c',d'} = J_{i_1,j_1;x_1,s_1}$, where x_1 is the origin of the S-ancestor of $R_{i_0}S_{j_0}$ with maximal length (or i_0 if there is none). In particular, $J_{a',b';c',d'} = J_{i_1,j_1;r_1,s_1}$ when $R_{i_1}R_{j_1}$ is equivalent with $S_{r_1}S_{s_1}$. The case, where $S_{r_1}S_{s_1}$ subsumes $R_{i_1}R_{j_1}$ is analogous.

In the following, a ts is denoted by $J^T_{i,j;h,\ell}$. If $J_{a',b';c',d'}$ is tight in $J_{a,b;c,d}$, then we call $J_{a,b;c,d}$ its envelope. By construction, the notion of ts is depending on its envelope. There are only four basic types of ts, see Fig. 5:

$\circ$: $\{R_iS_h\} = J^{\circ}_{i,j;h,\ell}$ and $i = j$, $h = \ell$;

$\bigtriangledown$: $R_iR_j \in J^{\bigtriangledown}_{i,j;h,\ell}$ and $S_hS_\ell \notin J^{\bigtriangledown}_{i,j;h,\ell}$;

$\square$: $\{R_iR_j, S_hS_\ell\} \in J^{\square}_{i,j;h,\ell}$;

$\bigtriangleup$: $S_hS_\ell \in J^{\bigtriangleup}_{i,j;h,\ell}$ and $R_iR_j \notin J^{\bigtriangleup}_{i,j;h,\ell}$.

In the Supplemental Material we prove:

PROPOSITION 2.1. *Let $J_{a,b;c,d}$ be a joint structure. Then*

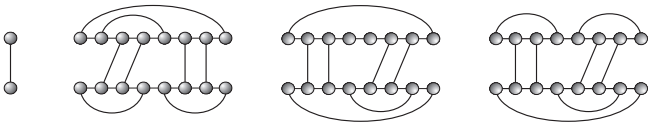

Fig. 5. From left to right: tights of type $\circ$, $\bigtriangledown$, $\square$ and $\bigtriangleup$.

1. *any exterior arc R_iS_j in $J_{a,b;c,d}$ is contained in a unique $J_{a,b;c,d}$-ts;*
2. *$J_{a,b;c,d}$ decomposes into a unique collection of $J_{a,b;c,d}$-ts and maximal segments.*

Given a ts, $J^{\bigtriangledown}_{i_0,j_0;r,s}$ (or $J^{\bigtriangleup}_{i,j;r_0,s_0}$), we introduce *double tight structures* as maximal substructure whose distinct leftmost and rightmost blocks are tights. By construction, therefore, each dts contains at least two tights.

More formally, given a ts $J^{\bigtriangledown}_{i_0,j_0;r,s}$, a *double-tight structure*, $J^{DT|\bigtriangledown}_{i,j;r,s}$, in $J^{\bigtriangledown}_{i_0,j_0;r,s}$, where $i_0 < i < j < j_0$, is defined as follows: there exists labels a, b, c, d where $i \le a < b \le j$ and $r \le c < d \le s$. Furthermore, two ts $J^{T}_{i,a;r,c}$ and $J^{T}_{b,j;d,s}$ in $J_{i_0+1,j_0-1;r,s}$ such that

$$J^{DT|\bigtriangledown}_{i,j;r,s} = J^{T}_{i,a;r,c} \dot\cup J_{a+1,b-1;c+1,d-1} \dot\cup J^{T}_{b,j;d,s}\,. \tag{2.1}$$

Here, the disjoint union $\dot\cup$ refers to both the vertex and arc sets of the joint structures, see Fig. 6. The case of a dts, $J^{DT|\bigtriangleup}_{i,j;r,s}$, within a ts, $J^{\bigtriangleup}_{i,j;r_0,s_0}$, is defined accordingly. By abuse of terminology, we simply use $J^{DT}_{i,j;r,s}$ in order to denote either $J^{DT|\bigtriangledown}_{i,j;r,s}$ or $J^{DT|\bigtriangleup}_{i,j;r,s}$.

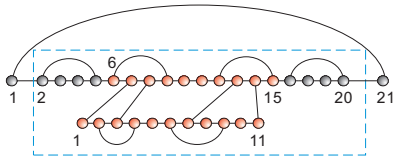

Fig. 6. A dts $J^{DT|\bigtriangledown}_{6,15;1,11}$ in $J_{2,20;1,11}$. Note that the joint structure $J_{1,21;1,11}$ itself is $\bigtriangledown$-tight. Here, $J_{2,15;1,11}$ is neither a ts nor a rts in $J_{2,20;1,11}$.

With the help of dts, as illustrated in Fig. 7, **Procedure (b)**, we decompose a ts as follows:
Let $J^{\bigtriangledown}_{i,j;r,s}$ be a ts of type $\bigtriangledown$ and let $R_{h_1}S_{\ell_1}$ and $R_{h_2}S_{\ell_2}$ be the leftmost and rightmost exterior arcs in $J_{i,j;r,s}$ and $i+1 \le i_1 \le j_1 \le j-1$. Then $J_{i+1,j-1;r,s}$ decomposes into

$$\begin{cases} R[i+1,i_1-1] \dot\cup J^{\{\bigtriangledown,\circ\}}_{i_1,j_1;r,s} \dot\cup R[j_1+1,j-1], \\ \qquad\qquad \text{if } J^{T}_{R_{h_1}S_{\ell_1}} = J^{T}_{R_{h_2}S_{\ell_2}}; \\ R[i+1,i_1-1] \dot\cup J^{DT}_{i_1,j_1;r,s} \dot\cup R[j_1+1,j-1], \\ \qquad\qquad \text{otherwise,} \end{cases} \tag{2.2}$$

where $J^{\{\bigtriangledown,\circ\}}_{i_1,j_1;r,s}$ denotes a $J_{i+1,j-1;r,s}$-ts of type $\bigtriangledown$ or $\circ$ and $J^{T}_{R_hS_\ell}$ denotes the unique ts in $J_{i+1,j-1;r,s}$ contain the exterior arc R_hS_ℓ.

Analogously, in case of a ts $J^{\bigtriangleup}_{i,j;r,s}$ with leftmost and rightmost exterior arcs $R_{h_1}S_{\ell_1}$ and $R_{h_2}S_{\ell_2}$, and $r+1 \le r_1 \le s_1 \le s-1$, $J_{i,j;r+1,s-1}$ can be decomposed in the form

$$\begin{cases} S[r+1,r_1-1] \dot\cup J^{\{\bigtriangleup,\circ\}}_{i,j;r_1,s_1} \dot\cup S[s_1+1,s-1], \\ \qquad\qquad \text{if } J^{T}_{R_{h_1}S_{\ell_1}} = J^{T}_{R_{h_2}S_{\ell_2}}; \\ S[r+1,r_1-1] \dot\cup J^{DT}_{i,j;r_1,s_1} \dot\cup S[s_1+1,s-1], \\ \qquad\qquad \text{otherwise,} \end{cases} \tag{2.3}$$

where $J^{\{\bigtriangleup,\circ\}}_{i_1,j_1;r,s}$ denotes a $J_{i,j;r+1,s-1}$-tight of type $\bigtriangleup$ or $\circ$.

For a ts $J^{\square}_{i,j;r,s}$ with $i+1 \le i_1 \le j_1 \le j-1$ we analogously derive

$$\begin{aligned} &J_{i+1,j-1;r,s} = \\ &R[i+1,i_1-1] \dot\cup J^{\{\bigtriangleup,\square\}}_{i_1,j_1;r,s} \dot\cup R[j_1+1,j-1], \end{aligned} \tag{2.4}$$

where $J^{\{\bigtriangleup,\square\}}_{i_1,j_1;r,s}$ denotes a $J_{i+1,j-1;r,s}$-tight of type $\bigtriangleup$ or $\square$.

Prop.(2.1) and equ. (2.1-2.4) establish, for each joint structure, a unique decomposition into interior and exterior arcs.

3 THE PARTITION FUNCTION

3.1 Refined Decomposition

The unique decomposition of ts would formally suffice to construct a partition function algorithm. Indeed, each decomposition step, such as equ. (2.1-2.4), corresponds to a multiplicative recursion relation for the partition function associated with the joint structures. However, this would result in an unwieldy expensive implementation. The reason are the multiple break points $a, b, c, d, \ldots$, each of which corresponding to a nested `for`-loop.

We therefore introduce a refined decomposition that reduces the number of break points. For this purpose we call a joint structure *right-tight* if its rightmost block is a ts. We adopt the point of view of Algebraic Dynamic Programming (Giegerich and Meyer, 2002) and regard each decomposition rule as a production in a suitable grammar. Fig. 7 summarizes two major steps in the decomposition: (I) the "arc-removal" reducing ts and dts. The scheme is complemented by the usual loop decomposition of secondary structures, and (II) the "block-decomposition" splitting joint structures into blocks.

The details of the decomposition procedures are collected in the SM, where we show that for each $J_{1,N;1,M}$ there exists a unique decomposition-tree (parse-tree), denoted by $T_{J_{1,N;1,M}}$. This tree has root $J_{1,N;1,M}$ and all other vertices correspond to specific substructures of $J_{1,N;1,M}$ obtained by the successive application of the decomposition steps of Fig. 7 and the loop decomposition of the secondary structures, see Fig. 8.

3.2 Extended Loop Model

The standard energy model for RNA folding (Mathews *et al.*, 1999), presented in the SM, is consistent with the basic decomposition of secondary structures. In addition, joint structures give rise to two further types of loops. Following Chitsaz *et al.* (2009), we call them *hybrid* and *kissing-loop*, Fig. 9.

- A *hybrid* is a maximal sequence of intermolecular interior loops formed by $\ell \ge 2$ exterior arcs $R_{i_1}S_{j_1}, \ldots, R_{i_\ell}S_{j_\ell}$ where $R_{i_h}S_{j_h}$ is nested within $R_{i_{h+1}}S_{j_{h+1}}$ and where the internal segments $R[i_h+1, i_{h+1}-1]$ and $S[j_h+1, j_{h+1}-1]$ consist of single-stranded nucleotides. I.e., a hybrid is the maximal unbranched stem-loop formed by external arcs.
- A *kissing-loop* is either a pair, $(R_iR_j, R[i+1,j-1])$, where the set of R_iR_j-children,, $R_{i_1}S_{j_1}, \ldots$ where $i < i_1 < j$ is nonempty, or a pair $(S_iS_j, S[i+1,j-1])$, where the set of S_iS_j-children $R_{i_1}S_{j_1}, \ldots$ where $i < j_1 < j$ is nonempty.

Kissing loops have been singled out for logical reasons and because some investigations into their thermodynamic properties have been reported in the literature (Gago *et al.*, 2005). For details of the parametrization employed in `rip` we refer to the SM.

Let us now have a closer look at the energy evaluation of $J_{i,j;h,\ell}$. Each decomposition step in Fig. 7 results in substructures whose energies we assume to contribute additively and generalized loops that need to be evaluated directly. There are the following two scenarios:

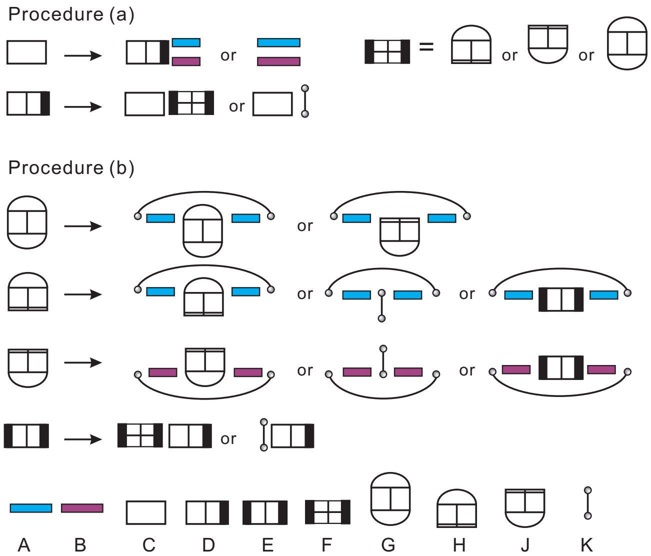

Fig. 7. Illustration of Procedure (a) the reduction of arbitrary joint structures and rts, and Procedure (b) the decomposition of tight structures. The panel below indicates the 10 different types of structural components: **A**, **B**: maximal secondary structure segments $R[i, j]$, $S[r, s]$; **C**: arbitrary joint structure $J_{i,j;r,s}$; **D**: right-tight structures $J^{RT}_{i,j;r,s}$; **E**: double-tight structure $J^{DT}_{i,j;r,s}$; **F**: tight structure of type $\triangledown$, $\triangle$ or $\square$; **G**: type $\square$ tight $J^{\square}_{i,j;r,s}$, the solid curved line (top and bottom) denotes an arc and a single horizontal line (top and bottom) denotes the backbone; **H**: type $\triangledown$ tight $J^{\triangledown}_{i,j;r,s}$, a solid curved line (top) denotes an arc, a single horizontal line (top) denotes the backbone and a double-horizontal line (bottom) denotes that the two terminals are not paired wit each other; **J**: type $\triangle$ tight $J^{\triangle}_{i,j;r,s}$, a solid curved line (bottom) denotes an arc, a single horizontal line (bottom) denotes the backbone and a double-horizontal line (top) indicates that the two terminals are not paired with each other; **K**: an exterior arc.

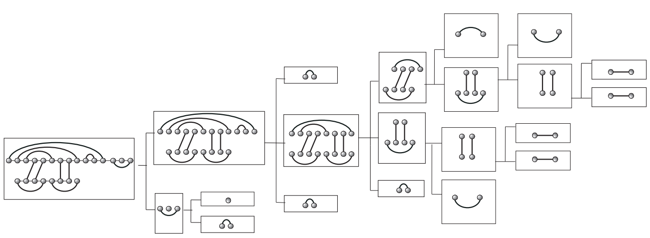

Fig. 8. The decomposition tree $T_{J_{1,15;1,8}}$ for the joint structure $J_{1,15;1,8}$.

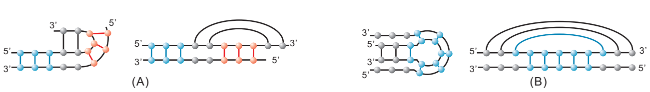

Fig. 9. The two new loop-types: hybrid **(A)** and kissing-loop **(B)**.

I. Arc removal. Most of the decomposition operations in Procedure (b) displayed in Fig. 7 can be viewed as the "removal" of an arc (corresponding to the closing pair of a loop in secondary structure folding) followed by decomposition. Both, loop-type as well as the subsequent decomposition steps depend on the newly exposed structural elements. Following the approach of Zuker and Stiegler (1981) for secondary structures, we treat the loop-decomposition problem by introducing additional matrices. Without loss of generality, we can assume that we open an interior base pair R_iR_j.

The set of base pairs on $R[i, j]$ consists of all interior pairs R_pR_q with $i \le p < q \le j$ and all exterior pairs R_pS_h with $i \le p \le j$. An interior arc is *exposed* on $R[i+1, j-1]$ if and only if it is not enclosed by any interior arc in $R[i, j]$. An exterior arc is *exposed* on $R[i+1, j-1]$ if and only if it is not a descendant of any interior arc in $R[i+1, j-1]$. Given R_{ij}, the arcs exposed on $R[i+1, j-1]$ corresponds to the base pairs *immediately interior* of R_iR_j. Let us write $E_{R[i,j]} = E^i_{R[i,j]} \dot\cup E^e_{R[i,j]}$ for this set of "exposed base pairs" and its subsets of interior and exterior arcs. As in secondary structure folding, the loop type is determined by $E_{R[i,j]} := E_R$ as follows: $E_R = \varnothing$, hairpin loop; $E_R = E^i_R$ and $|E_R| = 1$, interior loop (including bulge and stacks); $E_R = E^i_R$, $|E_R| \ge 2$, multi-branch loop; $E_R = E^e_R$, kissing-hairpin loop; $|E^i_R|, |E^e_R| \ge 1$, general kissing-loop.

This picture needs to be refined even further since the arc removal is coupled with further decomposition of the interval $R[i+1, j-1]$. This prompts us to distinguish ts and dts with different classes of exposed base pairs on one or both strands. It will be convenient, furthermore to include information on the type of loop in which it was found.

A ts $J^{\triangledown}_{i,j;h,\ell}$ is of type E, if $S[h, \ell]$ is not enclosed in any base pair ($J^{\triangledown,\mathsf{E}}_{i,j;h,\ell}$). Suppose $J^{\triangledown}_{i,j;h,\ell}$ is located immediately interior to the closing pair S_pS_q ($p < h < \ell < q$). If the loop closed by S_pS_q is a multiloop, then $J^{\triangledown}_{i,j;h,\ell}$ is of type M ($J^{\triangledown,\mathsf{M}}_{i,j;h,\ell}$). If S_pS_q is contained in a kissing-loop, we distinguish the types F and K, depending on whether or not $E^e_{S[h,\ell]} = \varnothing$. Fig. 10 displays this decomposition for $J^{\triangledown,\mathsf{M}}_{i,j;r,s}$.

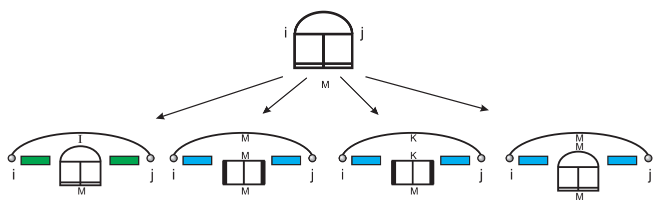

Fig. 10. Further refinement: the four decompositions of $J^{\triangledown,\mathsf{M}}_{i,j;r,s}$ via Procedure (b). The green rectangle denotes single-stranded segments. The letters I, M, etc denote the loop-type and the type of the exposed arc(s) of the dts. See Fig. 7 for more details on the notation. The four cases correspond to the four contributions in equ. (3.1).

For a dts $J^{DT}_{p,q,r,s}$ (denoted by "E" in Fig. 7) we need to determine the type of the exposed pairs of both $R[p, q]$ and $S[r, s]$. Hence each such structure will be indexed by two types. In total, we arrive at 18 distinct cases since some combinations cannot occur. For instance, a dts cannot be external in both R and S, that is, type EE does not exist, where E means external.

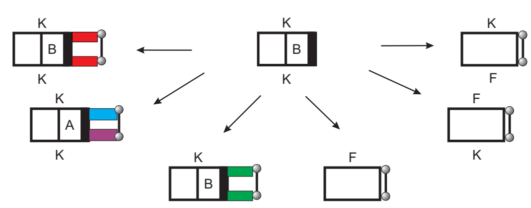

Fig. 11. Decomposition of $J^{RT,\mathsf{KKB}}_{i,j;h,\ell}$ by means of procedure (b). Here the red rectangle denotes a pair of secondary segments having the property that at least one of them is not single-stranded.

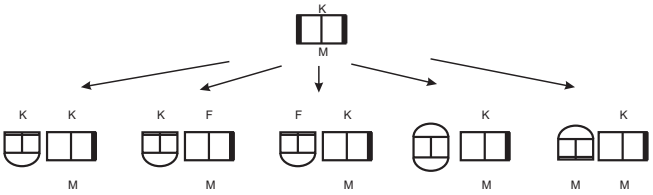

Fig. 12. Decomposition of $J^{DT,\mathsf{KM}}_{i,j;h,\ell}$ by means of procedure (b). The five alternatives correspond to the three additive terms in equ. (3.2).

II. Block decomposition. The second type of decomposition is the splitting of joint structures into "blocks", such as the decompositions of a rts in Procedure (a) and a dts in Procedure (b), see Fig. 7. A rts $J^{RT}_{i,j;h,\ell}$ may appear in two ways, depending on whether or not there exists an exterior arc $R_{i_1}S_{j_1}$ such that $R[i_1+1,j]$ and $S[j_1+1,\ell]$ are single-stranded segments. If such an exterior arc exists, $J^{RT}_{i,j;h,\ell}$ is of type (rB), otherwise it is of type (rA). Analogously, (ℓB) and (ℓA) are defined for dts. Fig. 11 shows the decomposition of $J^{RT,\mathsf{KKB}}_{i,j;h,\ell}$. Suppose $J^{DT}_{i,j;r,\ell}$ is a dts contained in a kissing-loop, that is we have either $E^e_{R[i,j]} \neq \varnothing$ or $E^e_{S[h,\ell]} \neq \varnothing$. W.l.o.g., we may assume $E^e_{R[i,j]} \neq \varnothing$. Then at least one of the two "blocks" contains the exterior arc belonging to $E^e_{R[i,j]}$ (i.e., direct bonds in the language of Chitsaz *et al.* (2009)) labeled by K. Otherwise the block is labeled F, see Fig. 12. The situation is analogous if we decompose $J^{RT}_{i,j;r,\ell}$ contained in a kissing-loop.

3.3 Recursions

The computation of the partition function is obtained "from the inside to the outside", see equs. (3.1,3.2). The recursions are initialized with the energies of individual external base pairs and empty secondary structures on subsequences of length up to four. In order to differentiate multi- and kissing-loop contributions, we introduce the partition functions $Q^{\mathsf{m}}_{i,j}$ and $Q^{\mathsf{F}}_{i,j}$ as a generalization of Zuker's algorithm (Zuker and Stiegler, 1981). Here, $Q^{\mathsf{m}}_{i,j}$ denotes the partition function of secondary structures on $R[i,j]$ or $S[i,j]$ having at least one arc contained in a multi-loop. Similarly, $Q^{\mathsf{F}}_{i,j}$ denotes the partition function of secondary structures on $R[i,j]$ or $S[i,j]$ in which at least one arc is contained in a kissing-loop.

For instance, the recursion for $Q^{\triangledown,\mathsf{M}}_{i,j;r,s}$ in Fig. 10 reads:

$$
\begin{aligned}
Q^{\triangledown,\mathsf{M}}_{i,j;r,s} = \sum_{h,\ell}\Big\{ & Q^{\triangledown,\mathsf{M}}_{h,\ell;r,s} e^{-G^{\mathsf{Int}}_{i,j;h,\ell}/kT} \\
& + Q^{DT,\mathsf{MM}}_{h,\ell;r,s} e^{-(\alpha_1+\alpha_2)/kT} \times (e^{-(h-i-1)\alpha_3/kT} + Q^{\mathsf{m}}_{i+1,h-1}) \\
& \quad \times (e^{-(j-\ell-1)\alpha_3/kT} + Q^{\mathsf{m}}_{\ell+1,j-1}), \\
& + Q^{DT,\mathsf{KM}}_{h,\ell;r,s} e^{-(\beta_1+\beta_2)/kT} \times (e^{-(h-i-1)\beta_3/kT} + Q^{\mathsf{F}}_{i+1,h-1}) \\
& \quad \times (e^{-(j-\ell-1)\beta_3/kT} + Q^{\mathsf{F}}_{\ell+1,j-1}), \\
& + Q^{\triangledown,\mathsf{M}}_{h,\ell;r,s} e^{-(\alpha_1+2\alpha_2)/kT} \big[e^{-(j-\ell-1)\alpha_3/kT} Q^{\mathsf{m}}_{i+1,h-1} \\
& \quad + e^{-(h-i-1)\alpha_3/kT} Q^{\mathsf{m}}_{\ell+1,j-1} + Q^{\mathsf{m}}_{\ell+1,j-1} Q^{\mathsf{m}}_{i+1,h-1} \big] \Big\}.
\end{aligned}
\tag{3.1}
$$

Analogously, the recursion for the dts $Q^{DT,\mathsf{KM}}_{i,j;r,s}$ of Fig.12 is given by

$$
\begin{aligned}
Q^{DT,\mathsf{KM}}_{i,j;r,s} = \sum_{i_1,j_1}\Big\{ & (Q^{\triangledown,\mathsf{M}}_{i,i_1;r,j_1} e^{-\beta_2/kT} + Q^{\triangle,\mathsf{K}}_{i,i_1;r,j_1} e^{-\alpha_2/kT} \\
& + Q^{\square}_{i,i_1;r,j_1} e^{-(\alpha_2+\beta_2)/kT} + Q^{\triangle,\mathsf{F}}_{i,i_1;r,j_1} e^{-\alpha_2/kT}) Q^{RT,\mathsf{KM}}_{i_1+1,j;j_1+1,s} \\
& + Q^{\triangle,\mathsf{K}}_{i,i_1;r,j_1} e^{-\alpha_2/kT} Q^{RT,\mathsf{FM}}_{i_1+1,j;j_1+1,s} \Big\}.
\end{aligned}
\tag{3.2}
$$

4 BASE PAIRING PROBABILITIES

Given two RNA sequences, our sample space is the ensemble of all zigzag-free joint interaction structures. Let Q^I denote the corresponding partition function. The probability of a joint structure $J_{1,N;1,M}$ is then given by $\mathbb{P}_{J_{1,N;1,M}} = Q_{J_{1,N;1,M}}/Q^I$.

4.1 Approach

While the computation of the partition function proceeds from smaller to larger subsequences, the computation of the substructure probabilities follows the order of the decomposition outlined in the previous section. That is, the longest-range substructures are computed first, analogous to McCaskill's algorithm for secondary structures (McCaskill, 1990).

Let $\mathbb{J}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell}$ be the set of substructures $J_{i,j;h,\ell} \subset J_{1,N;1,M}$ such that $J_{i,j;h,\ell}$ appears in $T_{J_{1,N;1,M}}$ as an interaction structure of type $\xi \in \{DT, RT, \triangledown, \triangle, \square, \circ\}$ with loop-subtypes $Y_1, Y_2 \in \{\mathsf{M},\mathsf{K},\mathsf{F}\}$ on the sub-intervals $R[i,j]$ and $S[h,\ell]$, $Y_3 \in \{\mathsf{A},\mathsf{B}\}$. Let $\mathbb{P}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell}$ be the probability of $\mathbb{J}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell}$. For instance, $\mathbb{P}^{RT,\mathsf{MKA}}_{i,j;h,\ell}$ is the sum over all the probabilities of substructures $J_{i,j;h,\ell} \in T_{J_{1,N;1,M}}$ such that $J_{i,j;h,\ell}$ is a rts of type $r\mathsf{A}$ and $R[i,j]$, $S[h,\ell]$ are enclosed by a multi-loop and kissing-loop, respectively. Then the computation of the pairing probabilities reduces to a trace-back routine in the decomposition tree constructed in Section 3.1.

Set $J = J_{1,N;1,M}$, $T = T_{J_{1,N;1,M}}$ and let $\Lambda_{J_{i,j;h,\ell}} = \{J | J_{i,j;h,\ell} \in T\}$ denote the set of all joint structures J such that $J_{i,j;h,\ell}$ is a vertex in the decomposition tree T. Then we have $\mathbb{P}_{J_{i,j;h,\ell}} = \sum_{J\in\Lambda_{i,j;h,\ell}} \mathbb{P}_J$ and furthermore

$$
\mathbb{P}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell} = \sum_{J_{i,j;h,\ell}\in\mathbb{J}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell}} \mathbb{P}_{i,j;h,\ell}. \tag{4.1}
$$

4.2 Case Study: Secondary Structures

In order to illustrate the logic of our backtracking procedure, we first consider the special case of secondary structures.

Let $\mathbb{P}_{R_iR_j}$ denote the base pairing binding probability of R_iR_j, i.e. $\mathbb{P}_{R_iR_j} = \sum_{R_iR_j\in W} Q_W Q^{-1}$, where the sums is taken over all the partition functions of secondary structures W in R such that $R_iR_j \in W$. Let T_W be the decomposition tree of a particular secondary structure W on $R[1,N]$ via Procedure (c), the key observation here is

$$
R_iR_j \in W \Longleftrightarrow R_iR_j \in T_W. \tag{4.2}
$$

Let $\Omega(R_iR_j)$ be the set of secondary structures whose decomposition tree contain the pair R_iR_j as a leaf. Clearly, via equ. (4.2), we obtain

$$
\mathbb{P}_{R_iR_j} = \sum_{W\in\Omega(R_iR_j)} Q_W Q^{-1}. \tag{4.3}
$$

In order to compute $\mathbb{P}_{R_iR_j}$, we express it as a sum over the probabilities of all substructures ξ, that are a parent of R_iR_j in the decomposition tree. Let $R^b(i,j)$ denote the set of secondary segments $R[i,j]$ in which R_i is connected with R_j and let $\mathbb{P}^b_{R_i,R_j}$ be its probability. By construction, we have $\mathbb{P}_{R_iR_j} = \mathbb{P}^b_{R_i,R_j}$, since the parent of R_iR_j in the decomposition tree must be a secondary segment $R[i,j]$ such that $R_iR_j \in R[i,j]$. Therefore

the computation of $\mathbb{P}_{R_iR_j}$ is reduced to the calculation of the substructure probability $\mathbb{P}^b_{R_i,R_j}$.

The decomposition is summarized in Procedure (c), Fig. 13. In view of the fact that the `rip` has $O(N^6)$ time complexity, we can differ here from the standard implementation of the RNA folding model. Inspection of Fig. 13 shows that for a $R^b(i,j)$-parent we have to distinguish the five cases displayed in the lower panel. Let $R^{\mathsf{m}}(i,j)$ be the set of segments $R[i,j] \in T_{R[1,N]}$ containing at least one arc with an outer loop of type M, and let $R^s(i,j)$ be the set of all segments $R[i,j] \in T_{R[1,N]}$. Furthermore, let $\mathbb{P}^{\mathsf{m}}_{R_i,R_j}$ and $\mathbb{P}^s_{R_i,R_j}$ be the corresponding probabilities. Note that it is possible, for (L1) and (L4) in Fig. 13, that $h = i$ and $j = \ell$ holds. However, via further backtracking for $R^s(i,j)$ and $R^m(i,j)$ we can recursively calculate the binding probability.

Following the logic of Fig. 13, we obtain

$$\begin{aligned}\mathbb{P}^b_{R_i,R_j} = \sum_{h,\ell}\Bigg\{&\mathbb{P}^s_{R_h,R_\ell}\frac{Q^s_{h,i-1}Q^b_{i,j}}{Q^s_{h,\ell}} + \mathbb{P}^b_{R_h,R_\ell}\frac{Q^b_{i,j}e^{-G^{\mathsf{Int}}_{h,\ell;i,j}/kT}}{Q^b_{h,\ell}}\\ &+\mathbb{P}^b_{R_h,R_\ell}\frac{Q^{\mathsf{m}}_{h+1,i-1}Q^b_{i,j}e^{-(\alpha_1+2\alpha_2+(\ell-j-1)\alpha_3)/kT}}{Q^b_{h,\ell}}\\ &+\mathbb{P}^{\mathsf{m}}_{R_h,R_\ell}\frac{Q^b_{i,j}e^{-(\alpha_2+(i-h+\ell-j)\alpha_3)/kT}}{Q^{\mathsf{m}}_{h,\ell}}\\ &+\mathbb{P}^{\mathsf{m}}_{R_h,R_\ell}\frac{Q^{\mathsf{m}}_{h,i-1}Q^b_{i,j}e^{-(\alpha_2+(\ell-j)\alpha_3)/kT}}{Q^{\mathsf{m}}_{h,\ell}}\Bigg\}.\end{aligned} \tag{4.4}$$

where the lines correspond to the five loop types (L1-L5) in Fig. 13. Analogously, the recursions for the base pairing probabilities $\mathbb{P}^{\mathsf{m}}_{R_i,R_j}$ and $\mathbb{P}^s_{R_i,R_j}$ are given by

$$\begin{aligned}\mathbb{P}^{\mathsf{m}}_{R_i,R_j} = \sum_{h,\ell}\Bigg\{&\mathbb{P}^b_{R_{i-1},R_\ell}e^{-(\alpha_1+2\alpha_2+(\ell-1-h)\alpha_3)/kT}\\ &\times\frac{Q^b_{j+1,h}Q^{\mathsf{m}}_{i,j}}{Q^b_{i-1,\ell}} + \mathbb{P}^{\mathsf{m}}_{R_i,R_\ell}\frac{Q^{\mathsf{m}}_{i,j}Q^b_{j+1,h}e^{-(\alpha_2+(\ell-h)\alpha_3)/kT}}{Q^{\mathsf{m}}_{i,\ell}}\Bigg\}\\ \mathbb{P}^s_{R_i,R_j} = \sum_{h,\ell}&\mathbb{P}^s_{R_i,R_\ell}\frac{Q^s_{i,j}Q^b_{j+1,h}}{Q^s_{i,\ell}}.\end{aligned} \tag{4.5}$$

4.3 Base pairing probabilities for joint structures

Set $\Sigma_1 = \{J \mid R_iR_j \in J\}$. We apply the same strategy to the joint structures appearing in Fig. 7. Let Q^I denote the partition function which sums over all the possible joint structures $J_{1,N;1,M}$. Then $\mathbb{P}_{R_i,R_j} = \sum_{J\in\Sigma_1} Q_J/Q^I$. In order to compute $\mathbb{P}_{R_iR_j}$ we classify Σ_1 according to the parent of R_iR_j in T:

$$\begin{aligned}\Sigma_1 = &\{J \mid R[i,j] \in T, R[i,j] \in R^b(i,j)\}\\ &\cup\bigcup_{h,\ell}\{J \mid J_{i,j;h,\ell} \in T, J_{i,j;h,\ell} \in \mathbb{J}^{\triangledown}_{i,j;h,\ell}\}\\ &\cup\bigcup_{h,\ell}\{J \mid J_{i,j;h,\ell} \in T, J_{i,j;h,\ell} \in \mathbb{J}^{\square}_{i,j;h,\ell}\},\end{aligned} \tag{4.6}$$

which translates to

$$\mathbb{P}_{R_iR_j} = \mathbb{P}^b_{R_i,R_j} + \sum_{h,\ell}\mathbb{P}^{\triangledown,\{\mathsf{E},\mathsf{M},\mathsf{F},\mathsf{K}\}}_{i,j;h,\ell} + \sum_{h,\ell}\mathbb{P}^{\square}_{i,j;h,\ell}, \tag{4.7}$$

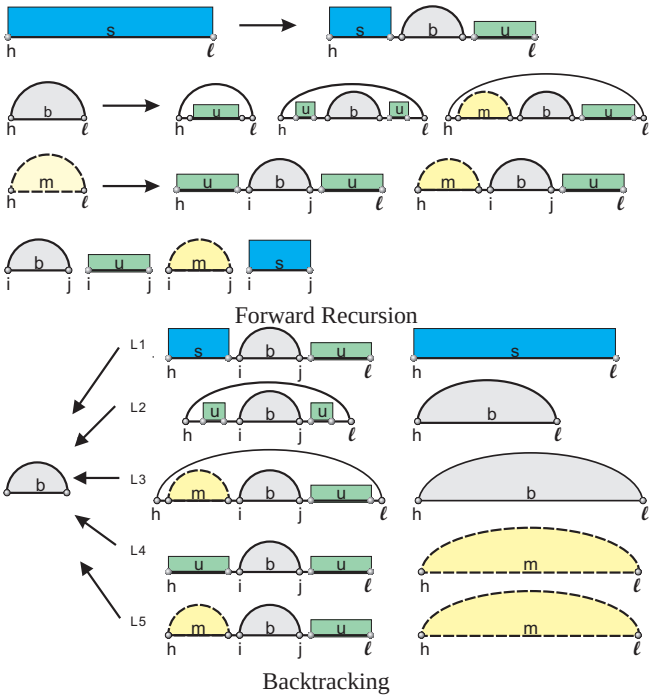

Fig. 13. Top. Extended version of Procedure (c) showing the productions for general structures, structures and enclosed by pairs, and multi-loops. The four types of segments are shown below: In $R^b(i,j)$, R_i, R_j is paired, $R^u(i,j)$ denotes stretches of unpaired bases, $R^{\mathsf{m}}(i,j)$ denotes parts of multiloops with containing least one arc, and $R^s(i,j)$ denotes arbitrary segments. **Below.** Backtracking for secondary structures: for a parent of $R^b(i,j)$ we have five cases according to Procedure (c): external (L1), interior loop (L2), closing pair of a multi-loop (L3), (L4) and (L5) denote the scenarios arising from decomposing a $R^m(h,\ell)$-segment. See equ. (4.4) for the corresponding recursions.

where $\mathbb{P}^{\triangledown,\{\mathsf{E},\mathsf{M},\mathsf{F},\mathsf{K}\}}_{i,j;h,\ell} = \mathbb{P}^{\triangledown,\mathsf{E}} + \mathbb{P}^{\triangledown,\mathsf{M}} + \mathbb{P}^{\triangledown,\mathsf{F}} + \mathbb{P}^{\triangledown,\mathsf{K}}$ for identical positions i, j, h, ℓ. Analogously, we obtain for pairs in S:

$$\begin{aligned}\Sigma_2 = &\{J \mid S[h,\ell] \in T, S[h,\ell] \in S^b[h,\ell]\}\\ &\cup\bigcup_{i,j}\{J \mid J_{i,j;h,\ell} \in T, J_{i,j;h,\ell} \in \mathbb{J}^{\triangle}_{i,j;h,\ell}\},\end{aligned} \tag{4.8}$$

and therefore $\mathbb{P}_{S_iS_j} = \mathbb{P}^b_{S_i,S_j} + \sum_{h,\ell}\mathbb{P}^{\triangle}_{h,\ell;i,j}$, with $\mathbb{P}^{\triangle} = \mathbb{P}^{\triangle,\mathsf{E}} + \mathbb{P}^{\triangle,\mathsf{M}} + \mathbb{P}^{\triangle,\mathsf{K}} + \mathbb{P}^{\triangle,\mathsf{F}}$.

Note that the expressions for $\mathbb{P}_{R_iR_j}$ and $\mathbb{P}_{S_iS_j}$ are not symmetric. This is due to the fact that our decomposition routine give preference to arc-removals in R over those in S. This asymmetry is necessary to ensure that the decomposition in Fig. 7 is unambiguous.

Finally, we calculate the binding probability of an exterior arc R_iS_j. Since R_iS_j is a ts of type $\circ$, $\mathbb{P}_{R_iS_j}$ is directly given by the probability of this special substructure in equ. (4.1).

In order to compute the binding probabilities of both interior and exterior arcs, the key is to employ an "inverse" grammar induced by tracing back in the decomposition tree as displayed in Supplement Materials [SM, Fig. 5].

5 RESULTS AND DISCUSSION

In this contribution we have introduced a framework in which both the partition function and the base pairing probabilities of zigzag-free RNA-RNA interactions can be derived in a natural

way. Our approach is implemented in the software package `rip` using the full standard energy model for RNA secondary structures together with a multi-loop-like additive parametrization for kissing-loops. In comparision with `piRNA` (Chitsaz *et al.*, 2009), `rip` is based on a different but equivalent decomposition grammar. The very encouraging data on the accuracy of the predicted interaction intergies reported by Chitsaz *et al.* (2009) therefore carry over to `rip` as well.

The notion of tights, which have a central role in our presentation, is also implicit in the work of Chitsaz *et al.* (2009). The focus on the underlying combinatorial aspects, however, leads us to highlight in particular the decomposition tree, which provides a natural framework in which to proceed beyond the algorithmic core of the partition function itself. Indeed, the decomposition tree facilitates the derivation of the base pairing probabilities. Related questions, such as that for the probability of complete hybrids (Huang *et al.*, 2009) can be answered along the same lines. While the current implmenentation of `piRNA` Chitsaz *et al.* (2009) concentrates on melting temperature in order to validate the partition function, `rip` focusses on a detailed analysis of the interaction structures themselves. To this end, we also compute the unweighted maximum expected accuracy structure, which is given as the maximum matching with weights given by the base pairing probabilites. On a more technical level, `piRNA` and `rip` differ in the decomposition of tights: `piRNA` utilizes a $4D$ gap-matrix by means of the Dirks-Pierce algorithm Dirks and Pierce (2003), while `rip` employs two distinct $2D$-matrices inspired by Zuker's recursion. For details, we refer to the SM.

Back-tracing of the base pairing patterns that underlie the free energy of RNA-RNA binding is of great importance in detailed studies of ncRNA-mRNA interactions. The details of the binding sites have a crucial impact on the interpretation of the computational results and on the comparison of the computational prediction and experimental data. It was shown by Mückstein *et al.* (2008), for instance, that positive and negative regulation of bacterial mRNAs can be distinguished depending on whether the interaction structure contains the Shine-Dalgarno sequence in stable stem or exposed in an predominantly unpaired region.

Only a small number of interaction structures have been described so far that are more complex than those computable by `RNAup`/`intaRNA`. It is not clear, however, whether complex interactions are truely rare in nature, or whether multi-point contacts such as that of the *fhlA-OxyS* interaction structure (Argaman and Altuvia, 2000) are rarely observed experimentally because they are typically excluded from candidate lists due to the lack of readily detectable pairing regions. A survey with `rip` may be suitable to provide us with a much more unbiased picture. Fig. 2 shows, that, modulo two base pairs, `rip` identifies the two distinct hybrids in *fhlA-OxyS*, correctly. Table 1 furthermore establishes, that the latter are indeed uniquely identified.

Fig. 14 compares the output of `rip` with several other, established, folding algorithms.

Following Chitsaz *et al.* (2009), we stipulate an independent initialization energy, σ_0, for each hybrid, and scaled energies for its base pairs. Many other RNA-cofolding algorithms, like `RNAcofold` (Bernhart *et al.*, 2006) and Dimitrov and Zuker (2004) assume a single initialization energy, ε. This energy model can *a posteriori* be derived from `rip`, once the partition function of the joint structures Q^I and the partition functions Q_R and Q_S

I	II	III	IV
58,47: 50.3%	21,69: 56.5%	34,55: 45.7%	83,45: 17.3%
59,46: 54.8%	20,70: 59.5%	35,54: 51.0%	82,46: 18.9%
60,45: 52.9%	19,71: 29.8%	36,53: 49.9%	81,47: 18.5%
61,44: 28.8%			

Table 1. The base pairing probabilities of the four alternative hybrids I, II, III and IV for *fhlA-OxyS*, predicted by `rip`, see Fig. 2. Each entry represents the positions in *fhlA*(R)-*OxyS*(S) and the base pair probability. For instance, $58,47:50.3\%$ is equivalent to $\mathbb{P}_{R_{58}S_{47}} = 0.503$.

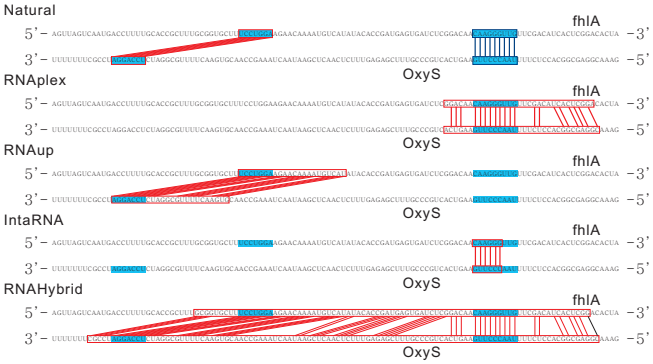

Fig. 14. The natural structure of the *fhlA-OxyS* interaction and the results predicted by several algorithms including `RNAplex`.

have been computed. Let Ω_1 be the set of all joint structures having at least one external arc and denote by Ω_0 the set of all structures that have none. Then $Q^{\mathtt{rip}} = Q(\Omega_1) + Q(\Omega_0)$, where $Q(\Omega_0) = Q_R Q_S$. Taking the initiation term into account, we compute $Q = Q(\Omega_1)\exp(-\varepsilon/kT) + Q_R Q_S$, from which we obtain the corrected value for $Q(\Omega_1)$. As shown by Bernhart *et al.* (2006), the base pairing probabilities can be rescaled via

$$\mathbb{P}^{\epsilon}_{ij} = \frac{\left[\mathbb{P}^{\mathtt{rip}}_{ij} Q^{\mathtt{rip}} - \mathbb{P}_{ij}(\Omega_0) Q_R Q_S\right] e^{-\varepsilon/kT} + \mathbb{P}_{ij}(\Omega_0) Q_R Q_S}{[Q^{\mathtt{rip}} - Q_R Q_S] e^{-\varepsilon/kT} + Q_R Q_S.} \tag{5.1}$$

We have focussed here on the algorithmic context for computing detailed models of RNA-RNA interactions in the most general framework that is computationally feasible at the moment. The current implementation of `rip` may, due to the computational costs incurred by several dozens of interdependent 4-dimensional arrays, be viewed "just" as a reference. However, in all computed examples, `rip` quite accurately reproduced the interaction regions. We are here in a similar position as with the Sankoff algorithm (which addresses the closely related dynamic programming problem of simultaneous alignment and structure prediction). While the full implementations are slow and of limited use in particular in large-scale studies, they are instrumental in optimizing the procedure and in devising efficient nearly exact pruning heuristics that can dramatically reduce the fraction of array entries that need to be computed (Havgaard *et al.*, 2007).

The constructions presented here give rise to several variations. Point in case being the computation of hybrid probabilities, i.e., the probabilities $\mathbb{P}^{\mathsf{Hy}}_{i,j;h,\ell}$ that $R[i,j]$ and $S[h,\ell]$ form an "interaction

stem" or a even an entire uninterrupted interaction region Huang *et al.* (2009). Another line of research concerns improved energy models for more complex types of loops Isambert and Siggia (2000).

The algorithmic approach taken here was motivated by a combinatorial analysis of zigzag-free interaction structures. From a mathematical point of view, our approach is centered around the notions of tight structures and their decomposition trees (the latter being described in the appendix). A detailed mathematical analysis, in particular the derivation of the generating function and further enumeration results, will be discussed elsewhere.

In order to store the partition function and the base pairing probabilities of joint structures in `rip`, we employ 4-dimensional arrays. For the recursion for the partition function, Q^I, we use 16 matrices, 24 matrices for Q^{RT}, 18 matrices for Q^{DT} and 45 matrices for Q^T, in the context of taking into account the loop energy. The complete set of partition function recursions and all details on the particular implementation of `rip` can be found at `http://www.combinatorics.cn/cbpc/rip.html`. The space complexity of `rip` is $O(N^4)$. Summations in our recursion equations run over at most two independent indices. Therefore, the time complexity in `rip` is $O(N^6)$. In order to obtain the pairing probabilities we trace back in the decomposition tree. Thus, we have the same space complexity and time complexity as for calculating the partition function.

ACKNOWLEDGEMENTS

We thank Bill Chen and Sven Findeiß for comments on the manuscript. This work was supported by the 973 Project of the Ministry of Science and Technology, the PCSIRT Project of the Ministry of Education, and the National Science Foundation of China to CMR and his lab, grant No. STA 850/7-1 of the Deutsche Forschungsgemeinschaft under the auspices of SPP-1258 "Small Regulatory RNAs in Prokaryotes", as well as the European Community FP-6 project SYNLET (Contract Number 043312) to PFS and his lab.

REFERENCES

Akutsu, T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Disc. Appl. Math.*, **104**, 45–62.

Alkan, C., Karakoc, E., Nadeau, J., Sahinalp, S. and Zhang, K. (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**, 267–282.

Andronescu, M., Zhang, Z. and Condon, A. (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**, 1101–1112.

Argaman, L. and Altuvia, S. (2000) *fhlA* repression by *OxyS* RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J. Mol. Biol.*, **300**, 1101–1112.

Bachellerie, J., Cavaillé, J. and Hüttenhofer, A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Banerjee, D. and Slack, F. (2002) Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays*, **24**, 119–129.

Benne, R. (1992) RNA editing in trypanosomes. the use of guide RNAs. *Mol. Biol.Rep.*, **16**, 217–227.

Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. and Hofacker, I. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3–3.

Busch, A., Richter, A. and Backofen, R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Cary, R. and Stormo, G. (1995) Graph-theoretic approach to RNA modeling using comparative data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 75–80.

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

Dimitrov, R. A. and Zuker, M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87**, 215–226.

Dirks, R. and Pierce, N. (2003) A partition function algorithm for nucleoic acid secondary structure inluding pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.

Gabow, H. (1973) *Implementation of algorithms for maximum matching on nonbipartite graphs*. Ph.D. thesis, Stanford University, Stanford (California). 248p.

Gago, S., De la Peña, M. and Flores, R. (2005) A kissing-loop interaction in a hammerhead viroid RNA critical for its in vitro folding and in vivo viability. *RNA*, **11**, 1073–1083.

Geissmann, T. and Touati, D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.*, **23**, 396–405.

Giegerich, R. and Meyer, C. (2002) *Lecture Notes In Computer Science*, volume 2422, chapter Algebraic Dynamic Programming, pp. 349–364. Springer-Verlag.

Hackermüller, J., Meisner, N., Auer, M., Jaritz, M. and Stadler, P. (2005) The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene*, **345**, 3–12.

Havgaard, J., Torarinsson, E. and Gorodkin, J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, 1896–1908.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Huang, F., Qin, J., Reidys, C. and Stadler, P. (2009) Target prediction for RNA-RNA interaction.

Isambert, H. and Siggia, E. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA*, **97**, 6515–6520.

Kugel, J. and Goodrich, J. (2007) An RNA transcriptional regulator templates its own regulatory RNA. *Nat. Struct. Mol. Biol.*, **3**, 89–90.

Majdalani, N., Hernandez, D. and Gottesman, S. (2002) Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol. Microbiol.*, **46**, 813–826.

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

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

McManus, M. and Sharp, P. (2002) Gene silencing in mammals by small interfering RNAs. *Nature Reviews*, **3**, 737–747.

Meisner, N., Hackermüller, J., Uhl, V., Aszódi, A., Jaritz, M. and Auer, M. (2004) mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure. *Chembiochem.*, **5**, 1432–1447.

Mneimneh, S. (2007) On the approximation of optimal structures for RNA-RNA interaction. *IEEE/ACM Trans. Comp. Biol. Bioinf.* In press, doi.ieeecomputersociety.org/10.1109/TCBB.2007.70258.

Mückstein, U., Tafer, H., Bernhard, S., Hernandez-Rosales, M., Vogel, J., Stadler, P. and Hofacker, I. (2008) Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In Elloumi, M., Küng, J., Linial, M., Murphy, R. F., Schneider, K. and Toma, C. T. (eds.), *BioInformatics Research and Development — BIRD 2008*, volume 13 of *Comm. Comp. Inf. Sci.*, pp. 114–127. Springer, Berlin.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhard, S., Stadler, P. and Hofacker, I. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182. Earlier version in: *German Conference on Bioinformatics 2005*, Torda andrew and Kurtz, Stefan and Rarey, Matthias (eds.), Lecture Notes in Informatics **P-71**, pp 3-13, Gesellschaft f. Informatik, Bonn 2005.

Narberhaus, F. and Vogel, J. (2007) Sensory and regulatory RNAs in prokaryotes: A new german research focus. *RNA Biol.*, **4**, 160–164.

Pervouchine, D. (2004) `IRIS`: Intermolecular RNA interaction search. *Proc. Genome Informatics*, **15**, 92–101.

Qin, J. and Reidys, C. (2008) A framework for RNA tertiary interaction. Submitted.

Rehmsmeier, M., Steffen, P., Höchsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *Gene*, **10**, 1507–1517.

Ren, J., Rastegari, B., Condon, A. and Hoos, H. (2005) Hotknots: heuristic prediction of microRNA secondary structures including pseudoknots. *RNA*, **11**, 1494–1504.

Rivas, E. and Eddy, S. (1999) A dynamic programming algorithms for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

Sharma, C., Darfeuille, F., Plantinga, T. and Vogel, J. (2007) A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes & Dev.*, **21**, 2804–2817.

Tafer, H., Kehr, S., Hertel, J. and Stadler, P. (2009) RNAsnoop: Efficient target prediction for box H/ACA snoRNAs. Submitted.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.