BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–12

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 subjects 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, ISMB 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).

Several restricted versions of RNA-RNA interaction have been considered in the literature. The simplest approach concatenates the two interacting sequences and subsequently employs a slightly modified standard 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) subscribe to this strategy. A 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). The resulting model, however, still does not generate all relevant interaction structures (Chitsaz *et al.*, 2009; Qin and Reidys, 2008). An alternative line of thought is to neglect all internal base-pairings in either strand and to compute the minimum free energy (mfe) secondary structure for their

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

hybridization under this constraint. For instance, RNAduplex and RNAhybrid (Rehmsmeier *et al.*, 2004) follows this line of thought. 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. To-date only a handful of interaction structures are known that are more complex than those covered by intaRNA/RNAup. The most famous example is the repression of *fhlA* by *OxyS* RNA, see Fig. 1, which involves to widely separated kissing hairpin loops (Argaman and Altuvia, 2000). A second important example is the binding of box H/ACA snoRNAs with their targets (Jin *et al.*, 2007). Due to the highly conserved interaction motif, snoRNA/target complexes are treated more efficiently using a specialized tool (Tafer *et al.*, 2009) however.

Pervouchine (2004) and Alkan *et al.* (2006) independently derived and implemented minimum free energy (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 "zig-zags", see Fig. 1 and 2 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.

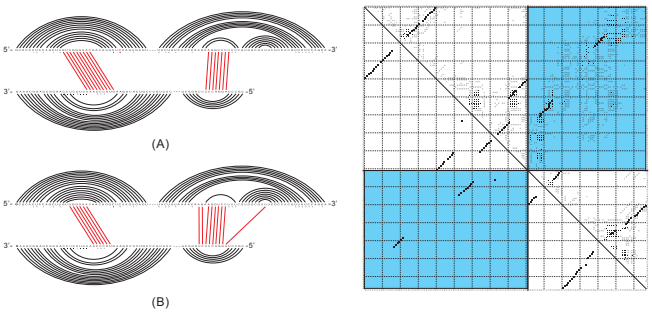

Fig. 1. Left: Natural structure (**A**) (Alkan *et al.*, 2006) and the joint structure predicted by rip, (**B**), formed by the the *OxyS* small RNA and its mRNA target *fhlA*. **Right:** Dot Plot showing the base pairing probabilities (proportional to the area of the squares) between the two structures (upper right triangle) and the interaction structure predicted by the maximum weighted matching algorithm (MWM) (Cary and Stormo, 1995; Gabow, 1973), in which the base pairs are weighted by its binding probability (lower left). The cut point between the two sequences is indicated by horizontal and vertical lines, intermolecular base pairs are depicted in the blue upper right and lower left rectangle, respectively.

Recently, Chitsaz *et al.* (2009) extended this approach to a dynamic programming algorithm that computes the partition function of "joint structures", also in $O(N^6)$ time. 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 the outline of McCaskill's approach for RNA secondary structure folding (McCaskill, 1990).

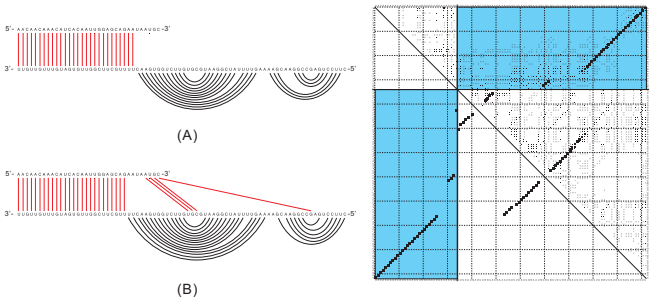

Fig. 2. Left: Joint structure predicted by IRIS (Pervouchine, 2004) (**A**) and rip (**B**) between the small RNA *gcvB* and its mRNA target *dppA* **Right:**: Dot Plot, as in Fig. 1.

The key idea is to identify a certain subclass of interaction structure that can serve as building blocks in a recursive decomposition generalizing the loop-decomposition of secondary structures. These "tight structures" are a generalization of the sub-secondary 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 energy minimization 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.

The output of rip consists of the partition function, the base pairing probability matrix and a specific joint structure. The latter is predicted by the maximal weighted matching algorithm (MWM) (Cary and Stormo, 1995; Gabow, 1973) in $O(N^3)$ time and $O(N^2)$ memory, where the base pairs are weighted by their respective binding probabilities. Finally, we discuss rip and showcase first example applications of the software rip.

2 JOINT STRUCTURES

2.1 Combinatorics of Interaction Structures

Two interacting RNAs are represented as *diagrams* (Chen *et al.*, 2008) R and S with N and M vertices, resp. In order to simply the notation in the following 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. The induced subgraph of S on a subsequence $(S_i, \ldots, S_j)$ is denoted by $S[i, j]$. In particular, $S[i, i] = S_i$ and $S[i, i-1] = \varnothing$.

A complex $C(R, S, I)$ is a graph consisting of two diagrams R and S (as induced subgraphs) and an additional set I of arcs of the form $R_i S_j$ such that each vertex has degree at most one, see Fig. 3. We shall draw $C(R, S, I)$ by arranging the vertices of R and S in two lines, showing the R-arcs in the upper, the S-arcs in the lower halfplane and I-arcs vertically. A subcomplex is a subgraph of C induced by the subsequences $(R_{i_1}, \ldots, R_{j_1})$ and $(S_{i_2}, \ldots, S_{j_2})$.

An arc is called *interior* if its start and endpoint are both contained in either R or S and *exterior*, otherwise. 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 $A_S(R_iS_j)$, resp. 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

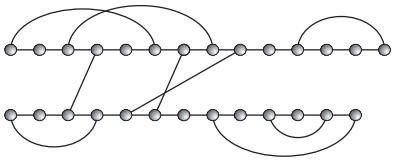

Fig. 3. A complex C induced by $R[1,14]$ and $S[1,13]$.

as R- and S-parents, see Fig. 4, (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.

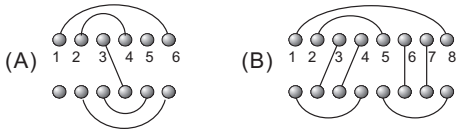

Fig. 4. (A) Ancestors and parents: for the exterior arc R_3S_4, we have the following ancestor sets $A_R(R_3S_4) = \{R_1R_6, R_2R_4\}$ and $A_S(R_3S_4) = \{S_2S_6, S_3S_5\}$. In particular, R_2R_4 and S_3S_5 are the R-parent and S-parent respectively.
(B) Subsumed and equivalent arcs: R_1R_8 subsumes S_1S_4 and S_5S_8. Furthermore, R_2R_5 is equivalent to S_1S_4.

Consider the subcomplex $C' = (R', S', I')$ induced by $R' = R[i_1, j_1]$ and $S' = S[i_2, j_2]$ and suppose there is an exterior arc R_aS_b with ancestors R_iR_j and $S_{i'}S_{j'}$. Then we say that R_iR_j is C'*-subsumed* in $S_{i'}S_{j'}$, if for any $R_kS_{k'} \in I'$, $i < k < j$ implies $i' < k' < j'$. In case of $C' = C$, we call R_iR_j simply "subsumed" in $S_{i'}S_{j'}$, see Fig. 4, (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*.

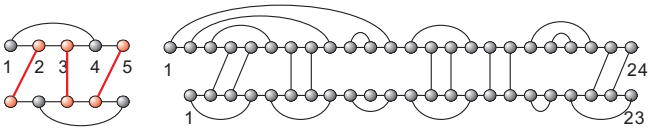

Fig. 5. Left: A zig-zag, generated by R_2S_1, R_3S_3 and R_5S_4 (red). **Right:** A joint structure induced by $R[1,24]$ and $S[1,23]$.

A *joint structure*, $J_{i,j;h,\ell} = J(R[i,j]; S[h,\ell], I')$, Fig. 5, is a subcomplex of $C(R,S,I)$ with the following properties:

1. R, S are secondary structures without internal pseudoknots.
2. There are no external pseudoknots, i.e., if $R_{i_1}S_{j_1}, R_{i_2}S_{j_2} \in I'$ where $i_1 < i_2$, then $j_1 < j_2$.
3. There are no "zig-zags", if $R_{i_1}R_{j_1}$ and $S_{i_2}S_{j_2}$ are dependent, then either $R_{i_1}R_{j_1}$ is subsumed by $S_{i_2}S_{j_2}$ or *vice versa*.

In absence of exterior arcs a joint structure reduces to a pair of pieces of secondary structure on R and S, to which we will refer as a pair of *segment*s for short. As segment $S[i_1, j_1]$ is maximal if there is no segment, $S[i,j]$ strictly containing $S[i_1, j_1]$.

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). In the following we introduce tight structures (ts), or tights, a specific class of joint structures. Tights form the basis of our algorithmic approach and can be viewed as the transitive closure of standard loops w.r.t. exterior arcs.

Fix an arbitrary joint structure $J_{a,b;c,d}$. Then $J_{i,j;h,\ell} \subset J_{a,b;c,d}$ is *tight* in $J_{a,b;c,d}$ if

1. it contains at least one exterior arc $R_{i_1}S_{j_1}$
2. for any exterior arc $R_{i_1}S_{j_1} \in J_{i,j;h,\ell}$ holds $\big(A_R(R_{i_1}S_{j_1}) \cup A_S(R_{i_1}S_{j_1})\big) \cap J_{a,b;c,d} \in J_{i,j;h,\ell}$
3. there does not exist any $J_{i_1,j_1;h_1,\ell_1} \subsetneq J_{i,j;h,\ell}$ containing at least one exterior arc, $R_{i_1}S_{j_1}$, such that for any such $R_{i_1}S_{j_1}$ $\big(A_R(R_{i_1}S_{j_1}) \cup A_S(R_{i_1}S_{j_1})\big) \cap J_{a,b;c,d} \in J_{i_1,j_1;h_1,\ell_1}$ holds.

Given a (ts) $J^T_{i,j;h,\ell}$, we observe that neither one i, j, h and ℓ, can be start or endpoints of a segment. In particular, neither i, j, h, and ℓ are isolated. In combination with the non-zig-zag property, we conclude that there are only the following four basic types of (ts), Fig. 6:

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

The latter case corresponds to a single external edge.

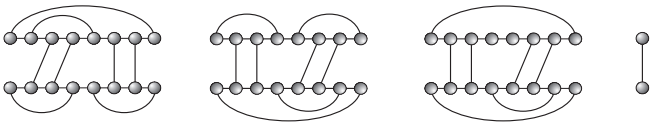

Fig. 6. Tight structures (ts) of type $\bigtriangledown$, $\bigtriangleup$, $\square$, and $\circ$.

In the Appendix, we will prove the following

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

1. *Any $J_{a,b;c,d}$-(ts) is of one of the four types $\bigtriangledown$, $\bigtriangleup$, $\square$, or $\circ$*
2. *Any exterior arc in $J_{a,b;c,d}$ is contained in a unique $J_{a,b;c,d}$-(ts)*
3. *$J_{a,b;c,d}$ decomposes into a unique collection of $J_{a,b;c,d}$-(ts) and maximal segments.*

Given a (ts) $J^{\bigtriangledown}_{i_0,j_0;r,s}$, a *double-tight joint structure* (dts), $J^{DT,|\bigtriangledown}_{i,j;r,s}$, 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 $r \le c < d \le s$ and (ts) $J^T_{i,a;r,c}$, $J^T_{b,j;d,s}$ in $J^T_{i_0+1,j_0-1;r,s}$ such that

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

where the disjoint union $\dot\cup$ refers to both the vertex and arc sets of the joint structures, see Fig. 7. The case of a double-tight joint structure, $J^{DT,|\bigtriangleup}_{i,j;r,s}$, in a (ts), $J^{\bigtriangleup}_{i,j;r_0,s_0}$, is defined accordingly. By abuse of language, we simply use $J^{DT}_{i,j;r,s}$ in order to to denote either $J^{DT,|\bigtriangledown}_{i,j;r,s}$ or $J^{DT,|\bigtriangleup}_{i,j;r,s}$.

With the help of (dts), we decompose (ts) in the following way:
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

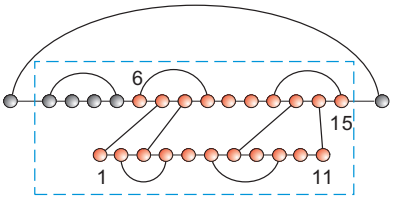

Fig. 7. A (dts) $J^{DT|\bigtriangledown}_{6,15;1,11}$ (red) in $J_{2,15;1,11}$ (blue box). Note that the joint structure $J_{1,16;1,11}$ itself is $\bigtriangledown$-tight.

$J_{i+1,j-1;r,s}$ decomposes into

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

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

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

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. In Fig. 8 we show how to decompose a tight of the types $\bigtriangledown$, $\bigtriangleup$, or $\square$ by means of equ. (2.2-2.4). The above decompositions form the combinatorial basis for the computation of the partition function.

2.2 Refined Decomposition

The unique (ts) decomposition would in principle already 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 functions associated with the joint structures. From a practical point of view, however, this would result in an unwieldy expensive implementation. The reason are the multiple break points a, b, c, d, ..., each of which correspond to a nested `for`-loop.

We therefore need a refined decomposition that reduced the number of break points. To this end 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. 8 summarizes two major steps in the decomposition: (I) "arc-removal" to reduce (dts). The scheme is complemented by the usual loop decomposition of secondary structures, and (II) "block-decomposition" to split a joint structure into two blocks.

The details of the decomposition procedures are collected in Appendix C, where it is shown that for each joint structure $J_{1,N;1,M}$ we indeed obtain a unique decomposition-tree (parse-tree), denoted by $T_{J_{1,N;1,M}}$. More precisely, $T_{J_{1,N;1,M}}$ has root $J_{1,N;1,M}$ and all other vertices correspond to a specific sub-complex of $J_{1,N;1,M}$ obtained by the successive application of the decomposition steps of Fig. 8 and the loop decomposition of the secondary structures. A concrete example is shown in Fig. 9.

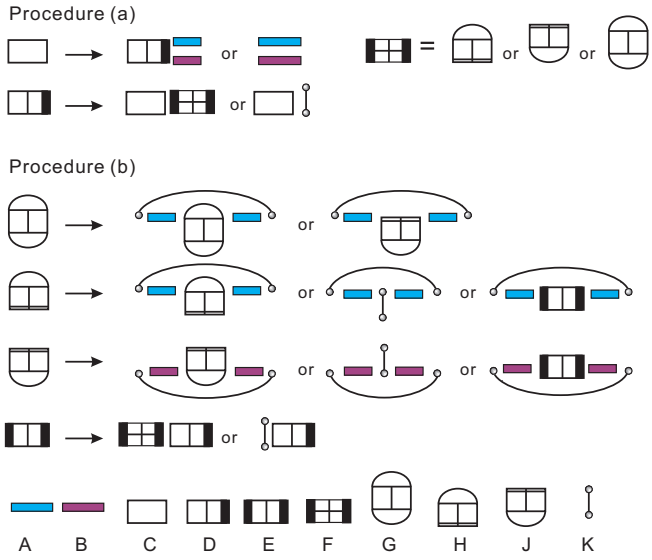

Fig. 8. Illustration of Procedure (a) the reduction of arbitrary joint structures and right-tight structures, and Procedure (b) the decomposition of tight structures. The panel below indicates the 10 different types of structural components: **A**, **B**: maximal secondary structure segments $R[i,j]$, $S[r,s]$; **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 $\bigtriangledown$, $\bigtriangleup$ or $\square$; **G**: type $\square$ tight structure $J^{\square}_{i,j;r,s}$; **H**: type $\bigtriangledown$ tight structure $J^{\bigtriangledown}_{i,j;r,s}$; **J**: type $\bigtriangleup$ tight structure $J^{\bigtriangleup}_{i,j;r,s}$; **K**: exterior arc.

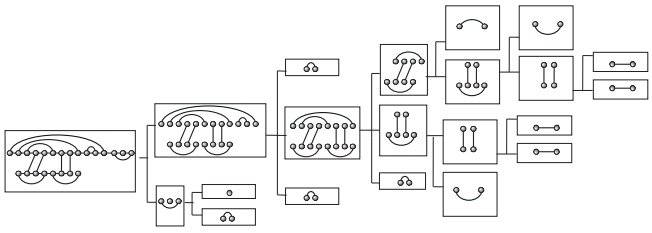

Fig. 9. The decomposition tree $T_{J_{1,15;1,8}}$ for the joint structure $J_{1,15;1,8}$.

2.3 Extended Loop Model

The standard energy model for RNA folding (Mathews *et al.*, 1999) is consistent with the basic decomposition of secondary structure diagrams in the following sense: for secondary structures, the analogue of Fig. 8 reads

$$S \to .S \mid PS \mid P \quad \text{and} \quad P \to (S) \tag{2.5}$$

representing the cases that either the first base pair is unpaired or paired. Here S denotes an arbitrary structure, while P is secondary structure enclosed by a base pair. In fact, we use this decomposition to evaluate the secondary structure segments A and B in Fig. 8.

The energy model, however, enforces a further refinement of the decomposition by distinguishing three different types of loops, for which energy contributions need to be computed by means of different rules:

hairpin loops $P \longrightarrow \mathsf{Ha}$, interior loops (including bulges and stacked base pairs) $P \longrightarrow \mathsf{Int}$, and multi-branched loops: $P \longrightarrow \mathsf{M}$. These are now expanded further

$$\mathsf{Ha} \longrightarrow (h) \qquad \mathsf{Int} \longrightarrow (i'Pi'') \qquad \mathsf{M} \longrightarrow (M'M'') \tag{2.6}$$

where h, i', i'' are the unpaired regions of the hairpin and interior loops. Multi-branch loops are further decomposed into components with a single branch M' and with multiple branches M'' for which the energy contributions are assumed to be additive. For completeness, we recall the productions $M' \longrightarrow .M'|P$ and $M'' \longrightarrow .M''|PM''|Pm$, where m is a stretch of unpaired nucleotides. The importance of this refined decomposition lies in the fact that the energy of each substructure can be obtained as a sum of the energies of the substructures associated with non-terminal symbols and an additional contribution that depends uniquely on the production and the terminals. The latter rules form the specific *energy parameters* (Mathews *et al.*, 1999).

The description of joint structures includes two further types of loops, both of which being arising in the context of exterior arcs. Following Chitsaz *et al.* (2009), we call them *hybrid* and *kissing loop*, Fig. 10.

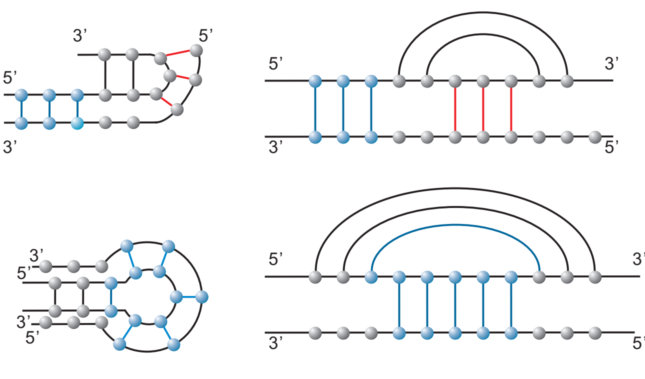

Fig. 10. The two new loop types in interaction structures: the hybrid (top) and the kissing loop (bottom).

- A *hybrid* ($\mathsf{Hy}_{i_1,i_\ell;j_1,j_\ell}$) consists of a series of intermolecular interior loops formed by $\ell \geq 2$ exterior arcs $R_{i_1}S_{j_1},\ldots,R_{i_\ell}S_{j_\ell}$ such that $R_{i_h}S_{j_h}$ is nested within $R_{i_{h+1}}S_{j_{h+1}}$ and the internal segments $R[i_h+1, i_{h+1}-1]$ and $S[j_h+1, j_{h+1}-1]$ consist of unpaired bases only. In other words, a hybrid is an unbranched stem-loop structure formed by external arcs
- A *kissing*-loop (K_{R_i,R_j}) is either a pair, $(R_iR_j, R[i+1,j-1])$, together with a nonempty set of R_iR_j-children, $R_{i_1}S_{j_1},\ldots$ where $i < i_1 < j$, or a pair $(S_iS_j, S[i+1,j-1])$, with a nonempty set of S_iS_j-children $R_{i_1}S_{j_1},\ldots$ where $i < j_1 < j$. Kissing loops have been singled out both for logical reasons and because some investigations into their thermodynamic properties have been reported in the literature (Gago *et al.*, 2005).

Let us now have a closer look at the energy evaluation of $J_{i,j;h,\ell}$. Each decomposition step in Fig. 8 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:

I. Arc removal. Most of the decomposition operations in Procedure (b) displayed in Fig. 8 can be viewed as the "removal" of an arc (corresponding to the closing pair of a loop in secondary structure folding) followed by a decomposition. Both: the loop-type and the subsequent possible decomposition steps depend on the newly exposed structural elements. W.l.o.g., we may 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 $i \leq p < q \leq j$ and all exterior pairs R_pS_h with $i \leq p \leq j$. 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:

$$
\begin{aligned}
E_R = \emptyset &\ \text{ hairpin loop}\\
E_R = E^i_R, |E_R| = 1 &\ \text{ interior loop (including bulge and stacks)}\\
E_R = E^i_R, |E_R| \geq 2 &\ \text{ multi-branch loop}\\
E_R = E^e_R &\ \text{ kissing hairpin loop}\\
|E^i_R|, |E^e_R| \geq 1 &\ \text{ general kissing loop}
\end{aligned}
$$

This picture needs to be refined even further since the arc removal is coupled with a 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.

For a (ts) $J^{\bigtriangledown}_{i,j;h,\ell}$ of type $\bigtriangledown$, $J^{\bigtriangledown}_{i,j;h,\ell}$ is of type E, if $S[h,\ell]$ is not enclosed in any base pair ($J^{\bigtriangledown,\mathsf{E}}_{i,j;h,\ell}$). Suppose $J^{\bigtriangledown}_{i,j;h,\ell}$ 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 $J^{\bigtriangledown}_{i,j;h,\ell}$ is of type M ($J^{\bigtriangledown,\mathsf{M}}_{i,j;h,\ell}$). 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. 11 displays the further decomposition for $J^{\bigtriangledown,\mathsf{M}}_{i,j;r,s}$.

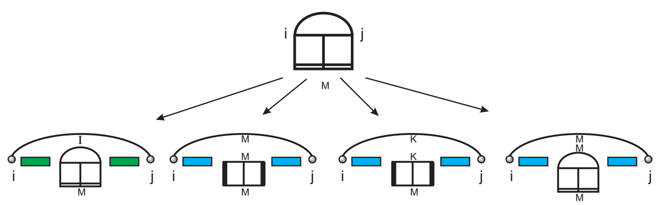

Fig. 11. Further refinement: the four decompositions of $J^{\bigtriangledown,\mathsf{M}}_{i,j;r,s}$ via Procedure (b). The green rectangle denotes isolated segments. The letters I, M, etc denote the loop-type and the type of of the exposed arc(s) of the double-tight structures. See also Fig. 8 for more details on the notation. The four cases correspond to the four contributions in equ. (3.3).

For a (dts) $J^{DT}_{p,q,r,s}$ (denoted by "E" in Fig. 8) 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, i.e., type EE does not exit, where E means external.

II. Block decomposition. The second type of decomposition is the splitting of joint structures into "blocks", such as the decompositions of a right-tight structure in Procedure (a) and a double-tight joint structure in Procedure (b) in Fig. 8. A right-tight structures $J^{RT}_{i,j;h,\ell}$ may appear in two ways, depending on whether or not the rightmost tight structure is of Type $\circ$. More precisely, let $R_{i_1}S_{j_1}$ denote the rightmost exterior arc in $J^{RT}_{i,j;h,\ell}$. There is a unique $J_{i,j;h,\ell}$-(ts) $J^T_{R_{i_1}S_{j_1}}$, such that $R_{i_1}S_{j_1} \in J^T_{R_{i_1}S_{j_1}}$. We distinguish type (r$\mathsf{B}$) loops induced by $J^T_{R_{i_1}S_{j_1}}$ being of type $\circ$ from all other cases (rA). Analogously, $(l\mathsf{B})$ and $(l\mathsf{A})$ are defined for double-tight joint structure. For instance, Fig. 12 displays the decomposition of

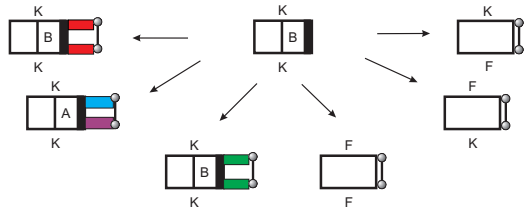

Fig. 12. Decomposition of $J^{RT,\mathsf{KKB}}_{i,j;h,\ell}$ by means of procedure (b). Here the red rectangle denotes a pair of secondary segments having the property that at least one of them is not isolated.

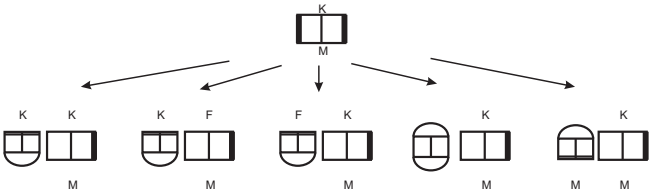

Fig. 13. Decomposition of $J^{DT,\mathsf{KM}}_{i,j;h,\ell}$ by means of procedure (b). The five alternatives correspond to the three additive terms in equ. (3.4).

$J^{RT,\mathsf{KKB}}_{i,j;h,\ell}$. 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 $E^e_{R[i,j]}$, see Fig. 13. The situation is analogous if we decompose a joint structure, $J^{RT}_{i,j;r,\ell}$, which is contained in a kissing loop.

Energy Parameters. The energy model implemented in `rip` in an extension of the standard parametrization. In recognizes the following loop-types:

1. *Hairpin-loop* A hairpin loop $\mathsf{Ha}_{i,j}$ has tabulated energies $G^{\mathsf{Ha}}_{i,j}$ depending on their sequence and length.
2. *Interior-loop* An interior loop $\mathsf{Int}_{i_1,j_1;i_2,j_2}$ also have tabulated energies $G^{\mathsf{Int}}_{i_1,j_1;i_2,j_2}$.
3. *Multi-loop* A multi-loop M_{i_0,j_0} has energy $\alpha_1 + \alpha_2(t+1) + \alpha_3 c_2$, where $t = |E^i_{R[i_0,j_0]}|$ ("branching order") inside $R[i_0,j_0]$ and c_2 is the number of isolated vertices contained in $R[i_0,j_0]$.
4. *Kissing-loop* A kissing-loop K_{i_0,j_0} has energy $\beta_1 + \beta_2(t+1) + \beta_3 c_2$, where $t = |E^i_{R[i_0,j_0]}|$ and c_2 is the number of isolated vertices contained in $R[i_0,j_0]$ in analogy with the parametrization of multiloops.
5. *Hybrid* A hybrid $\mathsf{Hy}_{i_1,i_\ell;j_1,j_\ell}$ has energy $G^{\mathsf{Hy}}_{i_1,i_\ell;j_1,j_\ell} = \sigma_0 + \sigma\sum_\theta G^{\mathsf{Int}}_{i_\theta,i_{\theta+1};j_\theta,j_{\theta+1}}$, where a intermolecular interior loop formed by $R_{i_\theta}S_{j_\theta}$ and $R_{i_{\theta+1}}S_{j_{\theta+1}}$ is treated like interior loop $\mathsf{Int}_{i_\theta,j_\theta;i_{\theta+1},j_{\theta+1}}$ with an affine scaling σ.

In the Discussion we also consider a different energy parametrization of hybrid loops (Bernhart *et al.*, 2006) and the approach of Dimitrov and Zuker (2004).

3 PARTITION FUNCTION

3.1 Recursions for secondary structures

The additivity of loop energies, see Section 2.3, translates immediately into the multiplicativity of the Boltzmann factors that contribute to the partition function $Q = \sum_S e^{-F(S)/kT}$, where the sum runs over all the secondary structures S that can be constructed from a given sequence of length M. This factorization of terms can be realized by introducing $Q^b_{h,\ell}$, where the sum is taken over all substructures $S[h,\ell]$ on the segment $[h,\ell]$ for which $S_hS_\ell \in S[h,\ell]$ is a base pair, and $Q^s_{h,\ell}$ for all the configurations on $[h,\ell]$, irrespective of whether or not S_h and S_ℓ are base paired. In particular, we have $Q = Q^s_{1,N}$ and arrive at the recursion (right to left block decomposition)

$$Q^s_{h,\ell} = 1 + \sum_{i,j} Q^s_{h,i-1}Q^b_{i,j}. \tag{3.1}$$

In order to compute $Q^b_{h,\ell}$ we need to distinguish between the three types of loops that are treated differently in standard RNA energy model (Mathews *et al.*, 1999): hairpin loops, interior loops (including bulges and stacked base pairs), and multi-branch loops, see Figure 14.

$$\begin{aligned}
Q^b_{h,\ell} =& e^{-G^{\mathsf{Ha}}_{h,\ell}/kT} + \sum_{i,j} e^{-G^{\mathsf{Int}}_{i,j;h,\ell}/kT} Q^b_{i,j} \\
&+ \sum_{i,j} Q^{\mathsf{m}}_{h+1,i-1} Q^b_{i,j} e^{-(\alpha_1+2\alpha_2+(\ell-j-1)\alpha_3)/kT}, \\
Q^{\mathsf{m}}_{h,\ell} =& \sum_{i,j} (Q^{\mathsf{m}}_{h,i-1} + e^{-(i-h)\alpha_3/kT}) Q^b_{i,j} e^{-(\alpha_2+(\ell-j)\alpha_3)/kT},
\end{aligned} \tag{3.2}$$

where Q^m is the auxiliary array representing multi-branched loops, as described in detail by McCaskill (1990).

3.2 Recursions for joint structures

The production rules described in the previous section are now translated into recursion equations for the partition function of each type. The computation of the partition function proceeds "from the inside to the outside", see equs. (3.3,3.4). 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 $Q^{\mathsf{k}}_{i,j}$. 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, $Q^{\mathsf{k}}_{i,j}$ 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. 11 reads:

$$\begin{aligned}
Q^{\triangledown,\mathsf{M}}_{i,j;r,s} = \sum_{h,\ell} \Big\{ & Q^{\triangledown,\mathsf{M}}_{h,\ell;r,s} e^{-G^{\mathsf{Int}}_{i,j;h,\ell}/kT} \\
&+ Q^{DT,\mathsf{MM}}_{h,\ell;r,s} e^{-(\alpha_1+\alpha_2)/kT} \times (e^{-(h-i-1)\alpha_3/kT} + Q^{\mathsf{m}}_{i+1,h-1}) \\
&\quad \times (e^{-(j-\ell-1)\alpha_3/kT} + Q^{\mathsf{m}}_{\ell+1,j-1}), \\
&+ Q^{DT,\mathsf{KM}}_{h,\ell;r,s} e^{-(\beta_1+\beta_2)/kT} \times (e^{-(h-i-1)\beta_3/kT} + Q^{\mathsf{k}}_{i+1,h-1}) \\
&\quad \times (e^{-(j-\ell-1)\beta_3/kT} + Q^{\mathsf{k}}_{\ell+1,j-1}), \\
&+ Q^{\triangledown,\mathsf{M}}_{h,\ell;r,s} e^{-(\alpha_1+2\alpha_2)/kT} \big[e^{-(j-\ell-1)\alpha_3/kT} Q^{\mathsf{m}}_{i+1,h-1} \\
&\quad + e^{-(h-i-1)\alpha_3/kT} Q^{\mathsf{m}}_{\ell+1,j-1} + Q^{\mathsf{m}}_{\ell+1,j-1} Q^{\mathsf{m}}_{i+1,h-1} \big] \Big\}.
\end{aligned} \tag{3.3}$$

Analogously, the recursion for the (dts) $Q^{DT,\mathsf{KM}}_{i,j;r,s}$ of Fig.13 is given by

$$Q^{DT,\mathsf{KM}}_{i,j;r,s} = \sum_{i_1,j_1} \Big\{ (Q^{\triangledown,\mathsf{M}}_{i,i_1;r,j_1} e^{-\beta_2/kT} + Q^{\triangle,\mathsf{K}}_{i,i_1;r,j_1} e^{-\alpha_2/kT} + Q^{\square}_{i,i_1;r,j_1} e^{-(\alpha_2+\beta_2)/kT} + Q^{\triangle,\mathsf{F}}_{i,i_1;r,j_1} e^{-\alpha_2/kT}) Q^{RT,\mathsf{KM}}_{i_1+1,j;j_1+1,s} + Q^{\triangle,\mathsf{K}}_{i,i_1;r,j_1} e^{-\alpha_2/kT} Q^{RT,\mathsf{FM}}_{i_1+1,j;j_1+1,s} \Big\}. \tag{3.4}$$

The complete set of recursions comprises for (ts) $Q^{T}_{i,j;r,s}$, 15 4D-arrays, for right-tight structures $Q^{RT}_{i,j;r,s}$, 24 4D-arrays, for (dts) $Q^{DT}_{i,j;r,s}$, 18 4D-arrays and 16 4D-arrays for arbitrary interaction structures $Q^{I}_{i,j;r,s}$. In addition, we need the usual matrices for the secondary structures R and S, and the above mentioned matrices for kissing loops. The full set of recursions is compiled in the Supplemental Material.

4 BASE PAIRING PROBABILITIES

4.1 Approach

In contrast to the computation of the partition function "from the inside to the outside", the computation of the base pairing probabilities are obtained "from the outside to the inside". That is, the longest-range pairs are computed first. This is 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 right-tight structure 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 2.2.

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

$$\mathbb{P}_{J_{i,j;h,\ell}} = \sum_{J \in \Lambda_{i,j;h,\ell}} \mathbb{P}_J \tag{4.1}$$

and furthermore

$$\mathbb{P}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell} = \sum_{J_{i,j;h,\ell} \in \mathbb{J}^{\xi,Y_1Y_2Y_3}_{i,j;h,\ell}} \mathbb{P}_{i,j;h,\ell}. \tag{4.2}$$

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

$$R_iR_j \in W \iff R_iR_j \in T_W. \tag{4.3}$$

Let $\Omega(R_iR_j) = \{W \mid R_iR_j \in T_W\}$, i.e. the set of secondary structures whose decomposition tree contain the pair R_iR_j as a leaf. Clearly, via equ. (4.3), we obtain

$$\mathbb{P}_{R_iR_j} = \sum_{W \in \Omega(R_iR_j)} Q_W Q^{-1}. \tag{4.4}$$

Next, in order to compute $\mathbb{P}_{R_iR_j}$, we need to express this probability via sum over the probabilities of the substructures ξ such that ξ is the 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

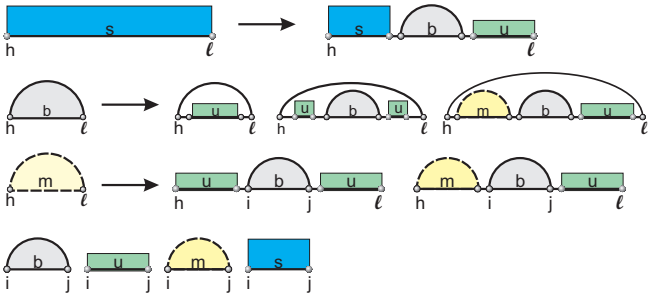

Fig. 14. Extended version of Procedure (c). The panel below indicates from left to right the segments $R^b(i,j)$, in which R_i, R_j is paired, the set of unpaired segments, $R^u(i,j)$, $R^{\mathsf{m}}(i,j)$, containing least one arc with an outer loop of type M and finally $R^s(i,j)$, the set of arbitrary segments.

is summarized in Procedure (c), Fig. 14. This representation differes from the usual implementation of the RNA standard folding model only it that we can afford more moving indices in each recursion since the entire `rip` algorithms requires $O(n^6)$ time. Inspection of Fig. 14 shows that for an $R^b(i,j)$-parent we have to distinguish the five cases displayed in Fig. 15. Denote by $R^{\mathsf{m}}(i,j)$ 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 write $R^s(i,j)$ for the set of all segments $R[i,j] \in T_{R[1,N]}$. Furthermore, set $\mathbb{P}^{\mathsf{m}}_{R_i,R_j}$ and $\mathbb{P}^{s}_{R_i,R_j}$ for the probabilities of $R^{\mathsf{m}}(i,j)$ and $R^s(i,j)$, respectively. For (L1) and (L4) in Fig. 15, it is possible 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 from the inside to the outside. Following the logic of

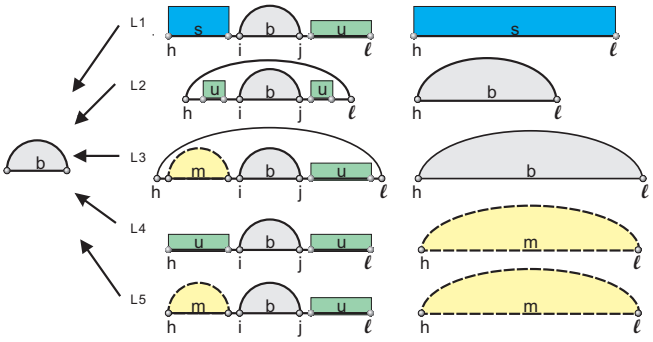

Fig. 15. Backtracing for secondary structures: for a parent of $R^b(i,j)$ we have five cases according to Procedure (c): external (L1), interior loop (L2), closing pair of a multi-loop (L3), (L4) and (L5) denote the scenarios arising from decomposing a $R^m(h,\ell)$-segment. See equ. (4.5) for the corresponding recursions.

Fig. 15, we obtain the following recursion for $\mathbb{P}^b_{R_i,R_j}$

$$\begin{aligned}\mathbb{P}^b_{R_i,R_j} = \sum_{h,\ell}\Bigg\{ &\underbrace{\mathbb{P}^s_{R_h,R_\ell}\frac{Q^s_{h,i-1}Q^b_{i,j}}{Q^s_{h,\ell}}}_{\text{L1}} \\ &+\underbrace{\mathbb{P}^b_{R_h,R_\ell}\frac{Q^b_{i,j}e^{-G^{\text{Int}}_{h,\ell;i,j}/kT}}{Q^b_{h,\ell}}}_{\text{L2}} \\ &+\underbrace{\mathbb{P}^b_{R_h,R_\ell}\frac{Q^{\text{m}}_{h+1,i-1}Q^b_{i,j}e^{-(\alpha_1+2\alpha_2+(\ell-j-1)\alpha_3)/kT}}{Q^b_{h,\ell}}}_{\text{L3}} \\ &+\underbrace{\mathbb{P}^{\text{m}}_{R_h,R_\ell}\frac{Q^b_{i,j}e^{-(\alpha_2+(i-h+\ell-j)\alpha_3)/kT}}{Q^{\text{m}}_{h,\ell}}}_{\text{L4}} \\ &+\underbrace{\mathbb{P}^{\text{m}}_{R_h,R_\ell}\frac{Q^{\text{m}}_{h,i-1}Q^b_{i,j}e^{-(\alpha_2+(\ell-j)\alpha_3)/kT}}{Q^{\text{m}}_{h,\ell}}}_{\text{L5}}\Bigg\}.\end{aligned} \tag{4.5}$$

Analogously, the recursions for the base pairing probabilities $\mathbb{P}^{\text{m}}_{R_i,R_j}$ and $\mathbb{P}^s_{R_i,R_j}$ are given by

$$\begin{aligned}\mathbb{P}^{\text{m}}_{R_i,R_j} &= \sum_{h,\ell}\Bigg\{\mathbb{P}^b_{R_{i-1},R_\ell}e^{-(\alpha_1+2\alpha_2+(\ell-1-h)\alpha_3)/kT} \\ &\quad\times\frac{Q^b_{j+1,h}Q^{\text{m}}_{i,j}}{Q^b_{i-1,\ell}}+\mathbb{P}^{\text{m}}_{R_i,R_\ell}\frac{Q^{\text{m}}_{i,j}Q^b_{j+1,h}e^{-(\alpha_2+(\ell-h)\alpha_3)/kT}}{Q^{\text{m}}_{i,\ell}}\Bigg\} \\ \mathbb{P}^s_{R_i,R_j} &= \sum_{h,\ell}\mathbb{P}^s_{R_i,R_\ell}\frac{Q^s_{i,j}Q^b_{j+1,h}}{Q^s_{i,\ell}}.\end{aligned} \tag{4.6}$$

4.3 Base pairing probabilities for joint structures

Set $\Sigma_1 = \{J \mid R_iR_j \in J\}$. Now we apply the same strategy to the joint structures appearing in Fig. 8. Let Q^I denote the partition function which sums over all the possible joint structures $J_{1,N;1,M}$. Then by definition, we have

$$\mathbb{P}_{R_i,R_j} = \frac{\sum_{J\in\Sigma_1}Q_J}{Q^I}. \tag{4.7}$$

In order to compute $\mathbb{P}_{R_iR_j}$ we classify Σ_1 according to the parent of R_iR_j in T:

$$\begin{aligned}\Sigma_1 = &\{J \mid R[i,j]\in T, R[i,j]\in R^b(i,j)\} \\ &\cup\bigcup_{h,\ell}\{J \mid J_{i,j;h,\ell}\in T, J_{i,j;h,\ell}\in\mathbb{J}^{\triangledown}_{i,j;h,\ell}\} \\ &\cup\bigcup_{h,\ell}\{J \mid J_{i,j;h,\ell}\in T, J_{i,j;h,\ell}\in\mathbb{J}^{\square}_{i,j;h,\ell}\},\end{aligned} \tag{4.8}$$

which translates to

$$\mathbb{P}_{R_iR_j} = \mathbb{P}^b_{R_i,R_j}+\sum_{h,\ell}\mathbb{P}^{\triangledown,\{\text{E,M,F,K}\}}_{i,j;h,\ell}+\sum_{h,\ell}\mathbb{P}^{\square}_{i,j;h,\ell}, \tag{4.9}$$

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

$$\begin{aligned}\Sigma_2 = &\{J \mid S[h,\ell]\in T, S[h,\ell]\in S^b[h,\ell]\} \\ &\cup\bigcup_{i,j}\{J \mid J_{i,j;h,\ell}\in T, J_{i,j;h,\ell}\in\mathbb{J}^{\triangle}_{i,j;h,\ell}\},\end{aligned} \tag{4.10}$$

and therefore

$$\mathbb{P}_{S_iS_j} = \mathbb{P}^b_{S_i,S_j}+\sum_{h,\ell}\mathbb{P}^{\triangle}_{h,\ell;i,j}, \tag{4.11}$$

with $\mathbb{P}^{\triangle} = \mathbb{P}^{\triangle,\text{E}}+\mathbb{P}^{\triangle,\text{M}}+\mathbb{P}^{\triangle,\text{K}}+\mathbb{P}^{\triangle,\text{F}}$.

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. 8 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. (4.2).

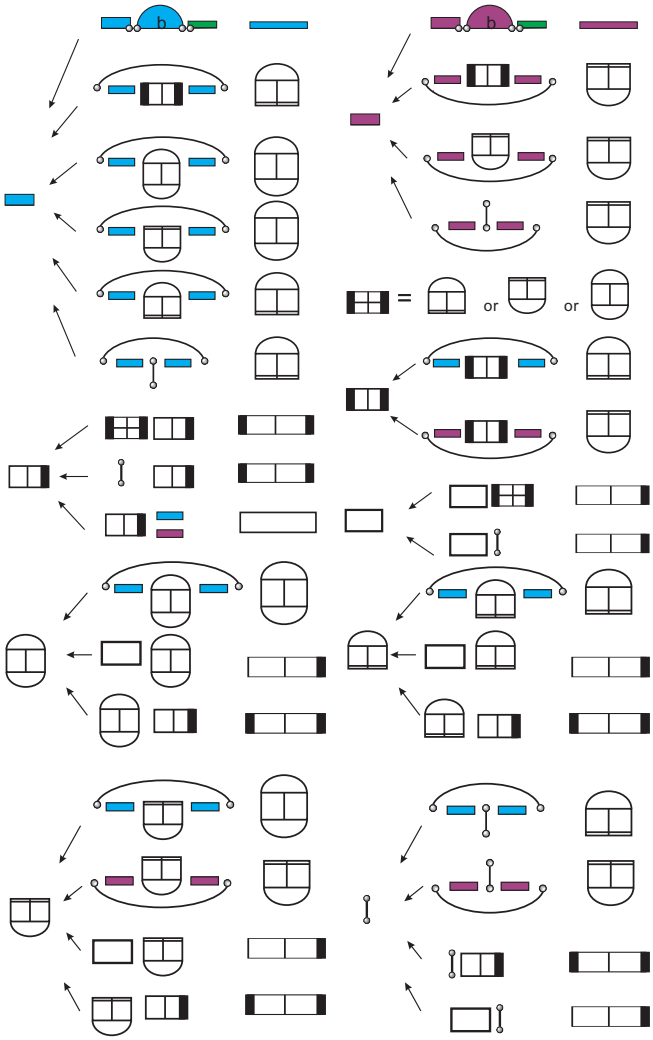

Fig. 16. Illustration of the back-tracing procedure. For each term, we need to add up the contributions of derivations that contain a particular substructure.

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 Fig. 16. By virtue of this backtracking, we obtain the recurrence formulae in analogy to the case of secondary structures, discussed above.

5 RESULTS AND DISCUSSION

In this contribution we have introduced a framework in which 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 and a multi-loop-like additive parametrization for kissing loops. The partition function computed by `rip` is equivalent to that derived by Chitsaz *et al.* (2009) based on a different mathematical framework.

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 decomposition trees (the latter described in full detail in the appendix). A detailed mathematical analysis, in particular the derivation of the generating function and further enumeration results, will be discussed elsewhere.

The computation of the partition function conceptually follows the logic of the McCaskill's approach (McCaskill, 1990) for RNA secondary structures. The generalization of the computation of the base pairing probabilities, however, is less straight-forward. The reason is that base pairs in joint structures are not always the unique closing pairs of loop, hence base pairing probabilities cannot be identified directly with the probabilities of certain (ts). Instead, one has to compute the pairing probabilities by explicitly "tracing back" all contributing joint structures.

For consistency with the approach of Chitsaz *et al.* (2009), we include an independent initialization energy σ_0 to each hybrid. A scaling of the energies in hybrids is also implemented. Alternatively, an single initialization energy ε is used in many other RNA-cofolding algorithms, including `RNAcofold` (Bernhart *et al.*, 2006) and the approach of Dimitrov and Zuker (2004). This initialization term can be introduced *a posteriori* once the partition function of the joint structures Q^I and the partition functions Q_R and Q_S of the isolated interaction partners have been computed. Let Ω_1 denote the set of all joint structures containing at least one external arc and denote by Ω_0 the set of all structures that have none. The partition function computed by `rip` is $Q^{\texttt{rip}} = Q(\Omega_1) + Q(\Omega_0)$. For the non-interacting contribution we have $Q(\Omega_0) = Q_R Q_S$. Taking the initiation term into account, we have to compute $Q = Q(\Omega_1)\exp(-\varepsilon/kT) + Q_R Q_S$, from which we easily obtain the corrected value for $Q(\Omega_1)$.

In addition to computing the equilibrium thermodynamics of RNA-RNA interactions (Chitsaz *et al.*, 2009), the `rip` software also predicts details of the interaction structures themselves. The base pairing probabilities are represented in "dotplots" analogous to those in `RNAcofold` (Bernhart *et al.*, 2006), Fig. 17. Two diagonal blocks (in white) contain the internal base pairs of the two pairing interaction molecules, the shaded rectangles display the interacting pairs. The upper-right triangle shows the base pairing probabilities. In the lower-left triangle, the unweighted maximum expected accuracy structure is displayed. It is given by the optimal RNA-RNA interaction structure with each possible base pair weighted by its base pairing probability. As shown by Bernhart *et al.* (2006), the pairing probabilities can also be rescaled quite easily for a single initiation energy contribution:

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

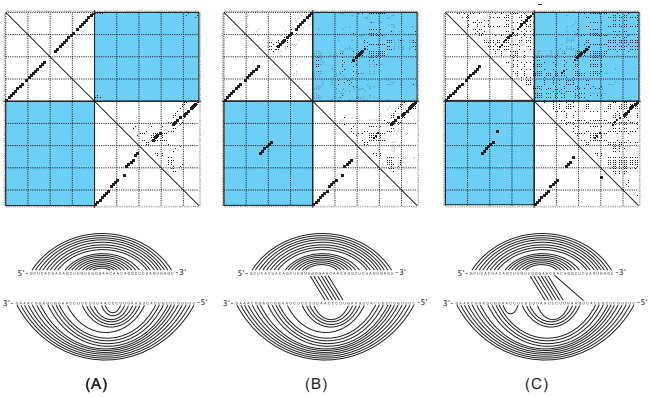

Fig. 17. `rip` versus `RNAcofold`: dot plots (top) and diagrams (bottom) of joint structures of the two RNA molecules GCUCACUACAGCUUGUUGGGAACAACAGGCUCUAGUGAGU and GAAACGGAGCGGAACCUCUUUUAACCCUUGAAGUCACUGCCCGUUUC as predicted by `RNAcofold` (Bernhart *et al.*, 2006) (left) and `rip` with parameters ($\sigma_0 = 5.1 kcal/mol$, $\sigma = 0.92$) used by (Chitsaz *et al.*, 2009) (middle) and be rescaled by a global $\epsilon = 4.1 kcal/mol$ and $\sigma_0 = 0$, $\sigma = 1$ (right). Dot Plot: Upper right: Base binding probability, the area of the squares is proportional to the corresponding pair probabilities. Lower left: structure predicted by the maximum weighted matching algorithm (MWM) (Cary and Stormo, 1995; Gabow, 1973), in which the base pairs are weighted by its binding probability. The two lines forming a cross indicate the cut point, intermolecular base pairs are depicted in the blue upper right and lower left rectangle.

This computation can easily be performed as a post-processing of the `rip` output.

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.

So far, very few interaction structures are known that are more complex than those computable by `RNAup`/`intaRNA`. It remains unclear, however, whether this is a correct representation of reality. It is entirely conceivable that multi-point contacts such as the one in the *OxyS*/*fhlA* system (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.

In this contribution we have focussed on the algorithmic background 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` is mostly intended as a reference implementation. Due to the immense computational costs incurred several dozens of interdependent 4-dimensional arrays, `rip` is a means of last resort for those cases where one suspects complex interaction structures beyond the reach

of simpler interaction models. 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 full implementation also serves a starting point for the exploration of further variations on the theme. One open question is the computation of "hybrid probabilities", i.e., probabilities $\mathbb{P}^{\mathsf{Hy}}_{i,j;h,\ell}$ that $R[i,j]$ and $[h,\ell]$ form an "interaction stem" or a even an entire uninterrupted interaction region. Another line of research concerns improved energy models for the more complex types of loops, possibly along the lines of Isambert and Siggia (2000).

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 15 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.

Chen, W., Qin, J. and Reidys, C. (2008) Crossings and nestings in tangled diagrams. *Electron. J. Comb.*, **15**, R86.

Chitsaz, H., Salari, R., Sahinalp, S. and Backofen, R. (2009) A partition function algorithm for interacting nucleic acid strands. In press.

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

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.

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.

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.

Jin, H., Loria, J. and Moore, P. (2007) Solution structure of an rRNA substrate bound to the pseudouridylation pocket of a box H/ACA snoRNA. *Mol. Cell*, **26**, 205–215.

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.

Waterman, M. and Smith, T. (1978) RNA secondary structure: A complete mathematical analysis. *Math. Biosci*, **42**, 257–266.

APPENDIX

A Proof of Proposition 2.1

Let $\prec_1$ be the partial order $\prec_1$ over the set of interior arcs, given by $S_{i_1}S_{j_1} \prec_1 S_{i_2}S_{j_2} \iff i_2 < i_1 < j_1 < j_2$. Similarly, let $\prec_2$ denote the partial order over the set of exterior arcs $R_{i_1}S_{j_1} \prec_2 R_{i_2}R_{j_2} \iff i_1 < i_2,\ j_1 < j_2$.

Let R_iS_j be the maximal (rightmost) exterior arc of $J_{a,b,c,d}$. We consider the set of maximal R_iS_j-ancestors, M. In case of $M = \varnothing$ we immediately observe $J_{i,j;h,\ell} = R_iS_j$, i.e. $J_{i,j;h,\ell}$ is of type $\circ$. Suppose next $|M| = 1$. By symmetry we can, without loss of generality, assume $M = \{R_{i_1}R_{j_1}\}$. Let $R_{i_0}S_{j_0}$ the minimal exterior arc being an descendant of $R_{i_1}R_{j_1}$ and let j_0^* denote either the start point of the maximal $R_{i_0}S_{j_0}$ S-ancestor or set $j_0^* = j_0$ if no such ancestor exists. Then, by construction, $J_{i_1,j_1;j_0^*,j}$ is tight in $J_{a,b,c,d}$. Finally, in case of $|M| = 2$, i.e. $M = \{R_{i_1}R_{j_1}, S_{r_1}S_{s_1}\}$. We may, without loss of generality, assume that $R_{i_1}R_{j_1}$ subsumes $S_{r_1}S_{s_1}$. Again we consider the minimal descendant of $R_{i_1}R_{j_1}$, R_zS_x. Let x^* be either the start point of the maximal S-ancestor of R_zS_x or $x^* = x$, otherwise. Then $J_{i_1,j_1;x^*,s_1}$ is tight. If $R_{i_1}R_{j_1}$ is equivalent to $S_{r_1}S_{s_1}$, then $J_{i_1,j_1;r_1,s_1}$ is tight. In the above procedure we have constructed a (ts), J^*, of type $\tau \in \{\triangledown, \triangle, \square, \circ\}$ that contains the maximal exterior $J_{a,b,c,d}$-arc. By definition of (ts) and the fact that we have non-crossing arcs it follows that any other (ts) in $J_{a,b,c,d}$ is disjoint to J^*. We proceed by considering the rightmost exterior arc of $J_{a,b,c,d}$ that is not contained in J^*, concluding assertion (c) by induction on the number of exterior arcs of $J_{a,b,c,d}$. Since any exterior arc of $J_{a,b,c,d}$ is contained in a unique (ts) generated by the above procedure, (b) follows, see Fig. 18.

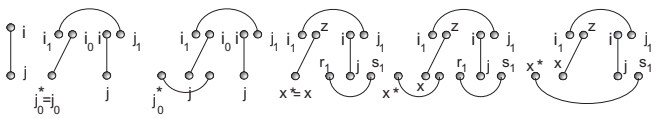

Fig. 18. Illustration of Prop.2.1.

B Derivation of equ. (2.2)-(2.4)

According to Prop. 2.1(ii), there exist unique $J_{i+1,j-1;r,s}$-tight structures $J^T_{i_1,i_2;r,r_1}$ and $J^T_{j_2,j_1;s_1,s}$ such that $J_{i_1,i_2;r,r_1} = J^T_{R_{h_1}S_{\ell_1}}$ and $J_{j_2,j_1;s_1,s} = J^T_{R_{h_2}S_{\ell_2}}$, respectively. We have the following two scenarios: in case of $J^T_{R_{h_1}S_{\ell_1}} = J^T_{R_{h_2}S_{\ell_2}}$, we have $r = s$, in which case $J^T_{i_1,j_1;r,s}$ is of type $\circ$ or in view of $S_rS_s \notin J^{\triangledown}_{i,j;r,s}$ $J_{i_1,j_1;r,s}$ is of type $\triangledown$, otherwise. In case of $J^T_{R_{h_1}S_{\ell_1}} \neq J^T_{R_{h_2}S_{\ell_2}}$, $J_{i_1,j_1;r,s}$ is a (dts) in $J_{i+1,j-1;r,s}$.

In order to arrive at equ. (2.4), we firstly observe that there exist only one $J_{i+1,j-1;r,s}$-tight structure, $J^T_{i_1,j_1;r,s}$ since $S_rS_s \in J_{i+1,j-1;r,s}$. Furthermore, consider the set M, consisting of arc that equivalent to S_rS_s. In case of $M = \varnothing$, we have $J^{\triangle}_{i_1,j_1;r,s}$ or $J^{\square}_{i_1,j_1;r,s}$, otherwise.

C Construction of the Decomposition Trees

Let us begin by giving an interpretation of Prop. 2.1.

Procedure (a) [Block Decomposition]

input: a joint structure $\vartheta_0 = J_{i,j;h,\ell}$, which is not ϑ_0-tight or a maximal secondary segment (ms).

output: a unique tree $T_a(\vartheta_0) = (V_a(T), E_a(T))$

Let $i \le j^* \le j+1$ and $R[j^*, j]$ be the ϑ_0-(ms) contain j. In particular, $j^* = j+1$ in case of such an ms does not exist and $j^* = 1$ if $R[i,j]$ itself is a (ms). Analogously, we define $S[\ell^*, \ell]$. We construct the tree $T_a(\vartheta_0)$ recursively as follows:

initialization: $V_a(T) = \{\vartheta_0\}$ and $E_a(T) = \varnothing$.

(a1): in case of $j^* = j+1$ and $\ell^* = \ell+1$, i.e. ϑ_0 is right-tight, then ϑ_0 decomposes via Prop. 2.1 (b) and (c) into a ϑ_0-tight structure $\vartheta_1 = J^{\{\triangledown,\triangle,\square,\circ\}}_{i_1,j;h_1,\ell}$ and a joint structure $\vartheta_2 = J_{i,i_1-1;h,h_1-1}$, where $i \le i_1 \le j$ and $h \le h_1 \le \ell$. Accordingly, we have

$$V_a(T) = V_a(T) \cup \{\vartheta_1, \vartheta_2\}, \tag{5.2}$$

$$E_a(T) = E_a(T) \cup \{\vartheta_0\vartheta_1, \vartheta_0\vartheta_2\}. \tag{5.3}$$

(a2) otherwise, ϑ_0 decomposes into a right-tight structure $\vartheta_3 = J^{RT}_{i,j^*-1;h,\ell^*-1}$ in ϑ_0 and two (ms) $\vartheta_4 = R[j^*, j]$, $\vartheta_5 = S[\ell^*, \ell]$. Accordingly, we have

$$V_a(T) = V_a(T) \cup \{\vartheta_3, \vartheta_4, \vartheta_5\}, \tag{5.4}$$

$$E_a(T) = E_a(T) \cup \{\vartheta_0\vartheta_3, \vartheta_0\vartheta_4, \vartheta_0\vartheta_5\}. \tag{5.5}$$

We iterate the process until all the leaves of $T_a(\vartheta_0)$ are either ϑ_0-tight structures or ϑ_0-ms.

We proceed by providing an interpretation of equ. (2.2)-(2.4).

Procedure (b): [Arc Removal and block decomposition]

input: a (ts) $\vartheta_0 = J_{i,j;h,\ell}$

output: a unique tree $T_b(\vartheta_0) = (V_b(T), E_b(T))$

initialization: $V_b(T) = \{\vartheta_0\}$ and $E_b(T) = \varnothing$.

We distinguish $J(i,j;h,\ell)$ by type:

$\circ$: do nothing.

$\square$: according to equ. (2.4), ϑ_0 decomposes into $\vartheta_1 = R_aR_b$, $\vartheta_2 = R[i+1, i_1-1]$, $\vartheta_3 = J^{\{\square,\triangle\}}_{i_1,j_1;h,\ell}$ and $\vartheta_4 = R[j_1+1, j-1]$, which gives rise to

$$V_b(T) = V_a(T) \cup \{\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4\}, \tag{5.6}$$

$$E_b(T) = E_a(T) \cup \{\vartheta_0\vartheta_1, \vartheta_0\vartheta_2, \vartheta_0\vartheta_3, \vartheta_0\vartheta_4\}. \tag{5.7}$$

$\triangledown$: according to equ. (2.2), we consider the set of $J_{i+1,j-1;h,\ell}$-tight structures, denoted by M. In case of $|M| = 1$, $J_{i+1,j-1;h,\ell}$ decompose into a sequence of a $J_{i+1,j-1;h,\ell}$-tight structure $\vartheta_6 = J^{\{\triangledown,\circ\}}_{i+1,j-1;h,\ell}$ and two $J_{i+1,j-1;h,\ell}$-ms, $\vartheta_7 = R[i+1, i_1-1]$ and $\vartheta_8 = R[j_1+1, j-1]$, where $i \le i_1 < j_1 \le j$. Accordingly,

$$V_b(T) = V_a(T) \cup \{\vartheta_1, \vartheta_6, \vartheta_7, \vartheta_8\}, \tag{5.8}$$

$$E_b(T) = E_a(T) \cup \{\vartheta_0\vartheta_1, \vartheta_0\vartheta_6, \vartheta_0\vartheta_7, \vartheta_0\vartheta_8\}. \tag{5.9}$$

In case of $|M| > 1$, $J_{i+1,j-1;h,\ell}$ decomposes into a sequence consisting of a (dts) in $J_{i+1,j-1;h,\ell}$, denoted by $\vartheta_9 = J^{DT}_{i+1,j-1;h,\ell}$ and two $J_{i+1,j-1;h,\ell}$-ms. $\vartheta_7 = R[i+1, i_1-1]$ and $\vartheta_8 = R[j_1+1, j-1]$, where $i \le i_1 < j_1 \le j$. Accordingly,

$$V_b(T) = V_a(T) \cup \{\vartheta_1, \vartheta_7, \vartheta_8, \vartheta_9\}, \tag{5.10}$$

$$E_b(T) = E_a(T) \cup \{\vartheta_0\vartheta_1, \vartheta_0\vartheta_7, \vartheta_0\vartheta_8, \vartheta_0\vartheta_9\}. \tag{5.11}$$

Furthermore, let $i_1 \le i_2 < j_1$ and $h \le j_2 < \ell$, a (dts) $\vartheta_9 = J^{DT}_{i+1,j-1;h,\ell}$ in $J_{i+1,j-1;h,\ell}$ decomposes into a $J_{i+1,j-1;h,\ell}$-tight structure $\vartheta_{10} = J^{\{\triangledown,\circ,\triangle,\square\}}_{i_1,i_2;h,j_2}$ and a right-tight joint structure $\vartheta_{11} = J^{RT}_{i_2+1,j_1;j_2+1,\ell}$ in

$J_{i+1,j-1;h,\ell}$. I.e.

$$V_b(T) = V_a(T) \cup \{\vartheta_{10}, \vartheta_{11}\}, \quad (5.12)$$

$$E_b(T) = E_a(T) \cup \{\vartheta_9\vartheta_{10}, \vartheta_9\vartheta_{11}\}. \quad (5.13)$$

$\triangle$: analogous to type $\bigtriangledown$ via symmetry.

Finally, we have the well-known (Waterman and Smith, 1978) secondary structure loop-decomposition

Procedure (c): [Secondary Structure]
input: a secondary structure $\vartheta_0 = R[i,j]$
output: a tree $T_c(\vartheta_0) = (V_c(T), E_c(T))$
initialization: $V_b(T) = \{\vartheta_0\}$ and $E_b(T) = \varnothing$.
We distinguish the following two cases:
$(c1)$: in case of $R_iR_j \notin R[i,j]$, let $\varnothing_a^b$ denote empty segment in which all the vertices are isolated. For $1 \leq j^* \leq j+1$, let $\varnothing_{j^*}^j$ be the maximal empty segment that contains R_j. In particular, if j is not isolated, we have $j^* = j+1$. Let $R^b(i_1, j^*-1)$ denote the segment in which R_{i_1} is connected with R_{j^*-1}. Then $R[i,j]$ decomposes as follows $R[i,j] = (\vartheta_1 = R[i, i_1 - 1], \vartheta_2 = R^b(i_1, j^* - 1), \vartheta_3 = \varnothing_{j^*}^j)$ and we set

$$V_c(T) = V_c(T) \cup \{\vartheta_1, \vartheta_2, \vartheta_3\}, \quad (5.14)$$

$$E_c(T) = E_c(T) \cup \{\vartheta_0\vartheta_1, \vartheta_0\vartheta_2, \vartheta_0\vartheta_3\}. \quad (5.15)$$

$(c2)$: in case of $R_iR_j \in R[i,j]$, i.e. for $R[i,j] = R^b(i,j)$, we have a decomposition into the pair $(\vartheta_4 = R_iR_j, \vartheta_5 = R[a+1, b-1])$. Accordingly, we have $V_c(T) = V_c(T) \cup \{\vartheta_4, \vartheta_5\}$ and $E_c(T) = E_c(T) \cup \{\vartheta_0\vartheta_4, \vartheta_0\vartheta_5\}$.
We iterate (c1) and (c2), until all the leaves in T are either isolated segments or single arcs.

For any joint structure, $J_{1,N;1,M}$, we can now construct a tree, with root $J_{1,N;1,M}$ and whose vertices are specific subgraphs of $J_{1,N;1,M}$. The latter are obtained by successive application of Procedure (a), (b) and (c), see Fig. 9. To be precise, let H be the graph rooted in $J_{1,N;1,M}$ defined inductively as follows: for the induction basis for fixed $J_{1,N;1,M}$ only one, Procedure (a), (b) or (c) applies. Procedure (a), (b) or (c) generates the (procedure-specific, nontrivial) subtrees, T_a, T_b and T_c. Suppose $\vartheta_\dagger$ is a leaf of T that has been constructed via Procedure (a), (b) or (c). As in case of the induction basis, each such leaf is input for exactly one procedure, which in turn generates a corresponding subtree. Prop. 2.1 and equ. (2.2-2.4) imply that H itself is a tree. We denote this decomposition tree by $T_{J_{1,N;1,M}}$, see Fig. 9.