

Target prediction and a statistical sampling algorithm for RNA-RNA interaction

Fenix W.D. Huang¹, Jing Qin¹, Christian M. Reidys^{1,2*}, and Peter F. Stadler^{3–7}

¹Center for Combinatorics, LPMC-TJKLC, Nankai University Tianjin 300071, P.R. China

²College of Life Science, Nankai University Tianjin 300071, P.R. China

³Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany.

⁴Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

⁵RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany

⁶Inst. f. Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria

⁷The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico, USA

Received on *****, revised on *****, accepted on *****

Associate Editor: *****

ABSTRACT

Motivation It has been proven that the accessibility of the target sites has a critical influence on RNA-RNA binding in general and the specificity and efficiency of miRNAs and siRNAs in particular. Recently, $O(N^6)$ time and $O(N^4)$ space dynamic programming algorithms have become available that compute the partition function of RNA-RNA interaction complexes, thereby providing detailed insights into their thermodynamic properties.

Results Modifications to the grammars underlying earlier approaches enables the calculation of interactions probabilities for any given interval on the target RNA. The direct computation of the “hybrid probabilities” is complemented by a stochastic backtracing algorithm that produces samples from the Boltzmann ensemble of RNA-RNA interaction structures. The sampling of k structures requires only negligible additional memory resources and runs in $O(k \cdot N^3)$.

Availability The algorithms described here are implemented in C as part of the `rip` package. The source code of `rip2` can be downloaded from <http://www.combinatorics.cn/cbpc/rip2.html> and <http://www.bioinf.uni-leipzig.de/Software/rip.html>.

Contact Christian Reidys duck@santafe.edu

1 INTRODUCTION

RNA-RNA binding is a major mode of action of various classes of non-coding RNAs and plays a crucial role in many regulatory processes in all living organisms. 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 (Bachelierie *et al.*, 2002), insertion editing (Benne, 1992), and transcriptional control (Kugel

and Goodrich, 2007). Emerging evidence suggests, furthermore, that RNA-RNA interactions also play a role for the functionality of long mRNA-like ncRNAs (Hekimoglu and Ringrose, 2009). A common theme in many RNA classes, including miRNAs, snRNAs, gRNAs, snoRNAs, and in particular many of the prokaryotic small RNAs is the formation of RNA-RNA interaction structures that are much more complex than simple complementary sense-antisense interactions. Thermodynamically, the binding of two RNA molecules A and B can be described by the binding energy $\Delta G^{\text{bind}} = G_{AB} - G_A - G_B$, i.e., by the difference of the energy of structure formation G_{AB} of the AB complex and the folding energies G_A and G_B of the two individual RNAs A and B . Thus, the binding or hybridization energy has been widely used as a criterion to predict RNA-RNA interactions (Rehmsmeier *et al.*, 2004; Tjaden *et al.*, 2006; Busch *et al.*, 2008).

The interaction between two RNAs is governed by the same physical principles that determine RNA folding: the formation of specific base pairing patterns whose energy is largely determined by base pair stacking and loop strains. Secondary structures, therefore, are an appropriate level of description to quantitatively understand the thermodynamics of RNA-RNA binding. Just as the general RNA folding problem with unrestricted pseudoknots (Akutsu, 2000), the RNA-RNA interaction problem (RIP) is NP-complete in its most general form (Alkan *et al.*, 2006; Mneimneh, 2007). 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 simplest approach concatenates the two interacting sequences and subsequently employs a slightly modified standard secondary structure folding algorithm. The software tools `RNACoFold` (Hofacker *et al.*, 1994; Bernhart *et al.*, 2006), `pairfold` (Andronescu *et al.*, 2005), and `NUPACK` (?) subscribe to this strategy. The main problem of this approach is that it cannot predict important motifs such as kissing-hairpin loops. The paradigm of concatenation has also

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

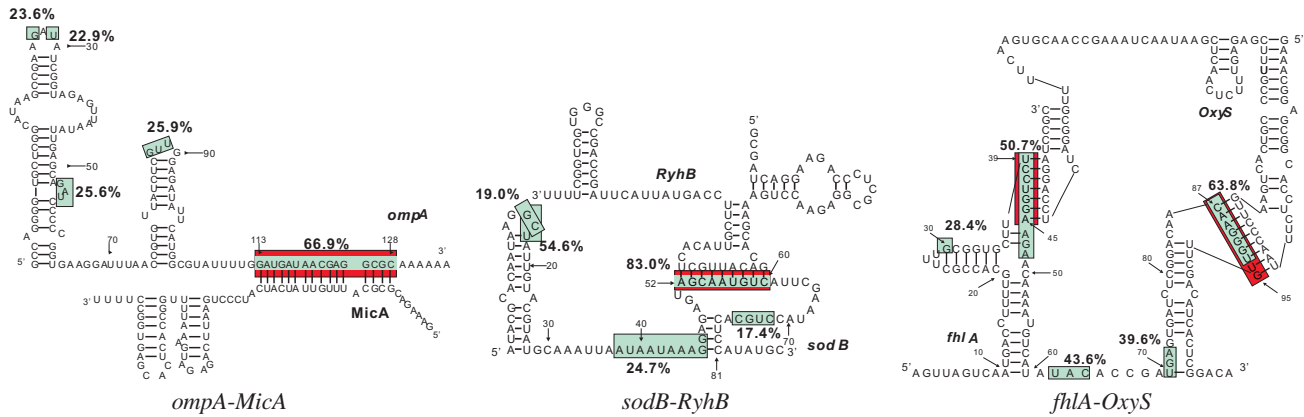


Fig. 1. Examples of RNA-RNA interactions structures. The primary interaction region(s) are highlighted in red in the experimentally supported structural models from the literature: *ompA-MicA*: Udekwu *et al.* (2005); *sodB-RyhB*: Geissmann and Touati (2004); *fhlA-OxyS*: Argaman and Altuvia (2000). Hybridization probabilities computed by `rip2` are annotated by green boxes for region with a probability larger than 10%. In many cases, the computational predictions identify additional hybridization regions that may further stabilize the interaction.

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.*, 2009b; Qin and Reidys, 2008). An alternative line of thought, implemented in `RNAduplex` and `RNAhybrid` (Rehmsmeier *et al.*, 2004), is to neglect all internal base-pairings in either strand, i.e., to compute the minimum free energy (MFE) secondary structure of hybridization of otherwise unstructured RNAs. `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. As a special case, `snoRNA/target` complexes are treated more efficiently using a specialized tool (Tafer *et al.*, 2009) due to the highly conserved interaction motif. Algorithmically, the approaches mentioned so far are close relatives of the RNA folding recursions Zuker and Sankoff (1984).

A different approach was taken independently by Pervouchine (2004) and Alkan *et al.* (2006), who proposed MFE folding algorithms for predicting the *joint structure* of two interacting RNA molecules. In this model, “joint structure” means that the intramolecular structures of each partner is pseudoknot-free, that the intermolecular binding pairs are noncrossing, and that there is no so-called “zig-zag” configuration (see below for details). The optimal joint structure can be computed in $O(N^6)$ time and $O(N^4)$ space by means of dynamic programming. More recently, extensions to the partition function were proposed by Chitsaz *et al.* (2009b) (`piRNA`) and Huang *et al.* (2009) (`rip1`). In contrast to the RNA folding problem, where minimum energy folding and partition functions can be obtained by very similar algorithms, this is much more complicated for joint structures. The reason is that simple unambiguous grammars are known for RNA secondary structures (Dowell and Eddy, 2004), while the disambiguation of grammar underlying the Alkan-Pervouchine algorithm requires the introduction of a large number of additional non-terminals (which algorithmically translate into additional dynamic programming tables). Although the partition function of joint structures can be computed in $O(N^6)$ time and $O(N^4)$ space, the current implementations require very large computational resources. Salari *et al.* (2009) recently achieved a substantial speed-up making use of

the observation that the external interactions mostly occur between pairs of unpaired regions of single structures. Chitsaz *et al.* (2009a), on the other hand, use tree-structured Markov Random Fields to approximate the joint probability distribution of multiple (≥ 3) contact regions.

The binding energies provides a useful over-all characterization of an RNA-RNA interaction. In many cases, however, the locations of the intermolecular base pairs and the detailed structure of the interaction complex is of crucial importance. Bacterial sRNAs, for example may either up- or down-regulate mRNA translation depending on the structural changes induced by the interaction (Urban and Vogel, 2007). In particular in RNA-RNA complexes with multiple interaction sites, i.e., in the class of structures for which the expensive computation of joint structures is necessary, Fig. 1, one is interested in the probabilities of hybridization in individual regions and in the interdependencies of alternative conformations.

We therefore extend our previous framework in two directions: (1) A modification of the underlying grammar explicitly treats hybrids, i.e., maximal regions with exclusively intermolecular interactions. This allows us to investigate local aspects in much more detail. (2) A stochastic backtracing algorithm, in analogy to similar approaches for RNA secondary structure prediction (Tacker *et al.*, 1996; Ding and Lawrence, 2003), which can be used to produce representative structure and to generate samples from the thermodynamic properties. These samples can be useful to assess complex structural features for which it would be too tedious or expensive to design and implement dedicated exact backtracing algorithms.

2 THE HYBRID-PARTITION FUNCTION

2.1 Some basic facts

We briefly review some basic concepts and outline the notation introduced in Huang *et al.* (2009). Full details are given in the Supplement Material (SM).

Given two RNA sequences $R = (R_i)_1^N$ and $S = (S_j)_1^M$ (e.g. an antisense RNA and its target or an mRNA and its sRNA regulator) with N

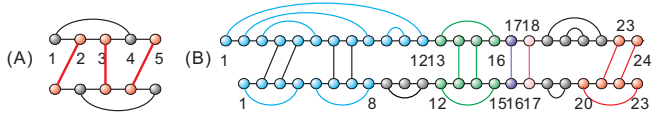


Fig. 2. A. A zig-zag, generated by R_2S_1 , R_3S_3 and R_5S_4 (red). B. the joint structure $J_{1,24;1,23}$, we color the different segments and tight structures in which $J_{1,24;1,23}$ decomposes.

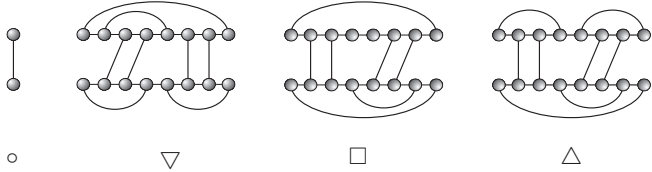


Fig. 3. The four basic types of tight.

and M vertices, we label the vertices such that R_1 is the 5' end of R and S_1 denotes the 3' end of S . The arcs of R and S then represent the respective, intramolecular base pairs. An arc is called *exterior* if it is of the form R_iS_j and *interior*, otherwise.

Next we formally define joint structures (Pervouchine, 2004; Alkan *et al.*, 2006; Chitsaz *et al.*, 2009b; Huang *et al.*, 2009). A *joint structure*, $J(R, S, I)$, see Fig. 2B, is a graph such that

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 exterior arcs without external pseudoknots, i.e., if $R_{i_1}S_{j_1}, R_{i_2}S_{j_2} \in I$ then $i_1 < i_2$ implies $j_1 < j_2$;
3. $J(R, S, I)$ contains no “zig-zags”, see Fig. 2A;

where a “zig-zag” is defined as follows: suppose there is an exterior arc R_aS_b with R_iR_j and $S_{i'}S_{j'}$, where $i < a < j$ and $i' < b < j'$. Then R_iR_j is *subsumed* in $S_{i'}S_{j'}$, if for any $R_kS_{k'} \in I$, $i < k < j$ implies $i' < k' < j'$. 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 (Fig. 2). Dependence here means that there exists at least one exterior arc R_hS_ℓ such that $i_1 < h < j_1$ and $i_2 < \ell < j_2$.

The (*induced*) *subgraph* of G induced by V has vertex set V and contains all G -edges having both incident vertices in V . The subgraph of a joint structure $J(R, S, I)$ induced by a pair of subsequences $(R_i, R_{i+1}, \dots, R_j)$ and $(S_h, S_{h+1}, \dots, S_\ell)$ is denoted by $J_{i,j;h,\ell}$. In particular, $J(R, S, I) = J_{1,N;1,M}$ and $J_{i,j;h,\ell} \subset J_{a,b;c,d}$ if and only if $J_{i,j;h,\ell}$ is a subgraph of $J_{a,b;c,d}$ induced by (R_i, \dots, R_j) and (S_h, \dots, S_ℓ) . In particular, we use $S[i, j]$ to denote the subgraph of the pre-structure $G(R, S, I)$ induced by $(S_i, S_{i+1}, \dots, S_j)$, where $S[i, i] = S_i$ and $S[i, i-1] = \emptyset$. Given a joint structure, $J_{a,b;c,d}$, its tight structure (TS) $J_{a',b';c',d'}$ 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 $c' \neq d'$.

In the following, a TS is denoted by $J_{i,j;h,\ell}^T$. If $J_{a',b';c',d'}$ is tight in $J_{a,b;c,d}$, 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. 3:

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

$$\nabla : R_iR_j \in J_{i,j;h,\ell}^\nabla \text{ and } S_hS_\ell \notin J_{i,j;h,\ell}^\nabla;$$

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

$$\triangle : S_hS_\ell \in J_{i,j;h,\ell}^\triangle \text{ and } R_iR_j \notin J_{i,j;h,\ell}^\triangle.$$

2.2 The hybrid-grammar

A *hybrid* structure, $J_{i_1,i_\ell;j_1,j_\ell}^{\text{Hy}}$, is a maximal sequence of intermolecular interior loops consisting of exterior arcs $(R_{i_1}S_{j_1}, \dots, R_{i_\ell}S_{j_\ell})$ 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 only. That is, a hybrid is the maximal unbranched stem-loop formed by external arcs. Each hybrid thus forms a distinctive region of interaction between the two RNAs. Note that we can interpret interactions admitted by `intarna/RNAup` (Busch *et al.*, 2008; Mückstein *et al.*, 2008) as joint structures with at most one hybrid.

In the following we re-design the grammar outlined by Huang *et al.* (2009) so that it explicitly makes use of hybrids. An efficient solution of the partition function problem for RIP requires an unambiguous context free grammar with the constraint that the number of break points, i.e., the number of non-terminals in each individual production, is as small as possible. This is achieved by introducing several specific types of joint structures that are described in detail in the following. We call a joint structure *right-tight* (RTS), $J_{i_1,j_1;r_1,s_1}^{RT}$ in $J_{i_1,j_1;r_1,s_1}$ if its rightmost block is a $J_{i_1,j_1;r_1,s_1}$ -TS and *double-tight* (DTS), $J_{i_1,j_1;r_1,s_1}^{DT}$ in $J_{i_1,j_1;r_1,s_1}$ if both of its leftmost and its rightmost blocks are $J_{i_1,j_1;r_1,s_1}$ -TS's. We remark that this definition is a bit different from the notion of the double-tight structured defined in Huang *et al.* (2009). In particular, we consider single interaction arcs as particular DTS. Adopting the point of view of Algebraic Dynamic Programming (Giegerich and Meyer, 2002) we regard each decomposition rule as a production in a suitable grammar. Fig. 4 summarizes the three basic steps of the hybrid-grammar: (I) “interior arc-removal” to reduce TS. The scheme is complemented by the usual loop decomposition of secondary structures, and (II) “block-decomposition” to split a joint structure into two smaller blocks.

The grammar in Fig. 4 corresponds to the decomposition (parsing) of a joint structure into interior arcs and hybrids. Figure 5A shows the corresponding parse tree. The full details of the decomposition procedures are described in Section 2 of the SM, where we show 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 substructure of $J_{1,N;1,M}$ obtained by the successive application of the decomposition steps of Fig. 4 and the loop decomposition of the secondary structures. Thus, the hybrid grammar is unambiguous. The two panels of Fig. 5 contrast the grammars of `rip1` (Huang *et al.*, 2009) and the hybrid-grammar of `rip2` introduced here. In `rip1`, hybrids were immediately decomposed into individual external base-pairs and their associated interior loops, so that individual hybrids were not tractable in a straightforward manner.

Let us now have a closer look at the energy evaluation of $J_{i,j;h,\ell}$. Each decomposition step in Fig. 4 results in substructures whose energies are assumed to contribute additively and generalized loops that can be evaluated directly. There are the following two basic scenarios:

I. Interior Arc removal. The first type of decomposition is derived from the decomposition of TS of Huang *et al.* (2009). Most of the decomposition operations in Procedure (b) displayed in Fig. 4 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 Sankoff (1984) 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 .

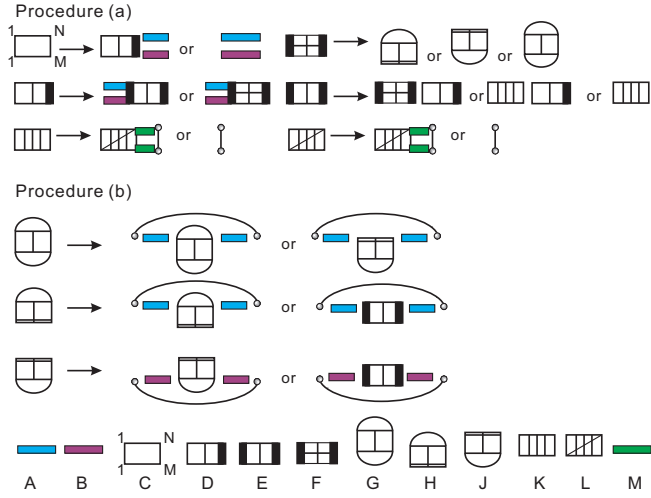


Fig. 4. Illustration of the reduction of arbitrary joint structures and of right-tight structures, Procedure (a), and of tight structures, Procedure (b). In the bottom row the symbols for the 10 distinct types of structural components are listed: **A, B**: maximal secondary structure segments $R[i, j]$, $S[r, s]$; **C**: arbitrary joint structure $J_{1, N; 1, M}$; **D**: right-tight structures $J_{i, j; r, s}^{RT}$; **E**: double-tight structure $J_{i, j; r, s}^{DT}$; **F** tight structure of type ∇ , \triangle or \square ; **G**: type \square tight structure $J_{i, j; r, s}^{\square}$; **H**: type ∇ tight structure $J_{i, j; r, s}^{\nabla}$; **J**: type \triangle tight structure $J_{i, j; r, s}^{\triangle}$; **K**: hybrid structure $J_{i, j; h, \ell}^{Hy}$; **L**: substructure of a hybrid $J_{i, j; h, \ell}^h$ such that $R_i S_j$ and $R_h S_\ell$ are exterior arcs and $J_{i, j; h, \ell}^h$ itself is not a hybrid since it is not maximal; **M**: isolated segment $R[i, j]$ or $S[h, \ell]$.

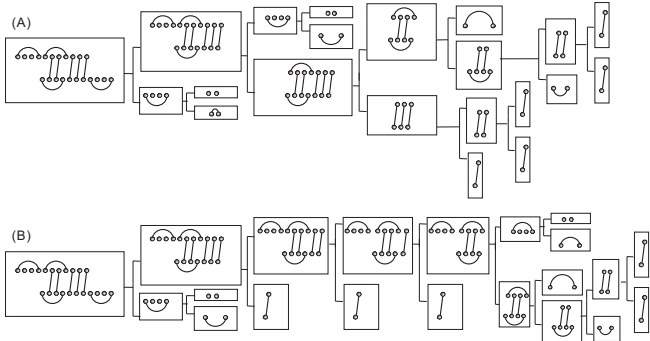


Fig. 5. Different grammars lead to different (parse) trees. We show the parse tree $T_{J_{1,11;1,11}}$ for the same joint structure $J_{1,11;1,11}$ according to the grammars of rip2 (top) and rip1 (bottom), respectively.

The set of base pairs on $R[i, j]$ consists of all interior pairs $R_p R_q$ with $i \leq p < q \leq j$ and all exterior pairs $R_p S_h$ 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 $R[i + 1, j - 1]$ correspond to the base pairs *immediately interior* of $R_i R_j$. Let us write $E_{R[i, j]} = E_{R[i, j]}^i \cup E_{R[i, j]}^e$ 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: $E_R = \emptyset$, hairpin loop; $E_R = E_R^i$ and $|E_R| = 1$, interior loop (including bulge and stacks); $E_R = E_R^i$, $|E_R| \geq 2$, multi-branch loop; $E_R = E_R^e$, kissing-hairpin loop; $|E_R^i|, |E_R^e| \geq 1$, general kissing-loop.

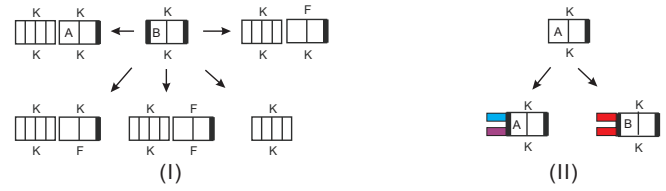


Fig. 6. Decomposition of $J_{i, j; h, \ell}^{DT, KKB}$ (l.h.s.) and $J_{i, j; h, \ell}^{RT, KKA}$ (r.h.s.).

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 $J_{i, j; h, \ell}^{\nabla}$ is of type E, if $S[h, \ell]$ is not enclosed in any base pair ($J_{i, j; h, \ell}^{\nabla, E}$). Suppose $J_{i, j; h, \ell}^{\nabla}$ is located immediately interior to the closing pair $S_p S_q$ ($p < h < \ell < q$). If the loop closed by $S_p S_q$ is a multiloop, then $J_{i, j; h, \ell}^{\nabla}$ is of type M ($J_{i, j; h, \ell}^{\nabla, M}$). If $S_p S_q$ is contained in a kissing-loop, we distinguish the types F and K, depending on whether or not $E_{S[h, \ell]}^e = \emptyset$.

Analogously, there are in total four types of a hybrid $J_{i, j; h, \ell}^{Hy}$, i.e. $\{J_{i, j; h, \ell}^{Hy, EE}, J_{i, j; h, \ell}^{Hy, EK}, J_{i, j; h, \ell}^{Hy, KE}, J_{i, j; h, \ell}^{Hy, KK}\}$.

II. Block decomposition. The second type of decomposition is the splitting of joint structures into “blocks”. Here the hybrid grammar differs from the grammar of Huang *et al.* (2009) in two ways. First, we use the hybrid as a new block of the grammar, decomposing a hybrid by removing its exterior arcs in parallel simultaneously starting from the right. Second, we split a joint structure into blocks via alternating decompositions of RTS and DTS as shown in the Procedure (a) of Fig. 4.

In order to guarantee the maximality hybrids, we observe that the RTS’s $J_{i, j; h, \ell}^{RT, KK}$, $J_{i, j; h, \ell}^{RT, KE}$, $J_{i, j; h, \ell}^{RT, EK}$ and $J_{i, j; h, \ell}^{RT, EE}$ can appear in two scenarios, depending on whether or not there exists an exterior arc $R_{i_1} S_{j_1}$ such that $R[i, i_1 - 1]$ and $S[j_1 - 1, j]$ are isolated segments. In case such an exterior arc exists, we say the RTS is of type (B) or (A), otherwise. Similarly, a DTS, $J_{i, j; h, \ell}^{RT, KK}$, $J_{i, j; h, \ell}^{RT, KE}$, $J_{i, j; h, \ell}^{RT, EK}$ or $J_{i, j; h, \ell}^{RT, EE}$ is of type (B) or (A) depending on whether $R_i S_h$ is an exterior arc. In Fig. 6 (I) we display the decomposition of $J_{i, j; h, \ell}^{DT, KKB}$ into hybrids and RTS of type (A) and in Fig. 6 (II) we display the decomposition of $J_{i, j; h, \ell}^{RT, KKA}$ into secondary structure segments and DTS accordingly.

Suppose $J_{i, j; r, \ell}^{DT}$ is a DTS contained in a kissing loop, that is we have either $E_{R[i, j]}^e \neq \emptyset$ or $E_{S[h, \ell]}^e \neq \emptyset$. Without loss of generality we may assume $E_{R[i, j]}^e \neq \emptyset$. Then at least one of the two “blocks” contains at least an exterior arc belonging to $E_{R[i, j]}^e$ labeled by K or F, otherwise, see Fig. 6 (I).

2.3 Forward Recursions

The computation of the partition function proceeds “from the inside to the outside”, see eq. (2.3). 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 $Q_{i, j}^m$ and $Q_{i, j}^k$. Here, $Q_{i, j}^m$ 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, $Q_{i, j}^k$ 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. Let $\mathbb{J}_{i, j; h, \ell}^{\xi, Y_1 Y_2 Y_3}$ be the set of substructures $J_{i, j; h, \ell} \subset J_{1, N; 1, M}$, induced from some joint-structure $J_{1, N; 1, M}$, such that $J_{i, j; h, \ell}$ appears in $T_{J_{1, N; 1, M}}$ as an interaction structure of type $\xi \in \{DT, RT, \nabla, \triangle, \square, \circ\}$ with loop-subtypes $Y_1, Y_2 \in \{M, K, F\}$ on the sub-intervals $R[i, j]$ and $S[h, \ell]$, $Y_3 \in \{A, B\}$. Let $Q_{i, j; h, \ell}^{\xi, Y_1 Y_2 Y_3}$ denote the partition function of the set

$J_{i,j;h,\ell}^{\xi, Y_1 Y_2 Y_3}$. All recursions for $Q_{i,j;h,\ell}^{\xi, Y_1 Y_2 Y_3}$ represent a reformulation of the hybrid-grammar specified in Fig. 4.

For instance, the recursion for $Q_{i,j;h,\ell}^{DT, KKB}$ displayed in Figure 6(I) is given by:

$$\begin{aligned} Q_{i,j;h,\ell}^{DT, KKB} &= \sum_{i_1, h_1} Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{RT, KKA} + Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{RT, KF} \\ &+ Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{RT, FF} + Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{RT, FK} + Q_{i, j; h, \ell}^{\text{Hy, KK}}, \end{aligned} \quad (2.1)$$

where the corresponding recursion for $Q_{i,j;h,\ell}^{\text{Hy, KK}}$ is

$$Q_{i,j;h,\ell}^{\text{Hy, KK}} = \sum_{i_1, h_1} Q_{i, i_1; h, h_1}^{\text{Hy, KK}} e^{-(\sigma_0 + \sigma G_{i_1, h_1, j, \ell}^{\text{int}} + (j + \ell - i_1 - h_1 - 2)\beta_3)}. \quad (2.2)$$

Analogously, the recursions for $Q_{i,j;h,\ell}^{\text{Hy, EE}}$, $Q_{i,j;h,\ell}^{\text{Hy, EK}}$ and $Q_{i,j;h,\ell}^{\text{Hy, KE}}$ read:

$$\begin{aligned} Q_{i,j;h,\ell}^{\text{Hy, EE}} &= \sum_{i_1, h_1} Q_{i, i_1; h, h_1}^{\text{Hy, EE}} e^{-(\sigma_0 + \sigma G_{i_1, h_1, j, \ell}^{\text{int}})}; \\ Q_{i,j;h,\ell}^{\text{Hy, EK}} &= \sum_{i_1, h_1} Q_{i, i_1; h, h_1}^{\text{Hy, EK}} e^{-(\sigma_0 + \sigma G_{i_1, h_1, j, \ell}^{\text{int}} + (\ell - h_1 - 1)\beta_3)}; \\ Q_{i,j;h,\ell}^{\text{Hy, KE}} &= \sum_{i_1, h_1} Q_{i, i_1; h, h_1}^{\text{Hy, KE}} e^{-(\sigma_0 + \sigma G_{i_1, h_1, j, \ell}^{\text{int}} + (j - i_1 - 1)\beta_3)}. \end{aligned} \quad (2.3)$$

2.4 Hybrid-probabilities

Since the probabilities of individual base pairs are not independent, it is not possible to compute the probabilities for particular hybrids directly from them. Hybrid probabilities thus cannot be obtained in a simple way from the backward recursions described by Huang *et al.* (2009).

Given two RNA sequences, our notion of probability is based on the ensemble of all possible joint interaction structures. Let Q^I denote the partition function of all these joint structures that can be formed by two input RNA sequences. The probability of a fixed joint structure $J_{1, N; 1, M}$ is given by

$$\mathbb{P}_{J_{1, N; 1, M}} = Q_{J_{1, N; 1, M}} / Q^I. \quad (2.4)$$

In difference to the computation of the hybrid-partition function “from the inside to the outside” (IO), the computation of probabilities of specific substructures is obtained “from the The same principle applies to the computation of base pairing computation of base pairing probabilities of secondary structures (McCaskill, 1990) and joint structures Huang *et al.* (2009).

Let $J = J_{1, N; 1, M}$, with associated decomposition tree $T(J)$ and let $\Lambda_{i,j;h,\ell} = \{J \mid J_{i,j;h,\ell} \in T(J)\}$ denote the set of all joint structures J such that $J_{i,j;h,\ell}$ is contained in the decomposition tree $T(J)$. Then we have, by construction,

$$\mathbb{P}_{J_{i,j;h,\ell}} = \sum_{J \in \Lambda_{i,j;h,\ell}} \mathbb{P}_J. \quad (2.5)$$

Following the (OI)-paradigm, the probability of a parent structure, \mathbb{P}_{θ_s} , is computed prior to the calculation of $\mathbb{P}_{J_{i,j;h,\ell}}$. The conditional probability $\mathbb{P}_{J_{i,j;h,\ell} \mid \theta_s}$ equals $Q_{\theta_s}(J_{i,j;h,\ell}) / Q(\theta_s)$, where $Q(\theta_s)$ is the partition function of θ_s , and $Q_{\theta_s}(J_{i,j;h,\ell})$ the partition function of all those θ_s , that have in addition $J_{i,j;h,\ell}$ as a child in their parse trees. Consequently, $\mathbb{P}_{J_{i,j;h,\ell}}$ can inductively be computed by summing over all probabilities \mathbb{P}_{θ_s} , i.e.

$$\mathbb{P}_{J_{i,j;h,\ell}} = \sum_{\theta_s} \mathbb{P}_{J_{i,j;h,\ell} \mid \theta_s} \mathbb{P}_{\theta_s} = \sum_{\theta_s} [Q_{\theta_s}(J_{i,j;h,\ell}) / Q(\theta_s)] \mathbb{P}_{\theta_s}. \quad (2.6)$$

Let $\mathbb{P}_{i,j;h,\ell}^{\text{Hy}}$ denote the probability of the set of substructures J such that the specific hybrid-substructure, $J_{i,j;h,\ell}^{\text{Hy}}$, appears in the decomposition tree $T(J)$, i.e. $J_{i,j;h,\ell}^{\text{Hy}} \in T(J)$. Since each joint-structure $J_{i,j;h,\ell}^{\text{Hy}}$ is either one

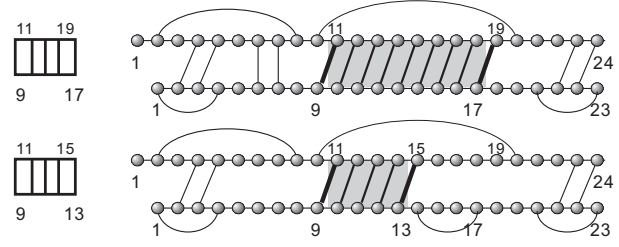


Fig. 7. Hybrid-probability: the maximality of hybrids implies that although the intervals $[h_1, l_1]$ and $[h_2, l_2]$ overlap, distinct hybrids (grey) are being induced.

of the four types $J_{i,j;h,\ell}^{\text{Hy, EE}}$, $J_{i,j;h,\ell}^{\text{Hy, EK}}$, $J_{i,j;h,\ell}^{\text{Hy, KE}}$ or $J_{i,j;h,\ell}^{\text{Hy, KK}}$, we arrive at

$$\mathbb{P}_{i,j;h,\ell}^{\text{Hy}} = \mathbb{P}_{i,j;h,\ell}^{\text{Hy, EE}} + \mathbb{P}_{i,j;h,\ell}^{\text{Hy, EK}} + \mathbb{P}_{i,j;h,\ell}^{\text{Hy, KE}} + \mathbb{P}_{i,j;h,\ell}^{\text{Hy, KK}}. \quad (2.7)$$

We remark that, by construction, for $[h_1, l_1] \neq [h_2, l_2]$ the hybrid probabilities $\mathbb{P}_{i,j;h_1, l_1}^{\text{Hy}}$ and $\mathbb{P}_{i,j;h_2, l_2}^{\text{Hy}}$ quantify disjoint classes of joint structures. This is a consequence of the maximality of hybrids, which implies that, for fixed interval $[i, j]$, each $[h_1, l_1]$ corresponds to a unique hybrid $J_{i,j;h_1, l_1}^{\text{Hy}}$. Based on the notion of hybrid probability, we can introduce

$$\mathbb{P}_{[i,j]}^{\text{target}} = \sum_{h,\ell} \mathbb{P}_{i,j;h,\ell}^{\text{Hy}}, \quad (2.8)$$

which is, according to the above, the probability of the target site $[i, j]$ and furthermore

$$\pi_R(i) = \sum_{p,q} \sum_{p \leq i \leq q} \mathbb{P}_{p,q;h,\ell}^{\text{Hy}}, \quad (2.9)$$

measuring, for each base i in R the probability that i is contained in a hybrid. A particularly instructive observable is the interaction base pairing matrix, given by

$$\pi_{i,k} = \sum_{p,q} \sum_{p \leq i \leq q} \sum_{r,s} \mathbb{P}_{p,q;r,s}^{\text{Hy}}. \quad (2.10)$$

Clearly, $\pi_{i,k}$ measures the probability that a pair of nucleotides (i, k) , located on different strands, is contained in an interaction region. In contrast to the base pairing probabilities, large values of $\pi_{i,k}$ do not imply that i and k actually form an exterior base pair. Instead, it highlights regions of intermolecular interactions.

2.5 Boltzmann sampling

A dynamic programming scheme for the computation of a partition function gives implies a corresponding stochastic backtracing procedure that can be used to sample from the associated distribution Tacker *et al.* (1996). The usefulness of this approach for RNA secondary structures is discussed by Ding and Lawrence (2003). The same ideas can of course also produce representative samples from the Boltzmann equilibrium distribution of RNA interaction structures, Fig. 8.

The basic data structure of the algorithm is a stack \mathcal{A} that stores tuples of the form $\{(i, j, h, \ell; \xi)\}$ describing a pair of intervals $[i, j]$ in R and $[h, \ell]$ in S and the type ξ of the – not further specified – joint structure formed by the two intervals. The stack \mathcal{A} , initialized with $(1, N; 1, M, ?)$ where “?” denotes the unspecified type, guides the backtracing which is complete as soon as \mathcal{A} is empty. A list \mathcal{L} is used to collect the interior and exterior arcs and unpaired bases generated by the decompositions and eventually define the sampled interaction structure. In the first step, $(1, N; 1, M, ?)$ is decomposed according to the grammar in Fig. 4 into either (1) a pair of secondary structures, or (2) a RTS $(i, N; j, M; RT\text{EE})$ with probabilities derived as explained above. Depending on the stochastic choice we push either (1) $(1, N; 0, 0; \text{sec})$ and $(0, 0; 1, M; \text{sec})$ or (2) $(1, i-1; 0, 0; \text{sec})$, $(0, 0; 1, j-1; \text{sec})$ and $(i, N; j, M; RT\text{EE})$ into the stack \mathcal{A} .

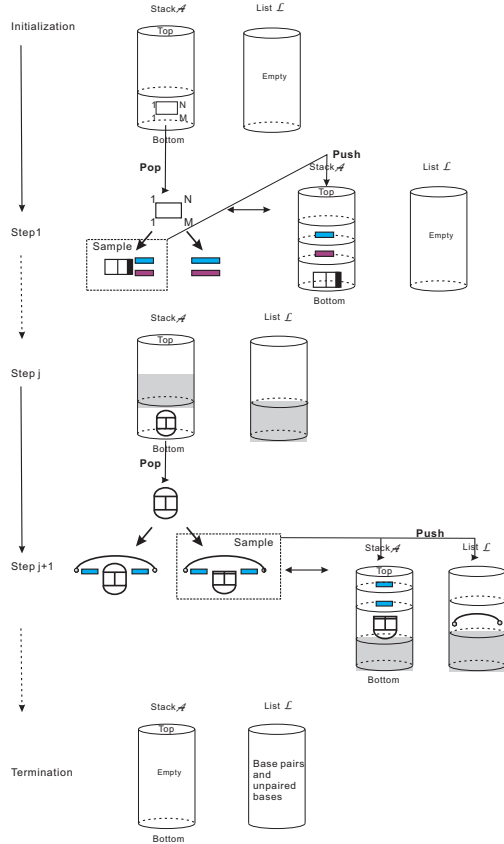


Fig. 8. Stochastic backtracing algorithm: elements of stack \mathcal{A} are successively decomposed according to the hybrid-grammar. The resulting arcs and unpaired vertices are stored in the list \mathcal{L} which, once \mathcal{A} is empty, eventually contains the Boltzmann-sampled interaction structure.

Given \mathcal{A} and \mathcal{L} , we can associate a probability by considering the decomposition of the particular type of joint structure. For instance, suppose we have extracted $(i, j; h, \ell, DT/KKB)$ from stack \mathcal{A} , see Fig. 6. Then the probabilities for continuing with one of the five decompositions displayed in Fig. 6, for each position of the break points $i_1 \in [i, j]$ and $j_1 \in [h, \ell]$, is given by

$$\begin{aligned} \mathbb{P}_{i_1, j_1}^0 &= Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{\text{RT, KKA}} / Q_{i, j; h, \ell}^{\text{DT, KKB}}, \\ \mathbb{P}_{i_1, j_1}^1 &= Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{\text{RT, KF}} / Q_{i, j; h, \ell}^{\text{DT, KKB}}, \\ \mathbb{P}_{i_1, j_1}^2 &= Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{\text{RT, FF}} / Q_{i, j; h, \ell}^{\text{DT, KKB}}, \\ \mathbb{P}_{i_1, j_1}^3 &= Q_{i, i_1; h, h_1}^{\text{Hy, KK}} Q_{i_1+1, j; h_1+1, \ell}^{\text{RT, FK}} / Q_{i, j; h, \ell}^{\text{DT, KKB}}, \\ \mathbb{P}_{i_1, j_1}^4 &= Q_{i, j; h, \ell}^{\text{Hy, KK}} / Q_{i, j; h, \ell}^{\text{DT, KKB}}. \end{aligned}$$

One of these decompositions is accordingly sampled and the respective output is pushed back into stack \mathcal{A} . For instance, if \mathbb{P}_{i_1, j_1}^1 is selected, then we push $(i, i_1; h, h_1; \text{HyKK})$ and $(i_1 + 1, j; h_1 + 1, \ell; \text{RTKF})$ back into stack \mathcal{A} .

3 RESULTS AND CONCLUSIONS

We presented here a modified and improved unambiguous grammar for the RNA-RNA interaction problem. Compared to our previous `rip1` approach Huang *et al.* (2009), it reduces the computational

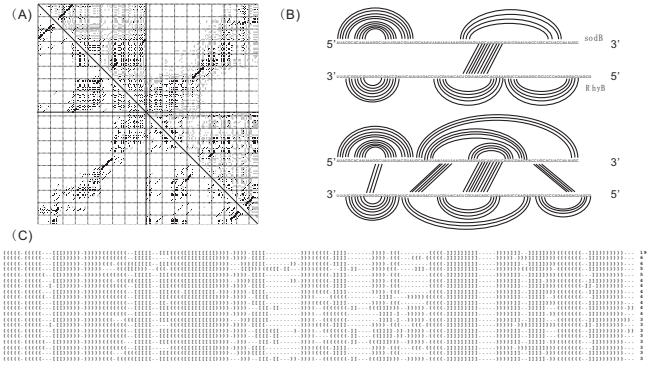


Fig. 9. Interaction of *sodB-RhyB*. (A) Base pairing probability matrix. The upper-right triangle shows the probabilities obtained from the exact backwards recursion, the lower-left triangle is the estimate from a sample of 10,000 structures obtained by stochastic backtracing, showing that the estimates converge quickly. (B) Comparison of the structure proposed in Geissmann and Touati (2004) and the `rip2` prediction. While the major stable hairpins agree and `rip2` correctly predicts the primary interaction region, `rip2` also identifies additional interaction regions that may stabilize the interaction. (C) Sampled joint structures (here the 20 most frequent ones) are represented as dot-bracket strings: $()$ and $[\]$ represent pairs of interior and exterior arc, respectively, while dots indicate unpaired bases. $|$ separates the two RNA sequences which are both written in $5' \rightarrow 3'$ direction.

efforts, and in particular the memory consumption, by about a third. The complete set of recursions is compiled in section 3 of the SM. It comprises 9 4D-arrays $Q_{i, j; r, s}^{\Delta, \nabla, \square}$ for tight structures of various types, 20 4D-arrays $Q_{i, j; r, s}^{\text{RT}}$ for right-tight structures, and 20 4D-arrays $Q_{i, j; r, s}^{\text{DT}}$ for double-tight structures. The implementation has been extended by a stochastic backtracing facility. Fig. 9 gives an example of some of the output than is produced by `rip2`, see also SM, Fig. 4. Despite algorithmic improvements, `rip2` still requires quite substantial computational resources for practical applications and is in practise limited to problem sizes of $N_1 + N_2 \lesssim 250$ on current hardware. Still not a tool large-scale routine applications, `rip2` is however suitable for investigating the fine details of particular interactions. Future work will thus focus on controlled approximations with the aim of a drastic reduction of both: CPU and memory consumption.

The major advantage of stochastic sampling is that it provides a generic and convenient means to estimate quantities that cannot be easily computed directly by backwards recursion (Ding and Lawrence, 2003). Both the *ompA-MicA* and the *sodB-RhyB* complexes show a primary, highly likely, hybrid region and several additional less stable points of contact, as shown by their “interaction maps” in Fig. 10. In these examples it is of interest to investigate in detail how the putative interaction regions influence each other: is the binding co-operative so that the major hybrids in Fig. 10 are positively correlated, or do they constitute mutually exclusive contacts? Once a sufficiently large Boltzmann sample we can easily compute e.g. correlations ρ_{PQ} between indicator variables P and Q that measure the existence of external base pairs in two different hybrids. The r.h.s. panels in Fig. 10 give typical examples, showing that typically there are strong correlations

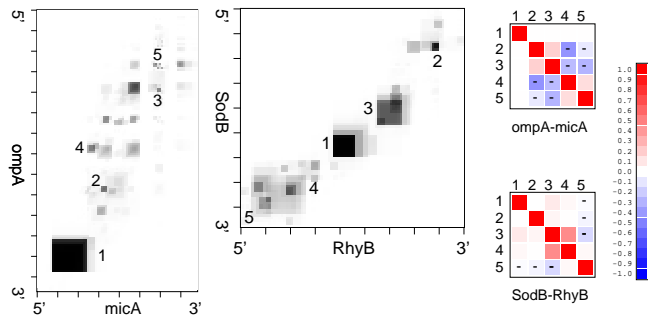


Fig. 10. Interaction maps. The *ompA-MicA* interaction (l.h.s.) has a dominating interaction region that brings together the 3' end of *ompA* and the 5' terminus of *MicA*. The *sodB-RhyB* interactions (middle) has two clear hybridization regions in the middle of the molecules and a diffuse contact area at the 3' end of *sodB*. The grayscale show the probabilities π_{ik} that (i, k) is contained in a hybrid. Tick marks indicate every 10th nucleotide. The correlations between the major binding regions can be computed easily from Boltzmann samples. The heatmaps show the correlation coefficients for the most probable interaction regions (indicated by numbers in the interaction maps). For *sodB-RhyB* we observe fairly weak correlations, except for the cooperative interaction between contacts 3 and 4. In the case of *ompA-MicA*, on the other hand, we observe strong negative correlations between conflicting hybridization regions.

between hybridization regions. The multiple contacts that can contribute substantially to the total interaction energy.

Acknowledgements. We want to thank Sven Findeiß for discussions. We are grateful to Sharon Selzo of the Modular and BICoC Benchmark Center, IBM and Kathy Tzeng of IBM Life Sciences Solutions Enablement. Their great support was vital for all computations presented here. 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 8507-1 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) *ftlA* 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., Cavaill e, J. and H utenhofer, 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., M uckstein, 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.

Chitsaz, H., Backofen, R. and Sahinalp, S. (2009a) biRNA: Fast RNA-RNA binding sites prediction. In *Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI)*. Accepted.

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

Ding, Y. and Lawrence, C. (2003) A statistical sampling algorithm for rna secondary structure prediction. *Nucleic Acid Res.*, **31**, 7280–7301.

Dowell, R. D. and Eddy, S. R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 7.

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.

Hekimoglu, B. and Ringrose, L. (2009) Non-coding RNAs in polycomb/trithorax regulation. *RNA Biol.*, **6**, 129–137.

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., Stadler, P. and Reidys, C. (2009) Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*.

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

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.

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.

M uckstein, 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 Eloumi, M., K ung, 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.

M uckstein, U., Tafer, H., Hackerm uller, J., Bernhard, S., Stadler, P. and Hofacker, I. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

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., H ochsmann, 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 algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

Salari, R., Backofen, R. and Sahinalp, S. (2009) Fast prediction of RNA-RNA interaction. In *Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI)*. Accepted.

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L. and Schuster, P. (1996) Algorithm independent properties of RNA structure prediction. *Eur. Biophys. J.*, **25**, 115–130.

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

Tjaden, B., Goodwin, S., Opdyke, J., Guillier, M., Fu, D., Gottesman, S. and Storz, G. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res.*, **34**, 2791–2802.

Udekwi, K., Darfeuille, F., Vogel, J., Reimeg ad, J., Holmqvist, E. and Wagner, E. G. H. (2005) Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev.*, **19**, 2355–2366.

Urban, J. H. and Vogel, J. (2007) Translational control and target recognition by *Escherichia coli* small RNAs in vivo. *Nucleic Acids Res.*, **35**, 1018–1037.

Zuker, M. and Sankoff, D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.