

Sequence analysis

Advance Access publication December 5, 2010

RNA–RNA interaction prediction based on multiple sequence alignmentsAndrew X. Li¹, Manja Marz², Jing Qin^{3,4} and Christian M. Reidys^{1,5,*}

¹Center for Combinatorics, the Key Laboratory of Pure Mathematics and Combinatorics, Tianjin Key Laboratory of Combinatorics (LPMC-TJKLC), Nankai University Tianjin 300071, PR China, ²RNA Bioinformatics Group, Philipps-University Marburg, Marbacher Weg 6, 34037 Marburg, ³Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, ⁴Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany and ⁵College of Life Science, Nankai University Tianjin 300071, PR China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Many computerized methods for RNA–RNA interaction structure prediction have been developed. 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. However, few of these methods incorporate the knowledge concerning related sequences, thus relevant evolutionary information is often neglected from the structure determination. Therefore, it is of considerable practical interest to introduce a method taking into consideration both: thermodynamic stability as well as sequence/structure covariation.

Results: We present the *a priori* folding algorithm *ripalign*, whose input consists of two (given) multiple sequence alignments (MSA). *ripalign* outputs (i) the partition function, (ii) base pairing probabilities, (iii) hybrid probabilities and (iv) a set of Boltzmann-sampled suboptimal structures consisting of canonical joint structures that are compatible to the alignments. Compared to the single sequence-pair folding algorithm *rip*, *ripalign* requires negligible additional memory resource but offers much better sensitivity and specificity, once alignments of suitable quality are given. *ripalign* additionally allows to incorporate structure constraints as input parameters.

Availability: The algorithm described here is implemented in C as part of the *rip* package. The supplemental material, source code and input/output files can freely be downloaded from <http://www.combinatorics.cn/cbpc/ripalign.html>.

Contact: duck@santafe.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 26, 2010; revised on October 5, 2010; accepted on November 29, 2010

1 INTRODUCTION

RNA–RNA interactions play a major role at many different levels of the cellular metabolism such as plasmid replication control, viral encapsidation, or transcriptional and translational regulation. With the discovery that a large number of transcripts in higher

eukaryotes are non-coding RNAs, RNA–RNA interactions in cellular metabolism are gaining in prominence. Typical examples of interactions involving two RNA molecules are snRNAs (Forne *et al.*, 1996); snoRNAs with their targets (Bachellerie *et al.*, 2002); microRNAs from the RNAi pathway with their mRNA target (Ambros, 2004; Murchison and Hannon, 2004); sRNAs from *Escherichia coli* (Hershberg *et al.*, 2003; Repoila *et al.*, 2003); and sRNA loop–loop interactions (Brunel *et al.*, 2003). The common feature in many ncRNA classes, especially prokaryotic small RNAs, is the formation of RNA–RNA interaction structures that are much more complex than the simple sense–antisense interactions.

As it is the case for 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, 2009). However, polynomial-time algorithms can be derived 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 algorithms RNACofold (Bernhart *et al.*, 2006; Hofacker *et al.*, 1994), pairfold (Andronescu *et al.*, 2005) and NUPACK (Dirks *et al.*, 2007; Zadeh *et al.*, 2010) 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.*, 2009b). 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 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 (Geissmann and Touati, 2004).

Pervouchine (2004) and Alkan *et al.* (2006) independently proposed MFE folding algorithms for predicting the *joint structure* of two interacting RNA molecules with polynomial time complexity. In their model, a ‘joint structure’ means that the intramolecular

*To whom correspondence should be addressed.

structures of each molecule are pseudoknot-free, the intermolecular binding pairs are non-crossing and there exist no so-called ‘zig-zags’, see Supplementary Material for detailed definition. The optimal joint structure is computed in $O(N^6)$ time and $O(N^4)$ space via a dynamic programming (DP) routine.

A more reliable approach is to consider the partition function, which by construction integrates over the Boltzmann-weighted probability space, allowing for the derivation of thermodynamic quantities, like e.g. equilibrium concentration, melting temperature and base pairing probabilities. The partition function of joint structures was independently derived by Chitsaz *et al.* (2009b) and Huang *et al.* (2009).

A key quantity here is the probability of hybrids, which cannot be recovered from base pairing probabilities since the latter can be highly correlated. Huang *et al.* (2010) presented a new hybrid-based decomposition grammar, facilitating the computation of the non-trivial hybrid probabilities as well as the Boltzmann sampling of RNA–RNA interaction structures. The partition function of joint structures can be computed in $O(N^6)$ time and $O(N^4)$ space and 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) introduced tree-structured Markov Random Fields to approximate the joint probability distribution of multiple (≥ 3) contact regions.

Unfortunately, incompleteness of the underlying energy model, in particular for hybrid- and kissing-loops, may result in prediction inaccuracy. One way of improving this situation is to involve phylogenetic information of multiple sequence alignments (MSAs).

In an MSA, homologous nucleotides are grouped in columns, where homologous is interpreted in *both*: structural as well as evolutionary sense. That is a column of nucleotides occupies similar structural positions and all diverge from a common ancestral nucleotide. Also, many ncRNAs show clear signs of undergoing compensatory mutations along evolutionary trajectories. In conclusion, it seems reasonable to stipulate that a non-negligible part of the existing RNA–RNA interactions contain preserved but covarying patterns of the interactions (Seemann *et al.*, 2010a, b). Therefore, we can associate a consensus interaction structure to pairs of interacting MSAs (see Section 2.1).

Along these lines, Seemann *et al.* (2010a, b) presented an algorithm PETcofold for prediction of RNA–RNA interactions including pseudoknots in given MSAs. Their algorithm is an extension of PETfold (Seemann *et al.*, 2008) using elements of RNACofold (Bernhart *et al.*, 2006) and computational strategies for hierarchical folding (Gaspin and Westhof, 1995; Jabbari *et al.*, 2007). However, PETcofold is an heuristics-based algorithm. A detailed comparative analysis also including the algorithm inteRNA (Salari *et al.*, 2009) will be given in Section 3.

Here, we present the algorithm ripalign which computes the partition function, base pairing as well as hybrid probabilities and performs Boltzmann sampling on the level of MSAs. ripalign represents a generalization of rip to pairs of interacting MSAs and a new grammar of canonical interaction structures. The latter is of relevance since there are no isolated base pairs in molecular complexes.

One important step consists in identifying the notion of a joint structure compatible to a pair of interacting MSAs. Our notion

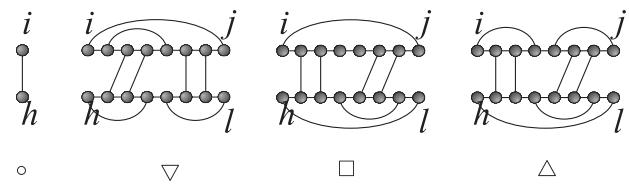


Fig. 1. The four basic types of tight structures are given as follows:
 \circ : $\{R_iS_h\} = J_{i,j;h,\ell}$ and $i=j$, $h=\ell$;
 ∇ : $R_iR_j \in J_{i,j;h,\ell}$ and $S_hS_\ell \notin J_{i,j;h,\ell}$;
 \square : $\{R_iR_j, S_hS_\ell\} \in J_{i,j;h,\ell}$;
 Δ : $S_hS_\ell \in J_{i,j;h,\ell}$ and $R_iR_j \notin J_{i,j;h,\ell}$.

is based on the framework of Hofacker *et al.* (2002), where a sophisticated cost function capturing thermodynamic stability as well as sequence covariation is employed. Furthermore, ripalign is tailored to take structure constraints, such as blocked nucleotides known e.g. from chemical probing, into account.

2 THEORY

2.1 MSAs and compatibility

An MSA, $\bar{\mathbf{R}}$, consists of $m_{\bar{\mathbf{R}}}$ RNA sequences of known species. Denoting the length of the aligned sequences by N , $\bar{\mathbf{R}}$ constitutes a $m_{\bar{\mathbf{R}}} \times N$ matrix, having 5'–3' oriented rows, $\bar{\mathbf{R}}^i$ and columns, $\bar{\mathbf{R}}_i$. Its (i,j) -th entry, $\bar{\mathbf{R}}_j^i$, is a nucleotide, A, U, G, C or a gap denoted by ..

For any pair $(\bar{\mathbf{R}}, \bar{\mathbf{S}})$, we assume that $\bar{\mathbf{S}}$ is a $m_{\bar{\mathbf{S}}} \times M$ matrix, whose rows carry 3'–5' orientation.

In the following, we shall assume that a pair of RNA sequences can only interact if they belong to the same species. A pair $(\bar{\mathbf{R}}, \bar{\mathbf{S}})$ can interact if for any row $\bar{\mathbf{R}}^i$, there exist at least one row in $\bar{\mathbf{S}}$ that can interact with $\bar{\mathbf{R}}^i$.

Given a pair of interacting MSAs $(\bar{\mathbf{R}}, \bar{\mathbf{S}})$, let m be the total number of potentially interacting pairs. ripalign exhibits a pre-processing step which generates a $m \times N$ -matrix \mathbf{R} and a $m \times M$ -matrix \mathbf{S} such that $(\mathbf{R}^i, \mathbf{S}^j)$ range over all m potentially interacting RNA pairs, see Section 1.2. of Supplementary Material for details.

In the following, we shall refer to \mathbf{R} and \mathbf{S} as MSAs ignoring the fact that they have multiple sequences.

We proceed by defining joint structures that are compatible to a fixed (\mathbf{R}, \mathbf{S}) . To this end, let us briefly review some concepts introduced in Huang *et al.* (2009).

A joint structure $J(\mathbf{R}, \mathbf{S}, I)$ is a graph consisting of

- (j1) Two secondary structures R and S , whose backbones are drawn as horizontal lines on top of each other and whose arcs are drawn in the upper and lower halfplane, respectively. We consider R over a 5' to 3' oriented backbone (R_1, \dots, R_N) and S over a 3' to 5' oriented backbone (S_1, \dots, S_M) and refer to any R - and S -arcs as interior arcs.
- (j2) An additional set I , of non-crossing arcs of the form R_iS_j (exterior arc), where R_i and S_j are unpaired in R and S .
- (j3) $J(\mathbf{R}, \mathbf{S}, I)$ contains no ‘zig-zags’.

The subgraph of a joint structure $J(\mathbf{R}, \mathbf{S}, I)$ 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, $J(\mathbf{R}, \mathbf{S}, I) = J_{1,N;1,M}$ 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_1, \dots, R_j) and (S_h, \dots, S_ℓ) . In particular, we use $S[i, j]$ to denote the subgraph of $J_{1,N;1,M}$ induced by $(S_i, S_{i+1}, \dots, S_j)$, where $i \leq j$. In particular, in case of $i=j$, we identify $S[i, i]$ with the vertex S_i .

Given a joint structure, $J_{a,b;c,d}$, a tight structure (TS), $J_{i,j;h,\ell}$, (Huang *et al.*, 2009) is a specific subgraph of $J_{a,b;c,d}$ indexed by its type $\in \{\circ, \nabla, \square, \Delta\}$, see Figure 1. For instance, we use $J_{i,j;h,\ell}^\square$ to denote a TS of type \square .

A *hybrid* is a joint structure $J_{i_1,i_\ell;j_1,j_\ell}^{\text{Hy}}$, i.e. a maximal sequence of intermolecular interior loops consisting of a set 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.

A joint structure $J(R, S, I)$ is called *canonical* if and only if:

- (c1) each stack in the secondary structures R and S is of size at least two, i.e. there exist no isolated interior arcs,
- (c2) each hybrid contains at least two exterior arcs.

In the following, we always assume a joint structure to be canonical.

Next, we come to (\mathbf{R}, \mathbf{S}) -compatible joint structures. In difference to single sequence compatibility, this notion involves statistical information of the MSAs.

The key point consists in specifying under which conditions two vertices contained in $(R_1, \dots, R_N, S_1, \dots, S_M)$ can pair. This is obtained by a generalization of the RNAalifold approach (Hofacker *et al.*, 2002). We specify these conditions for interior $(c_{i,j}^{\mathbf{R}})$, $(c_{i,j}^{\mathbf{S}})$ and exterior pairs $(c_{i,j}^{\mathbf{R}, \mathbf{S}})$ in Equation (2.3)–(2.5).

For interior arcs (R_i, R_j) , let $X, Y \in \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$. Let $f_{ij}^{\mathbf{R}}(XY)$ be the frequency of (X, Y) which exists in the 2-column submatrix $(\mathbf{R}_i, \mathbf{R}_j)$ as a row-vector and

$$C_{i,j}^{\mathbf{R}} = \sum_{XY, X'Y'} f_{ij}^{\mathbf{R}}(XY) D_{XY, X'Y'}^{\mathbf{R}} f_{ij}^{\mathbf{R}}(X'Y'). \quad (2.1)$$

Here, XY and $X'Y'$ independently range over all 16 elements of $\{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\} \times \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$ and $D_{XY, X'Y'}^{\mathbf{R}} = d_H(XY, X'Y')$, i.e. the Hamming distance between XY and $X'Y'$ in case of XY and $X'Y'$ being Watson–Crick, or \mathbf{GU} wobble base pair and 0, otherwise. Furthermore, we introduce $q_{i,j}^{\mathbf{R}}$ to deal with the inconsistent sequences

$$q_{i,j}^{\mathbf{R}} = 1 - \frac{1}{m} \sum_h \{\Pi_{i,j}^h(\mathbf{R}) + \delta(\mathbf{R}_i^h, \cdot) \delta(\mathbf{R}_j^h, \cdot)\}, \quad (2.2)$$

where $\delta(x, y)$ is the Kronecker delta and $\Pi_{i,j}^h(\mathbf{R})$ is equal to 1 if \mathbf{R}_i^h and \mathbf{R}_j^h are Watson–Crick or \mathbf{GU} wobble base pair and 0, otherwise. Now we obtain $B_{i,j}^{\mathbf{R}} = C_{i,j}^{\mathbf{R}} - \phi_1 q_{i,j}^{\mathbf{R}}$. Based on sequence data, the threshold for pairing $B_*^{\mathbf{R}}$ as well as the weight of inconsistent sequences ϕ_1 are computed. We have

$$(c_{i,j}^{\mathbf{R}}) \quad B_{i,j}^{\mathbf{R}} \geq B_*^{\mathbf{R}}. \quad (2.3)$$

The case of two positions S_i and S_j is completely analogous

$$(c_{i,j}^{\mathbf{S}}) \quad B_{i,j}^{\mathbf{S}} \geq B_*^{\mathbf{S}}, \quad (2.4)$$

where $B_{i,j}^{\mathbf{S}}$ and $B_*^{\mathbf{S}}$ are analogously defined.

As for $(c_{i,j}^{\mathbf{R}, \mathbf{S}})$ a further observation factors in: since many ncRNA show clear signs of undergoing compensatory mutations in the course of evolution (Marz *et al.*, 2008; Seemann *et al.*, 2010b), we postulate the existence of a non-negligible amount of RNA–RNA interactions containing conserved pairs, consistent mutations, compensatory mutations as well as inconsistent mutations. Based on this observation we arrive at

$$(c_{i,j}^{\mathbf{R}, \mathbf{S}}) \quad B_{i,j}^{\mathbf{R}, \mathbf{S}} \geq B_*^{\mathbf{R}, \mathbf{S}}, \quad (2.5)$$

where $B_{i,j}^{\mathbf{R}, \mathbf{S}}$ and $B_*^{\mathbf{R}, \mathbf{S}}$ are analogously defined as the case $B_{i,j}^{\mathbf{R}}$ and $B_*^{\mathbf{R}}$.

A joint structure J is compatible to (\mathbf{R}, \mathbf{S}) if for any J -arc, the corresponding intra- or inter-positions can according to Equation (2.3)–(2.5) pair.

2.2 Energy model

According to Huang *et al.* (2009), joint structures can be decomposed into disjoint loops. These loop types include standard hairpin-, bulge-, interior- and multi-loops found in RNA secondary structures as well as *hybrids* and *kissing-loops*. Following the energy parameter rules of Mathews *et al.* (1999), the energy of each loop can be obtained as a sum of the energies associated with non-terminal symbols, i.e. graph properties (sequence independent) and an additional contributions that depend uniquely on the terminal bases (sequence dependent).

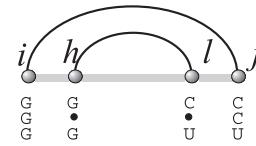


Fig. 2. Interior loop energy: an interior loop formed by $R_i R_j$ and $R_h R_\ell$, where $i < h < \ell < j$ are the alignment positions. Gray bands are used to denote the positions we omit between segment (i, h) , (h, ℓ) and (ℓ, j) .

Suppose we are given a joint structure J , compatible to a pair $\mathcal{P} = (\mathbf{R}, \mathbf{S})$. Let $L \in J$ be a loop and let $\mathcal{F}_{L,i}$ represent the loop energy of the i -th interaction pair $(\mathbf{R}^i, \mathbf{S}^i)$. Then the loop energy of \mathcal{P} is

$$\mathcal{F}_{\mathcal{P}, \mathcal{P}} = 1/m \sum_i \mathcal{F}_{L,i}. \quad (2.6)$$

We consider the energy of the structure as the sum of all loop contributions:

$$\mathcal{F}_J = \sum_{L \in J} \mathcal{F}_{L, \mathcal{P}}. \quad (2.7)$$

To save computational resources, gaps are treated as bases in *riparig*. Thus, only alignment positions contribute as indices and loop sizes. Since no measured energy parameters for non-standard base pairs are available at present time, additional terminal-dependent contributions for the latter are ignored. For instance, let $\text{Int}_{i,j;h,l}$ denote an interior loop formed by $R_i R_j$ and $R_h R_\ell$ and $\mathcal{F}_{\text{Int}, \mathcal{P}}$ denote the free energy of $\text{Int}_{i,j;h,l}$ with respect to the aligned sequences in \mathcal{P} . Then $\mathcal{F}_{\text{Int}, \mathcal{P}}$ associated to the three aligned subsequences of Figure 2 reads

$$\mathcal{F}_{i,j;h,l}^{\text{Int}, \mathcal{P}} = \frac{1}{3} (3G_{i,j;h,l}^{\text{Int}} + G_{*,\mathbf{G},\mathbf{C};\mathbf{G},\mathbf{C}}^{\text{Int}} + G_{*,\mathbf{G},\mathbf{U};\mathbf{G},\mathbf{U}}^{\text{Int}} + G_{*,\mathbf{G},\mathbf{C};\dots}^{\text{Int}}). \quad (2.8)$$

Here $G_{i,j;h,l}^{\text{Int}}$ represents contributions related exclusively to the positions of the interior loop while $G_{*,\mathbf{A},\mathbf{B};\mathbf{C},\mathbf{D}}^{\text{Int}}$ represents additional contributions related to the specific nucleotides which form the interior loop. We set $G_{*,\mathbf{G},\mathbf{C};\dots}^{\text{Int}}$ to be zero.

2.3 The grammar of canonical joint structures and the partition function

The partition function algorithm is easily extended to work with the modified energy functions given in Equation (2.7). The reformulation of the original hybrid-grammar into a grammar of canonical joint structures represents already for single interaction pairs a significant improvement in prediction quality. The original *riparig*-grammar would oftentimes encounter joint structures having a hybrid composed by a single isolated exterior arc. In order to decompose canonical joint structures via the unambiguous grammar, detailed in the Section 2 of Supplementary Material, we distinguish the two types (Type cc and Type c) of TS’s of type ∇ , Δ or \square . Given a TS of type ∇ , denoted by $J_{i,j;h,\ell}^{\nabla}$, we write depending on whether $R_{i+1} R_{j-1} \in J_{i,j;h,\ell}^{\nabla}$, $J_{i,j;h,\ell}^{\nabla,cc}$ and $J_{i,j;h,\ell}^{\nabla,c}$, respectively. Analogously, we define $J_{i,j;h,\ell}^{\square,cc}$, $J_{i,j;h,\ell}^{\square,c}$ and $J_{i,j;h,\ell}^{\triangle,cc}$, $J_{i,j;h,\ell}^{\triangle,c}$, see Figure 3.

As illustrated in the Figure 4 of Supplementary Material, there are two basic steps of the canonical-grammar: (i) *interior arc-removal* to reduce TS and (ii) *block-decomposition* to split a joint structure into two smaller blocks. The key feature here is, that since J is canonical, the smaller blocks are still canonical after block decomposition. Each decomposition step results in substructures which eventually break down into generalized loops whose energies can be directly computed. More details of the decomposition procedures are described in Section 2 of Supplementary Material, where we prove that for any canonical joint structure J , there exists a unique decomposition tree (parse-tree), denoted by T_J , see Figure 4.

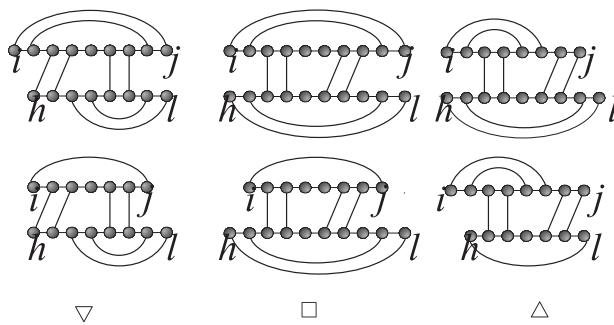


Fig. 3. Examples of two TS-types. We display ∇ , \square , or \triangle -tight structures: Type cc (top) and Type c (bottom).

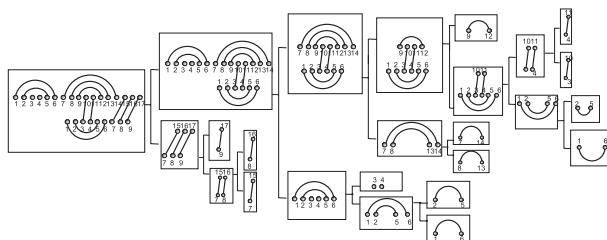


Fig. 4. Example of the parse tree. The parse tree of the canonical joint structure $J_{1,17;1,9}$.

2.4 Probabilities and the Boltzmann sampling

A dynamic programming scheme for the computation of a partition function implies a corresponding computation of probabilities of specific substructures is obtained ‘from the outside to the inside’ and a stochastic backtracing procedure that can be used to sample from the associated distribution (Ding and Lawrence, 2003; Huang *et al.*, 2010; McCaskill, 1990). We remark that the time complexity does not increase linearly as a function of m (see Supplementary Table S6). Along the lines of the design of the Vienna software package (Hofacker *et al.*, 1994), ripalign now offers the following features as optional input parameters:

- (i) a position i can be restricted to form an interior or an exterior arc (denoted by ‘–’ and ‘^’, respectively);
- (ii) a position i can be forced to be unpaired (denoted by ‘x’);
- (iii) a position i can be restricted to form an (interior or an exterior) arc with *some* position j (denoted by ‘*’);
- (iv) a pair of positions i and j can be forced to form an interior or exterior arc (denoted by ‘()’ or ‘[]’, respectively).

However, the above features are optional. Thus, ripalign can deal with both scenarios: the absence of any *a priori* information and the existence of specific information, e.g. the location of the Sm-binding site, see Figure 7.

3 RESULTS AND DISCUSSION

This discussion is organized into three parts: Firstly, we apply ripalign to two well-known ncRNA interactions: the *fhlA/OxyS* and the *CopT/CopA*. In the process, we provide a comparative analysis with rip, PETcofold with stem constraints (PETcofold⁺), without stem constraints (PETcofold⁻) and InteRNA. The stem extension constraint of PETcofold represents a specific programming technique of Seemann *et al.* (2010a, b) to avoid incomplete stems. Secondly we study two interactions whose consensus structures are still under investigation: (i) the interaction of the spliceosomal RNAs U4/U6 snRNA and (ii) the interaction of the *SmY-10/SL-1* RNA

Table 1. Comparative analysis of ripalign with rip Huang *et al.* (2009), PETcofold Seemann *et al.* (2010b); PETcofold⁻ (PC⁻) and PETcofold⁺ (PC⁺) and InteRNA Salari *et al.* (2009)

	rip	ripalign	PC ⁻	PC ⁺	InteRNA
<i>fhlA/OxyS</i> (entire joint structure predicted)					
Sensitivity	0.6286	0.7714	0.4143	0.6286	0.7143
Specificity	0.9987	0.9990	0.9992	0.9995	0.9985
PPV	0.5714	0.6835	0.6042	0.7857	0.5747
MCC	0.5981	0.7253	0.4991	0.7020	0.6396
ACC	0.9976	0.9983	0.9976	0.9985	0.9977
MEF	-61.96	-89.91	-32.75	-57.48	-81.13
Runtime	4.11d	3.91d	35 s	33 s	4.13 h
<i>CopT/CopA</i> (entire joint structure predicted)					
Sensitivity	0.9091	0.9091	0.9091	0.8864	0.9091
Specificity	0.9984	0.9984	0.9981	0.9984	0.9986
PPV	0.8000	0.8000	0.7692	0.7959	0.8163
MCC	0.8517	0.8517	0.8350	0.8388	0.8604
ACC	0.9978	0.9978	0.9975	0.9976	0.9979
MEF	-57.64	-57.64	-52.11	-57.20	-56.12
Runtime	8.63 h	10.16 h	3.589 s	3.065 s	5.20 m

We consider sensitivity, specificity, PPV, ACC, MCC and the MFE of the predicted structures. The *fhlA/OxyS*-joint structures of ripalign and PETcofold are based on the MSAs displayed in Figure 5. The *fhlA/OxyS*-joint structures predicted by rip and InteRNA are based on the 8th pair of sequences displayed in Figure 5. All quantities are based on the natural structure derived in Seemann *et al.* (2010b), free energy is given in kcal/mol. The *CopT/CopA*-joint structures are based on the sequences presented in Wagner and Flärdh (2002). All calculations were performed on an Intel(R) Xeon(R) E5410 @ 2.33 GHz (eight cores), 48 GB memory, CentOS release 5.3. The runtime is given in day (d), hour (h), minute (m) and second (s).

proposed by MacMorris *et al.* (2007). Thirdly, we integrate our findings and give concluding remarks.

3.1 The *fhlA/OxyS* interaction

The *OxyS* RNA represses *fhlA* mRNA translation initiation through base pairing with two short sequences (Argaman and Altuvia, 2000), one of which overlaps the ribosome binding sequence and the other resides further downstream, within the coding region of *fhlA*. We can report that ripalign improves the computational results in terms of sensitivity, specificity, positive prediction value (PPV), accuracy (ACC), Matthew’s correlation coefficient (MCC) in comparison with rip. Furthermore, we can report improvements compared to the existing alignment interaction programs, see Table 1. The exception here is the PPV 0.7857 achieved by PETcofold⁺.

Comparison of the predicted interaction (Fig. 5) shows that ripalign and PETcofold⁻ produce a similar result and agree well with the natural structure derived in Argaman and Altuvia (2000). In contrast, rip and PETcofold⁺ predict joint structures that differ significantly from Argaman and Altuvia (2000). We display for ripalign each contact region having a probability > 10% and accordingly highlight in Figure 5D four distinct contact regions. The two additional contact regions, identified in the partition function, exhibit a significantly lower probability. An additional hairpin over $R[72,89]$ is predicted in *OxyS*, instead of the unpaired segment occurring in the natural structure, can be understood in the context of minimizing free energy.

3.2 The *CopT/CopA* interaction

It is known that the antisense RNA *CopA* binds to the leader region of the *repA* mRNA. The target here is named *CopT* (Kolb *et al.*, 2000; Wagner and Flärdh, 2002). In lack of MSA for *CopT/CopA*, we cannot utilize the full potential of ripalign; however, the interaction is well

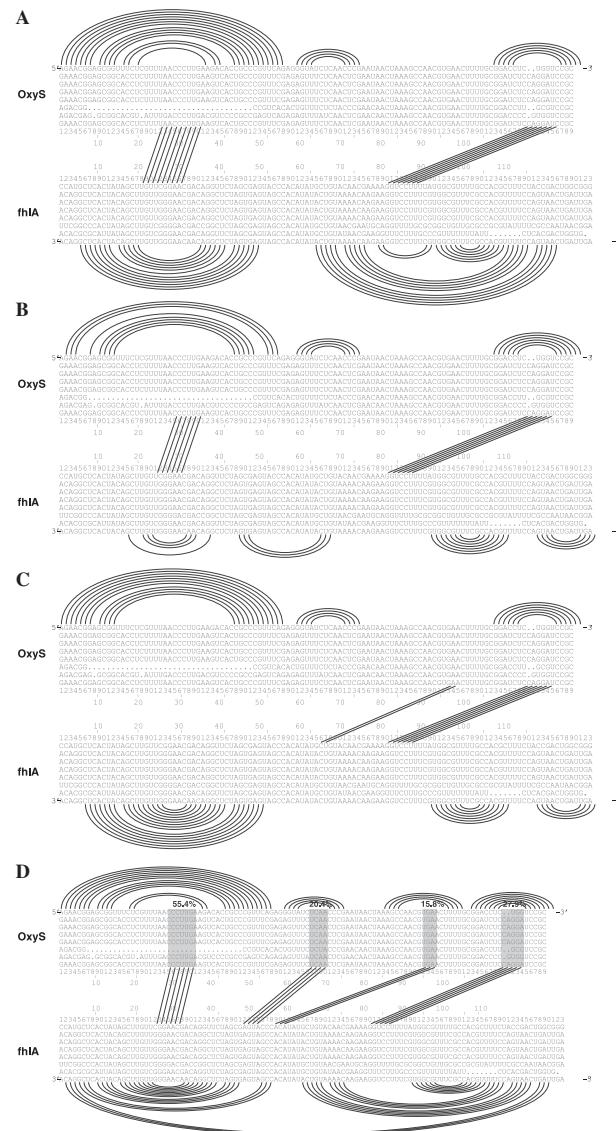


Fig. 5. Alignment-based prediction of *fhlA/OxyS* (**A**) the annotated structure of Argaman and Altuvia (2000), (**B**) the PETcofold⁻ structure, (**C**) the PETcofold⁺ structure and (**D**) the joint structure as predicted by ripalign based on the entire MSA. Target site (grey blocks) probabilities [as detailed in Section 4 of Supplementary Material, Equation (5.5)], computed by ripalign, are annotated explicitly if they exceed 10% or just by $\leq 10\%$, otherwise.

known and included for reference purposes. The quality of the prediction is comparable to that of PETcofold⁺, PETcofold⁻ or InteRNA, see Table 1. *CopT/CopA* exhibits two distinct interactions. While the first (main) interaction is correctly identified as $J_{25,30;20,27}^H$ by ripalign, the second interaction is ranked fifth, $J_{54,57;53,56}^H$ with a probability of 6.14%, see Table 2.

3.3 The *U4/U6* interaction

U4 and *U6* are two spliceosomal RNAs (snRNAs) that are known to be involved in splicing pre-mRNA. For at least a quarter century, these two ncRNA molecules have been conjectured to interact. The precise nature of their interaction and related proteins is still the subject of investigation.

Table 2. Top five hybrids and their hybrid-probabilities (H-prob)

Rank	<i>fhlA/OxyS</i>	<i>CopT/CopA</i>		
	Hybrid	H-prob (%)	Hybrid	H-prob (%)
1	$J_{25,30;20,27}^H$	55.41	$J_{17,38;17,38}^H$	38.71
2	$J_{98,102;69,73}^H$	27.33	$J_{17,37;17,37}^H$	13.12
3	$J_{34,36;42,44}^H$	21.37	$J_{16,39;16,39}^H$	10.33
4	$J_{56,59;41,44}^H$	20.17	$J_{18,37;18,37}^H$	8.43
5	$J_{101,102;98,99}^H$	16.22	$J_{54,57;53,56}^H$	6.14

The hybrid-probabilities of *CopT/CopA* and *fhlA/OxyS* listed in the table as predicted by ripalign.

Adopting an evolutionary perspective, we notice that there are significant distinctions between the different clades (ranging from simple protostomes to higher deuterostomes) in which this interaction is described (López *et al.*, 2008; Otake *et al.*, 2002; Shambaugh *et al.*, 1994; Shukla *et al.*, 2002; Thomas *et al.*, 1990). We find for, instance, a discussed 5' stem of *U6* as well as stem I between *U4* and *U6*. Furthermore, even within e.g. human a third interaction (stem III) is discussed (Brow and Vidaver, 1995; Jakab *et al.*, 1997 A. Bindereif and S. Rader, personal communication). Figure 6A shows a consensus of the previously described *U4/U6* interactions.

We divide all known metazoan *U4* and *U6* snRNAs into three distinct groups and alignments: protostomia without insects, insects and deuterostomia. Marz *et al.* (2008) observed that insects behave in their secondary structure different from other protostomes. Comparing all predicted *U4/U6* interactions, displayed in Figure 6B–E, we draw the following conclusions:

- (i) The secondary partial structures of the *U4/U6* complex for all three groups predicted by ripalign agree predominantly with the described secondary structures in metazoans: Stem I and II are conserved at present, as well as the four commonly described hairpins.
- (ii) For all three groups, Stem II (Fig. 6, top) is highly conserved. External ascendancies, such as protein interactions are discussed to stabilize Stem II additionally. Stem I is also highly conserved except for insects, which agrees with the literature.
- (iii) For all three groups, the 5' hairpin of *U4* snRNA seems highly conserved to interact with the *U6* snRNA ('N'). This RNA feature is not fully understood, since this element is also believed to contain intraloop interactions and may bind to a 15.5-kDa protein (Vidovic *et al.*, 2000).
- (iv) For all metazoans, the *U6* snRNA shows conserved intramolecular interactions (3'stem '3'st).
- (v) Stem III (Brow and Vidaver, 1995; Jakab *et al.*, 1997) seems to be not a conserved feature; however, the same region of *U6* interacts in deuterostomes significantly with a probability of 24.3% in higher metazoans with the central stem loop of *U4* (Fig. 6E, 'J').
- (vi) For both: protostomia (without insects) and deuterostomes, the 5' hairpin of *U6* snRNA seems to interact with the *U4* 3' hairpin ('M'). However, this observation does not hold for insects, which agrees with a systematically different secondary structure of spliceosomal RNAs in insects (Marz *et al.*, 2008).

Since the three divergent groups of metazoans independently exhibit analogous secondary structure features, ripalign has led to observations (3)–(6), potentially having biological relevance.

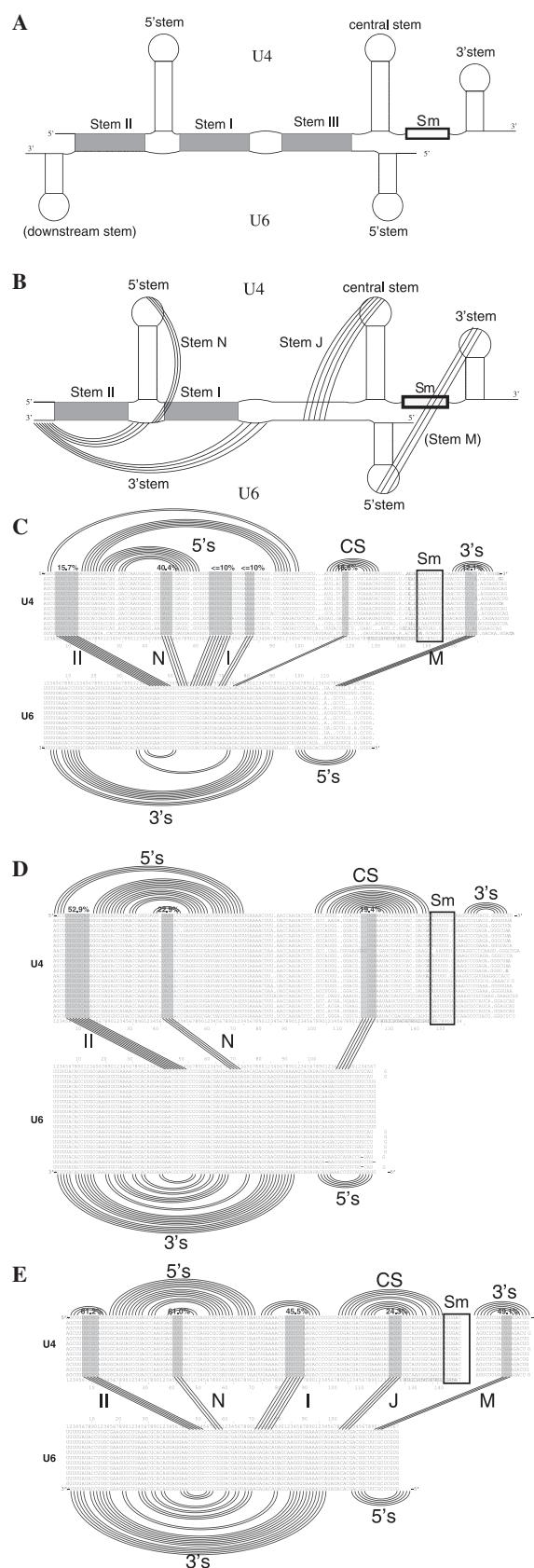


Fig. 6. The *U4/U6* interaction prediction with Sm-binding site constraint in *U4*. The Sm-binding site in molecule *U4* is 5'-AAUUUUUG-3' (black frames

3.4 The *SmY-10/SL-1* interaction of *Caenorhabditis elegans*

MacMorris *et al.* (2007) stipulated that *SmY-10* RNA, is possibly involved in *trans*-splicing, interacts with the splice leader RNA (*SL-1* RNA). This interaction has never been identified before. In Figure 7, we show that the Sm-binding sites of the RNA molecules *SmY-10* and *SL-1* are *R*[56, 62] and *S*[25, 31], respectively. In Figure 7A, the structure is being predicted by *rip* (Huang *et al.*, 2010). We observe two unsatisfying phenomena: a stack in *SmY-10* is formed by a single intra arc *S*₂₄*S*₆₇ and there exist intra arcs in the Sm-binding sites. The canonical grammar presented here restricts the configuration ensemble to canonical joint structures, resulting in the structure presented in Figure 7B in which the peculiar isolated interaction arc disappears. However, the nucleotides of the Sm-binding sites still form either intra or inter-molecular base pairs. Incorporating the structural constraints option, we derive the structure displayed in Fig. 7C. Here the Sm-binding sites are constrained to be single stranded. Since the predicted interactions have either a low probability or consist of two base pairs only, we conclude that if *SmY-10* is interacting with *SL-1* *in vivo*, this interaction has to be stabilized further by e.g. proteins.

Our case studies show that *ripalign* improves the prediction of interaction regions significantly at the expense of computational cost. The case studies of the *fhlA/OxyS* and the *CopT/CopA* interaction show that *ripalign*'s full potential comes to bear when MSAs are present. It is then that our method can reliably predict the joint secondary structures. We remark that structure prediction is just part of the *ripalign*-functionalities. Derived thermodynamic quantities, in particular the base pairing and hybrid probabilities, are also useful in a variety of ways. For instance, they help to improve the accuracy of other software tools such as *RactIP* (Kato *et al.*, 2010).

The quality of prediction depends critically on the quality of the MSAs. The issue of alignment quality is not easily solved: creating an alignment without knowing the structure is unlikely to produce a structural alignment. In this work, we mainly used handmade alignments considering sequence and structure derived from literature knowledge. In case of automatically derived alignments, it might be an option to realign the sequences of an RNA family taking both the predicted secondary structures and predicted joint structure with other RNA families into consideration.

Clearly, *ripalign* is limited by its a priori output class of joint structures and can in particular not identify any joint structures exhibiting pseudoknots. To save computational resources, we stipulate that only alignment positions contribute as indices and loop sizes. This assumption may cause the existence of some interior arcs *R*_i*R*_j having arc length smaller than three. However, Bernhart *et al.* (2008) showed that this problem can be improved substantially by introducing a different, more rational handling of alignment gaps, and by replacing the rather simplistic model of covariance scoring with more sophisticated RIBOSUM-like scoring matrices.

Table 1 makes evident that approximation algorithms are much faster. For instance *PETcofold* (Seemann *et al.*, 2010a, b), has a time complexity of $O(m(N+M)^3 n)$, where m is the number of sequences in the MSA, N and M being the sequence lengths of the longer and shorter alignment, respectively, and $n < N/2$ is the number of iterations for the adaption of the threshold value to find likely partial secondary structures. Their basic assumption of *PETcofold* being that the two secondary structures fold independently and that intra-loop evaluation differences are negligible.

The flip side of reducing the complexity of a folding problem by introducing additional assumptions, is the uncertainty of the quality of the

labeled with Sm). (A) Consensus structure of previously published *U4/U6* interactions. (B) The consensus structure of *ripalign* results shows several possible conserved interactions. (C–E) The joint structures of protostomia (without insects), insects and deuterostomia are predicted by *ripalign* under the Sm-binding site constraint. The target site (grey blocks) probabilities computed by *ripalign* are annotated explicitly if > 10% or just by $\leq 10\%$, otherwise.

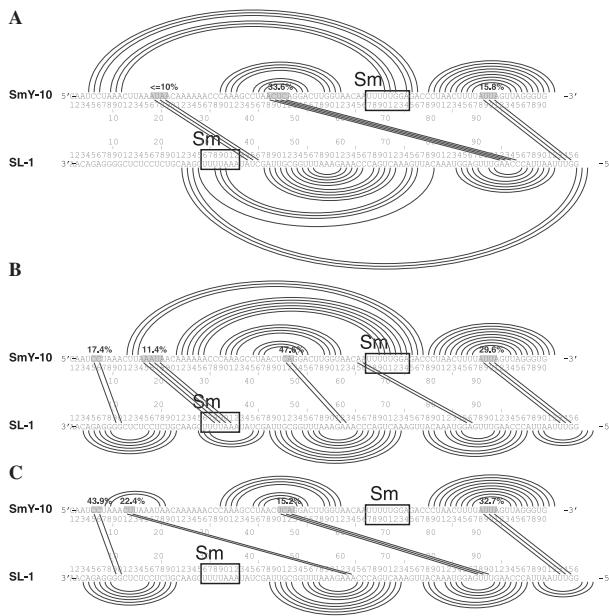


Fig. 7. ripalign versus rip: interaction of two specific RNA molecules, *SL-1* and *SmY-10* of *C.elegans*. The Sm-binding sites (black frames labeled with Sm) in the RNA molecules *SmY-10* and *SL-1* are 5'-AAUUUUG-3'(*R*[56,62]) and 3'-GUUUUA-5'(*S*[25,31]), respectively. The joint structure contains a single interior arc $S_{24}S_{67}$ (A), which is predicted by rip implemented by Huang *et al.* (2010). The joint structure (**B**) is predicted by ripalign without any structural constraints. The joint structure (**C**) is predicted by ripalign under the structural constraints that 5'-AAUUUUG-3'(*R*[56,62]) and 3'-GUUUUA-5'(*S*[25,31]) are Sm-binding sites in the RNA molecules *SmY-10* and *SL-1*, respectively. The target site (grey blocks) probabilities computed by ripalign are annotated explicitly if $> 10\%$ or just by $\leq 10\%$, otherwise.

solution. Point in case here is that arguably the two secondary structures did not evolve independently, but rather correlated by means of their functional interaction.

The RNA–RNA interaction problem exhibits a challenge long absent from the folding of RNA secondary structures: the time efficient folding of MFE configurations. While the $O(N^3)$ time complexity of the latter is, even for long sequences, easy to work with, it is clear that the $O(N^6)$ time complexity of the former has to be improved. At present, the only alternative to the DP-paradigm are heuristic-based algorithms and both approaches complement each other well. When encountering new scenarios, like the *U4/U6* interaction or the *SmY-10/SL-1* interaction of *C.elegans*, it seems however safer to explore the entire configuration space and make use of statistical information via the partition function—even at the high computational cost. For the *U4/U6* and the *SmY-10/SL-1* interaction, ripalign offers a detailed picture of the interaction and identifies potential interaction regions. Within its complexity limitations, ripalign is capable capturing the entire space of RNA interaction structures and allows to reconstruct the latter via Boltzmann sampling.

It is likely that the next milestone for RNA–RNA folding algorithms lies neither in DP-foldings nor in heuristics. Deeper structural understanding of the landscape generated by arc-configurations over fixed sequence alignments is the key here. Advances on this topic will eventually lead to provable error bounds and thereby simultaneously make the DP-paradigm and heuristics obsolete. At the same time, this might allow us to understand the *ad hoc* definition of joint structures due to Alkan *et al.* (2006), which is intimately connected to the topology of the configuration.

ACKNOWLEDGEMENTS

We want to thank Fenix W.D. Huang and Jan Engelhardt for their helpful suggestions. We thank Sharon Selzo of the Modular and BICoC Benchmark Center, IBM and Kathy Tzeng of IBM Life Sciences Solutions Enablement. Their support was vital for all computations presented here. We thank Albrecht Bindereif, Elizabeth Chester and Stephen Rader for their *U4/U6* analysis.

Funding: 973 Project of the Ministry of Science and Technology; the PCSIRT Project of the Ministry of Education; National Science Foundation of China to C.M.R. and his laboratory.

Conflict of Interest: none declared.

REFERENCES

- Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Disc. Appl. Math.*, **104**, 45–62.
- Alkan,C. *et al.* (2006) RNA–RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**, 267–282.
- Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
- Andronescu,M. *et al.* (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**, 1101–1112.
- Argaman,L. and Aluvia,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. *et al.* (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–779.
- Bernhart,S. *et al.* (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.
- Bernhart,S. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474–487.
- Brow,D. and Vidaver,R. (1995) An element in human U6 RNA destabilizes the U4/U6 spliceosomal RNA complex. *RNA*, **1**, 122–131.
- Brunel,C. *et al.* (2003) RNA loop-loop interactions as dynamic functional motifs. *Biochimie*, **84**, 925–944.
- Busch,A. *et al.* (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
- Chitsaz,H. *et al.* (2009a) biRNA: Fast RNA–RNA binding sites prediction. In *Proceedings of the 9th Workshop on Algorithms in Bioinformatics (WABI)*, Vol. 5724 of *LNCS*, Springer, Berlin/Heidelberg, pp. 25–36.
- Chitsaz,H. *et al.* (2009b) A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, **25**, i365–i373.
- Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acid Res.*, **31**, 7280–7301.
- Dirks,R. *et al.* (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65–88.
- Forne,T. *et al.* (1996) Structural features of U6 snRNA and dynamic interactions with other spliceosomal components leading to pre-mRNA splicing. *Biochimie*, **78**, 434–442.
- Gaspin,C. and Westhof,E. (1995) An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J. Mol. Biol.*, **254**, 163–174.
- 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.
- Hershberg,R. *et al.* (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.
- Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
- Hofacker,I. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
- Huang,F.W.D. *et al.* (2009) Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics*, **25**, 2646–2654.
- Huang,F. *et al.* (2010) Target prediction and a statistical sampling algorithm for RNA–RNA interaction. *Bioinformatics*, **26**, 175–181.
- Jabbari,H. *et al.* (2007) Hfold: RNA pseudoknotted secondary structure prediction using hierarchical folding. In Giancarlo,R. and Hannenhalli,S. (eds), *In Algorithms in Bioinformatics, 7th International Workshop, WABI 2007*. Springer, Philadelphia, PA, USA.

- Jakab,G. *et al.* (1997) Chlamydomonas U2, U4 and U6 snRNAs. An evolutionary conserved putative third interaction between U4 and U6 snRNAs which has a counterpart in the U4atac-U6atac snRNA duplex. *Biochimie*, **79**, 387–395.
- Kato,Y. *et al.* (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**, i460–i466.
- Kolb,F.A.*et al.* (2000) An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA*, **6**, 311–324.
- López,M. *et al.* (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.*, **36**, 3001–3010.
- MacMorris,M. *et al.* (2007) A novel family of *C. elegans* snRNPs contains proteins associated with *Trans*-splicing. *RNA*, **13**, 511–520.
- Marz,M. *et al.* (2008) Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, **67**, 594–607.
- Matthews,D. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
- McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
- Mneimneh,S. (2009) On the approximation of optimal structures for RNA-RNA interaction. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, **6**, 682–688.
- Mückstein,U. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
- Mückstein,U. *et al.* (2008) Translational control by RNA-RNA interaction: improved computation of RNA-RNA binding thermodynamics. In Elloumi,M. (eds), *BioInformatics Research and Development — BIRD 2008*, Vol.13 of *Comm. Comp. Inf. Sci.* Springer, Berlin, pp. 114–127.
- Murchison,E. and Hannon,G. (2004) miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell. Biol.*, **16**, 223–229.
- Otake,L. *et al.* (2002) The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. *Mol. Cell*, **9**, 439–446.
- Pervouchine,D. (2004) IRIS: Intermolecular RNA interaction search. *Proc. Genome Informatics*, **15**, 92–101.
- Rehmsmeier,M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *Gene*, **10**, 1507–1517.
- Repolo,F. *et al.* (2003) Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: The RpoS paradigm. *Mol. Microbiol.*, **48**, 855–861.
- Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithms for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
- Salari,R. *et al.* (2009) Fast prediction of RNA-RNA interaction. In *Proceedings of the 9th Workshop on Algorithms in Bioinformatics (WABI)*, Vol. 5724 of *LNCS*, pp. 261–272. Springer, Berlin/Heidelberg.
- Seemann,S. *et al.* (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.
- Seemann,S. *et al.* (2010a) PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, doi:10.1093/bioinformatics/btq034.
- Seemann,S. *et al.* (2010b) Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. *Algorithms Mol. Biology*, **5**, 22.
- Shambough,J. *et al.* (1994) The spliceosomal U small nuclear RNAs of *Ascaris lumbricoides*. *Mol. Biochem. Parasitol.*, **64**, 349–352.
- Shukla,G. (2002) Domains of human U4atac snRNA required for U12-dependent splicing in vivo. *Nucleic Acids Res.*, **30**, 4650–4657.
- Thomas,J. *et al.* (1990) The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **18**, 2633–2642.
- Vidovic,I. *et al.* (2000) Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell*, **6**, 1331–1342.
- Wagner,E.G. and Flärdh,K. (2002) Antisense RNAs everywhere? *Trends Genet.*, **18**, 223–226.
- Zadeh,J. *et al.* (2010) Nupack: analysis and design of nucleic acid systems. *J. Comput. Chem.*, doi:10.1002/jcc.21596.