RNA-RNA interaction prediction based on multiple sequence alignments

Andrew X. Li¹, Manja Marz², Jing Qin³, Christian M. Reidys¹,4*

¹Center for Combinatorics, LPMC-TJKLC, Nankai University Tianjin 300071, P.R. China
²RNA Bioinformatics Group, Philipps-University Marburg, Marbacher Weg 6, 34037 Marburg, Germany
³Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
⁴College of Life Science, Nankai University Tianjin 300071, P.R. China.

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

ABSTRACT

Motivation Many computerized methods for RNA-RNA interaction structure prediction have been developed. Recently, \(O(N^2)\) time and \(O(N^3)\) 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 and sequence covariation.

Results We present the a priori folding algorithm ripalign, whose input consists of two (given) multiple sequence alignments (MSA). ripalign outputs (1) the partition function, (2) base-pairing probabilities, (3) hybrid probabilities and (4) 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. Furthermore, we incorporate possible structure constraints as input parameters into our algorithm.

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 Christian Reidys duck@santafe.edu

Keywords multiple sequence alignment, RNA-RNA interaction, joint structure, dynamic programming, partition function, base pairing probability, hybrid, loop, RNA secondary structure.

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 noncoding RNAs, RNA-RNA interactions in cellular metabolism are gaining in prominence. Typical examples of interactions involving two RNA molecules are snRNAs (Förne et al. 1996); snoRNAs with their targets (Bachellerie et al. 2002); micro-RNAs from the RNAi pathway with their mRNA target (Ambros 2004; Murchison and Hannon 2004); sRNAs from Escherichia coli (Hershberg et al. 2003; Repollet 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 2004), 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 (Hofacker 1994; Bernhart et al. 2006), pairfold (Andronescu et al. 2005), and NUIFack (Ren et al. 2009) 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 (Chitase et al. 2009). 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 (Muckstein et al. 2008, 2008) and interRNA (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 structures of each molecule are pseudoknot-free, the intermolecular binding pairs are noncrossing and there exist no so-called “zig-zags”, see supplement material (SM) 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 Huang et al. (2009b) and Huang et al. (2009), while the base-pairing probabilities are due to Huang et al. (2009a).

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 nontrivial 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 ($\geq 3$) 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 (MSA).

In an MSA homologous nucleotides are grouped in columns, where homologous is interpreted in both: structural as well as evolutionary sense. i.e. 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. (2010). Therefore we can associate a consensus interaction structure to pairs of interacting MSAs (see Section 2.1).

Along these lines Seemann et al. (2010) 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. (2004) and computational strategies for hierarchical folding Gaspin and Westhof (1995) Jabbari et al. (2007). However, PETcofold is an approximation algorithm and further differences between the two approaches will be discussed in Section 2.2.

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 zip 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.

---

<table>
<thead>
<tr>
<th>sp.</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>AGAACGGA</td>
<td>AGAACGGA</td>
</tr>
<tr>
<td>S</td>
<td>GGGCCG</td>
<td>AAGAACGGA</td>
</tr>
<tr>
<td>S</td>
<td>AGAACGGA</td>
<td>GGCCG</td>
</tr>
</tbody>
</table>

Table 1. Preprocessing in ripalign: Given a pair of MSAs $(R, S)$, where $R$ consists of three aligned RNA sequences of species (sp.) $\theta_1$ or $\theta_2$. $S$ in turn consists of four aligned sequences of species $\theta_1$ and $\theta_2$. Then we obtain the matrix-pair $(R, S)$, where $(R_i, S_j), 1 \leq i \leq 6$, ranges over all the six potentially interacting RNA-pairs.

---

One important step consists in identifying the notion of a joint structure compatible to a pair of interacting MSAs. Our notion 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 Multiple sequence alignments and compatibility

A MSA, $R$, consists of $m_R$ RNA sequences of known species. Denoting the length of the aligned sequences by $N$, $R$ constitutes a $m_R \times N$ matrix, having $5' - 3'$ oriented rows, $R_i$ and columns, $R_j$, its $(i,j)$-th entry, $R_{ij}$, is a nucleotide, A, U, G, C or a gap denoted by ..

For any pair $(R, S)$ we assume that $S$ is a $m_S \times M$ 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 $(R, S)$ can interact if for any row $R_i$ there exist at least one row in $S$ that can interact with $R_i$.

Given a pair of interacting MSAs $(R, S)$, let $m$ be the total number of potentially interacting pairs. ripalign exhibits a pre-processing step which generates a $m \times M$-matrix $R$ and a $m \times M$-matrix $S$ such that $(R_i, S_j)\forall i \leq m, j \leq M$ range over all $m$ potentially interacting RNA-pairs, see Tab. 1 and the SM, Section 1.2.

We shall refer in the following to $R$ and $S$ as MSAs ignoring the fact that they have multiple sequences.

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

A joint structure $J(R, S, I)$ is a graph consisting of:

(1) 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 half-plane, respectively. We consider $R$ over a $5'$ to $3'$ oriented backbone $(R_1, \ldots, R_N)$ and $S$ over a $3'$ to $5'$ oriented backbone $(S_1, \ldots, S_M)$ and refer to any $R$- and $S$-arcs as interior arcs.

(2) An additional set $I$, of noncrossing arcs of the form $R_iS_j$ (exterior arc), where $R_i$ and $S_j$ are unpaired in $R$ and $S$.

(3) $J(R, S, I)$ contains no “zig-zags” (see SM).

The subgraph of a joint structure $J(R, S, I)$ induced by a pair of subsequences $(R_i, R_{i+1}, \ldots, R_j)$ and $(S_h, S_{h+1}, \ldots, S_k)$ is denoted by $J_{i,j,h,k}$. In particular, $J(R, S, I) = J_{1,N+1,M}$ and $J_{i,j,h,k} \subset J_{a,b,c,d}$ if and only if $J_{i,j,h,k}$ is a subgraph of $J_{a,b,c,d}$ induced by $(R_i, \ldots, R_j)$ and $(S_h, \ldots, S_k)$. In particular, we use $S[i,j]$ to denote the subgraph of $J_{1,N+1,M}$ induced by $(S_i, S_{i+1}, \ldots, S_j)$, where $S[i,j] = S_i$ and $S[i,j - 1] = \emptyset$. 

---
Given a joint structure, \( J_{a,b,c,d} \), a tight structure (TS), \( J_{i,j,h,l} \), \( \text{Huang et al.} [2009] \) is a specific subgraph of \( J_{a,b,c,d} \) indexed by its type \( \in \{ \circ, \triangledown, \square, \triangle \} \), see Fig. 1 for instance, we use \( J_{i,j,h,l} \) to denote a TS of type \( \square \).

A hybrid is a joint structure \( J_{h}^{HY} \), i.e. a maximal sequence of intermolecular interior loops consisting of a set of exterior arcs \( \{R_{i}, S_{j,1}, \ldots, R_{i}, S_{j,l}\} \) where \( R_{i}, S_{j,h} \) is nested within \( R_{i}, S_{j,h} \) and where the internal segments \( R_{i, j_{h} + 1, j_{h} + 1 - 1} \) 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 \( (R, S) \)-compatible joint structures. In difference to single sequence compatibility, this notion involves statistical information of the MSAs.

The key point is in specifying under which conditions two vertices contained in \( \{R_{i,1}, \ldots, R_{i,y} \} \) \( \times \) \( \{S_{j,1}, \ldots, S_{j,y} \} \) can pair. This is obtained by a generalization of the RNAalifold approach \( \text{Hofacker et al.} [2002] \). We specify these conditions for interior (\( c_{R}^{i,j} \)), exterior (\( c_{R}^{i,j} \)) and interior pairs (\( c_{R}^{i,j} \)) in eq. (2.1-2.5).

For interior arcs \( (R_{i}, R_{j}) \), let \( X,Y \in \{A, U, G, C\} \). Let \( f_{R}^{i,j}(XY) \) be the frequency of \( (X,Y) \) which exists in the 2-column sub-matrix \( (R_{i}, R_{j}) \) as a row-vector and

\[
C_{R}^{i,j} = \sum_{X \in X, Y \in Y} f_{R}^{i,j}(XY)D_{XY \times Y}^{i,j}(X', Y').
\]

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

\[
d_{R}^{i,j} = 1 - \frac{1}{m} \sum_{x \in X, y \in Y} \left( \Pi_{R}^{i,j}(x, y) + \delta(R_{i}^{h}, \text{gap}) \delta(R_{j}^{h}, \text{gap}) \right),
\]

where \( \delta(x, y) \) is the Kronecker delta and \( \Pi_{R}^{i,j}(x, y) \) is equal to 1 if \( R_{i}^{h} \) and \( R_{j}^{h} \) are Watson-Crick or GU wobble base pair and 0, otherwise. Now we obtain \( B_{R}^{i,j} = c_{R}^{i,j} - \phi_{4} \phi_{4} \). Based on sequence data, the threshold for pairing \( B_{R}^{i,j} \) as well as the weight of inconsistent sequences \( \phi_{4} \) are computed we have

\[
(c_{R}^{i,j} \geq B_{R}^{i,j} \geq c_{R}^{i,j} - B_{R}^{i,j} + B_{R}^{i,j})
\]

The case of two positions \( S_{i} \) and \( S_{j} \) is completely analogous

\[
(S_{R}^{i,j} \geq B_{R}^{i,j} \geq S_{R}^{i,j} - B_{R}^{i,j} + B_{R}^{i,j})
\]

where \( B_{R}^{i,j} \) and \( B_{R}^{i,j} \) are analogously defined.

As for \( c_{R}^{i,j}(RS) \) a further observation factors in: since many ncRNA show clear signs of undergoing compensatory mutations in the course of evolution.

\textbf{Fig. 2. Interior loop energy:} An interior loop formed by \( R_{i}, R_{j} \) and \( R_{l}, R_{j} \), where \( i < h < l < j \) are the alignment positions. Grey bands are used to denote the positions we omit between segment \((i, h), (h, l) \) and \((l, j)\).

\textbf{Seemann et al.} [2013], \textbf{Marz et al.} [2008], 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

\[
B_{R}^{i,j} \geq B_{R}^{i,j} \geq B_{R}^{i,j} + G_{\text{int}}^{i,j}(h, l, \epsilon, \Delta),
\]

where \( B_{R}^{i,j} \) and \( B_{R}^{i,j} \) are analogously defined as the case for \( B_{R}^{i,j} \) and \( B_{R}^{i,j} \).

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

\section{Energy model}

According to \( \text{Huang et al.} [2004] \) 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 hybrid and kissing-loops. Following the energy parameter rules of \( \text{Mathews et al.} [1999] \), 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 which depend uniquely on the terminal bases (sequence dependent).

Suppose we are given a joint structure \( J \), compatible to a pair \( P = (R, S) \). Let \( L \in J \) be a loop and let \( J_{L} \) represent the loop energy of the \( i \)-th interaction-pair \( (R_{i}, S_{j}) \). Then the loop energy of \( P \) is

\[
J_{L} = \frac{1}{m} \sum_{i} J_{L_{i}}.
\]

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

\[
J = \sum_{L \in J} J_{L}.
\]

To save computational resources, gaps are treated as bases in \texttt{zipalign}. Thus only alignment positions contribute as indices and loop sizes. Since no measured energy parameters for nonstandard 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 \( R_{l}, R_{j} \) and \( \text{Int}_{i,j,h,l} \) denote the free energy of \( \text{Int}_{i,j,h,l} \) with respect to the aligned sequences in \( P \). Then \( \text{Int}_{i,j,h,l} \) associated to the three aligned subsequences of Fig. 2 reads

\[
\text{Int}_{i,j,h,l} = \frac{1}{3} \left( 3 G_{\text{int}}^{i,j}(h, l, \epsilon, \Delta) + G_{\text{int}}^{\text{g}, \text{G}, \text{C}, \text{G}, \text{G}, \text{G}, \text{G}} + G_{\text{int}}^{\text{g}, \text{G}, \text{G}} + G_{\text{int}}^{\text{g}, \text{G}} \right).
\]

Here \( G_{\text{int}}^{i,j}(h, l, \epsilon, \Delta) \) represents contributions related exclusively to the positions of the interior loop while \( G_{\text{int}}^{\text{g}, \text{G}, \text{G}, \text{G}, \text{G}, \text{G}, \text{G}, \text{G}} \) represents additional contributions related to the specific nucleotides which form the interior loop. We set \( G_{\text{int}}^{\text{g}, \text{G}, \text{G}} \) to be zero.

\section{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 eq. (2.1). 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 ripalign-gra mmar would oftentimes encounter joint structures having a hybrid composed by a single isolated external arc, see Fig. 3.

In order to decompose canonical joint structures via the unambiguous grammar introduced in Section 2.3, we distinguish the two types (Type cc and Type c) of TS's of type ▽, △ or □. Given a TS of type ▽, denoted by $J_{i,j,h,t}$, we write depending on whether $R_{i+1}j_{j-1} \in J_{i,j,h,t}$, $J_{i,j,h,t}$ and $J_{i,j,h,t}$, respectively. Analogously, we define $J_{i,j,h,t}$ and $J_{i,j,h,t}$, respectively.

Fig. 3. Examples of two TS-types. We display ▽, □, or △-tight structures: Type cc (top) and Type c (bottom).

### 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 backtracking procedure that can be used to sample from the associated distribution: McCaskill 1990, Ding and Lawrence 2003, Huang et al. 2010. We remark that the time complexity does not increase linearly as a function of $n$ (see SM Table. 5).

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:

1. a position $i$ can be restricted to form an interior or an exterior arc, denoted by “−” and “△”, respectively;
2. a position $i$ can be forced to be unpaired (denoted by “X”);
3. a position $i$ can be restricted to form an interior or an exterior arc with some position $j$ (denoted by “△”);
4. a pair of positions $i$ and $j$ can be forced to form an interior or exterior arc (denoted by “△” or “△”).

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 Fig. 3.

### 3 RESULTS AND DISCUSSION

In this paper we present an a priori $O(N^6)$ time and $O(N^4)$ space dynamic programming algorithm ripalign, whose input consists of a pair of interacting MSAs. ripalign requires only marginally more computational resources but is, without doubt, still computationally costly. Approximation algorithms are much faster, for instance PECofold Seemann et al. 2010, having a time complexity of $O(m(N + M)^3 n)$, where $m$ is the number of sequences in MSA, $N$ and $M$ being the sequence lengths 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 is 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 however, the uncertainty of the quality of the solution. Point in case here is that the two secondary structures did not evolve independently, but rather correlated by means of their functional interaction. We remark that ripalign (within its complexity limitations) is capable to describe the space of RNA interaction structures, for instance via Boltzmann sampling, in detail and transparency.

ripalign represents significant improvements in the following aspects:
(a) we incorporate evolutionary factors into the RNA-RNA interaction structure prediction via alignments as input,
(b) we introduce the grammar of canonical joint structures of interacting-alignments,
(e) we a priori factor in structural-constraints, like for instance, knowledge on Sm-binding sites.

Below we shall discuss (a), (b) and (e) in more detail in the context of concrete examples. All the MSAs involving in (a), (b) and (e) are listed in SM, Section 2.

(a): The flhA/OxyS RNA-repressing flhA mRNA translation initiation through base-pairing with two short sequences [Agrawam and Altunia 2000], one of which overlaps the ribosome binding sequence and the other resides further downstream, within the coding region of flhA. Our algorithm predicts correctly both interaction sites based on MSAs, see Fig. 5. In addition, most predicted stacks in the secondary structures of flhA and OxyS agree well with the most frequent Boltzmann sampled structure. Two more hybrids, $J_{11}^h$ and $J_{48}^h$, are predicted in our output. 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 flhA, instead of the unpaired segment occurring in the natural structure, can be understood in the context of minimizing free energy. Comparing the prediction based on the MSAs (Fig. 6 middle) with the one based on the consensus sequence (Fig. 6 bottom), we observe:
(1) the secondary structure of flhA agrees better with the annotation joint structure (Fig. 6 top),
(2) the leftmost hybrid agrees better with that of the annotated structure,
(3) the binding-site probability (see SM, Section 5, eq. (5.5)) of the leftmost hybrid increases by nearly 40%.

On the flip side, due to the gaps in seven out of eight subsequences induced by R[98, 102] (Column 98-102 in flhA), the prediction quality of the right-most hybrid and its corresponding contact-region probability decreases slightly.

Let us next contrast our results with those of PETcofold, see Fig. 7. The latter predicts one of the two interaction sites. The second site is predicted subject to the condition that constrained stems were not extended [Seemann et al. 2010]. It can furthermore be observed that in order to predict the second hybrid, at the same time the secondary structures prediction of both flhA and OxyS gets worse. ripalign predicts both: the interaction sites situated in flhA and comes close to predicting the secondary structures of flhA as well as OxyS without any additional constraints.

(b): The SmY-10/SL-1 interaction of C. elegans

MacMorris et al. 2007 stipulated that SmY-10 RNA, possible involved in trans-splicing, interacts with the splice leader RNA (SL1 RNA). In Fig. 8 we show that the Sm-binding sites (colored in red) of the RNA molecules SmY-10 and SL-1 are R[56, 62] and S[25, 31], respectively. In Fig. 8 the top structure is being predicted by ripalign without any structure constraint. The hybrids listed in Column III are predicted by ripalign under the structural constraints that $5'-AUAUUUG-3'$ and $3'-GUUUA-5'$ (S[25, 31]) are Sm-binding sites (colored in red) in SmY-10 and SL-1, respectively. Here, we use $J_{i,j}^h, J_{h,l}$ to denote the hybrid induced by $R[i, j]$ and $S[h, l]$.

(c): The U4/U6 interaction

Two of the snRNAs involved in pre-mRNA splicing, U4 and U6, are known to interact by base pairing [Zucker-Aprison et al. 1988]. We divided all known metazoan U4 and U6 snRNAs into three distinct groups and alignments: protostomia without insects, insects and deuterostomia [Mart et al. 2008; Mart et al. 2008] observed that insects behave in their secondary structure different from other protostomes, see Fig. 8. Comparing all the predicted U4/U6 interactions, displayed in Fig. 8 we can conclude:
(1) the secondary partial structures of the U4/U6 complex for all three groups predicted by ripalign agree predominantly with the described secondary structures in metazoa, [Thomas et al. 1999; Drake et al. 2002; Shambeau et al. 1994; López et al. 2008; Shukla et al. 2008], e.g. as depicted in Fig. 6 (top) for C. elegans [Zucker-Aprison et al. 1988].
(2) for all three groups, Stem I and II (Fig. 6) are highly conserved. External ascendancies, such as protein interactions may stabilize stem II additionally.
(3) for all three groups, the 5‘ hairpin of U4 snRNA seems highly conserved to interact with the U6 snRNA. This RNA feature is not fully understood, since this element is also believed to contain intraloop interactions and may bind to a 15.8kDa protein [Vidovic et al. 2003].
(4) for all metazoans, the U6 snRNA shows conserved intramolecular interactions between the 5‘ part and the region downstream of the 5‘-hairpin.
(5) for deuterostomes (Fig. 9) bottom, with a contact-region probability of 45.5%, our algorithm identifies a third U4/U6 interaction, Stem III, to be conserved, which agrees with the findings in Jakab et al. 1997.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J_{41}^h</td>
<td>J_{40}^h</td>
<td>J_{40}^h</td>
</tr>
<tr>
<td>2</td>
<td>J_{40}^h</td>
<td>J_{40}^h</td>
<td>J_{40}^h</td>
</tr>
<tr>
<td>3</td>
<td>J_{76}^h</td>
<td>J_{78}^h</td>
<td>J_{78}^h</td>
</tr>
<tr>
<td>4</td>
<td>R_{11}S_{10}</td>
<td>J_{11}^h</td>
<td>J_{11}^h</td>
</tr>
<tr>
<td>5</td>
<td>J_{16}^h</td>
<td>J_{78}^h</td>
<td>J_{78}^h</td>
</tr>
<tr>
<td>6</td>
<td>J_{34}^h</td>
<td>J_{57}^h</td>
<td>J_{16}^h</td>
</tr>
</tbody>
</table>
We finally remark that the quality of prediction of \( \text{PETcofold} \) for both: protostomia (without insects) and deuterostomes, the region probability of \( \text{PETcofold} \) can also be assumed.

REFERENCES


Brow and Vidaver (1995). For protostomes, a similar feature with a contact-

Brow and Vidaver (1995). For protostomes, a similar feature with a contact-

the rather simplistic model of covariance scoring with more sophisticated

the rather simplistic model of covariance scoring with more sophisticated

We want to thank Fenix W.D. Huang and Jan

We want to thank Fenix W.D. Huang and Jan

We finally remark that the quality of prediction of \( \text{PETcofold} \) for both: protostomia (without insects) and deuterostomes, the region probability of \( \text{PETcofold} \) can also be assumed.

REFERENCES


Brow and Vidaver (1995). For protostomes, a similar feature with a contact-

Brow and Vidaver (1995). For protostomes, a similar feature with a contact-

the rather simplistic model of covariance scoring with more sophisticated

the rather simplistic model of covariance scoring with more sophisticated

We want to thank Fenix W.D. Huang and Jan

We want to thank Fenix W.D. Huang and Jan
**Fig. 8. ripalign versus rip:** Interaction of two specific RNA molecules, SL-1 and SmY-10 of Caenorhabditis elegans. The Sm-binding sites (colored in red) in the RNA molecules SmY-10 and SL-1 are 5’-AAUUUUUG-3’(R[56, 62]) and 3’-GUUUUA-A-S5(S[25, 31]), respectively. The joint structure contain a single interior arc R24,576(top) is predicted by rip implemented by Huang et al. [2014]. The joint structure (middle) is predicted by ripalign without any structural constraint. The joint structure (bottom) is predicted by ripalign under the structural constraints that 5’-AAUUUUUG-3’(R[56, 62]) and 3’-GUUUUAA-S5(S[25, 31]) are Sm-binding sites in the RNA molecules SmY-10 and SL-1, respectively. The target site (green boxes) probabilities computed by ripalign are annotated explicitly if > 10% or just by ≤ 10%, otherwise.


**Fig. 9. The U4-U6 interaction prediction with Sm-binding site constraint in U4.** The Sm-binding site in molecule U4 is 5’-AAUUUUUG-3’(colored in red). Top of the figure is the natural structure of U4/U6 of C. elegans depicted by Zucker-Apinson et al. [1988], in which the stem I, stem II and Sm-binding site are colored in green and red, respectively. The joint structures of protostomia (without insects), insects and deuterostomia (from top to bottom) are predicted by ripalign under the Sm-binding site constraint. The target site (green boxes) probabilities computed by ripalign are annotated explicitly if > 10% or just by ≤ 10%, otherwise.