Concurrent prediction of RNA secondary structures with pseudoknots and local 3D motifs in an integer programming framework

Gabriel Loyer [1] and Vladimir Reinharz [1,*]
[1]Computer Science, Université du Québec à Montréal, Canada.
[*] reinharz.vladimir@uqam.ca

Abstract

Motivation: The prediction of RNA structure canonical base pairs from a single sequence, especially pseudoknotted ones, remains challenging in a thermodynamic models that approximates the energy of the local 3D motifs joining canonical stems. It has become more and more apparent in recent years that the structural motifs in the loops, composed of non-canonical interactions, are essential for the final shape of the molecule enabling its multiple functions. Our capacity to predict accurate 3D structures is also limited when it comes to the organization of the large intricate network of interactions that form inside those loops.
Results: We previously developed the integer programming framework RNA-MoIP (RNA Motifs over Integer Programming) to reconcile RNA secondary structure and local 3D motif information available in databases. We further develop our model to now simultaneously predict the canonical base pairs (with pseudoknots) from base pair probability matrices with or without alignment. We benchmarked our new method over the all non-redundant RNAs below 150 nucleotides. We show that the joined prediction of canonical base pairs structure and local conserved motifs (i) improves the ratio of well predicted interactions in the secondary structure, (ii) predicts well canonical and Wobble pairs at the location where motifs are inserted, (iii) is greatly improved with evolutionnary information and (iv) can recover kink-turn locations.
Availability: The source code of the framework is available at: `https://gitlab.info.uqam.ca/cbe/RNAMoIP`

1 Introduction

The rise of RNA therapeutics [1, 2] is due to technical and computational advances in our understanding of RNA sequence-structure-function paradigm. While the prediction of all-atoms RNA structure from sequence still remains a challenge [3], many different approaches have allowed to reach interesting results in various facets of the problem.

Taking advantage of the hierarchical folding of RNA, the secondary structure composed of strong stems of canonical and Wobble base pairs form first [4], many efficient theoretical approaches have been developed to predict this secondary structure. Nonetheless, the most accurate and feasible models as in the ViennaRNA package [5] or RNAstructure [6] assume that there are no crossing interactions, no pseudoknots, since that assumption adds complexity and decreases accuracy in the thermodynamic parameters making it often impractical to use. Yet pseudoknots are abundant and important. To predict accurate secondary structure with

them would be invaluable for the main 3D reconstruction tools that rely on that secondary structure [7].

The prediction of RNA with pseudoknots is in all generality in the nearest neighbour model NP-Hard [8]. The dynamic programming algorithm solving exactly the minimal free energy (MFE) structure problem and with the most general classes of pseudoknots is PKnots [9] with prohibitive complexity in time of $\mathcal{O}(n^6)$ and space of $\mathcal{O}(n^4)$. Instead of exact methods heuristics have also been developed as Hotknot [10]. Reducing the set of achievable pseudoknot configurations (while keeping known important ones) combined with sparsification techniques Knotty [11] is able to achieve a time complexity of $\Theta(n^3 + Z)$ (with in practice $Z < \mathcal{O}(n^4)$). While deep neural networks as SPOT-RNA [12] on the subjects have been published, rigorous benchmarks show that they still lack generability [13]. By formulating the problem as an Integer Programming problem IPknot [14] has proven to be effective to predict general pseudoknotted secondary structures when maximising base pair probabilities generated by RNAfold [5], leveraging fast general solvers and without sacrificing possible outputs. More recent work improved IPknot using clever heuristic to solve long sequences in reasonable time, combining a dynamic threshold with a linear-time approximation of the RNA folding partition [15].

Beyond the secondary structure RNA are composed of many different important interactions. The Leontis-Westhof (LF) classification [16] defines 12 types of possible base pairs, between any nucleotides. When describing the loops between the rigid stems using those non-canonical interactions, different methods have shown that conserved sub structures are present and important [17, 18, 19]. These databases of motifs can be leverage for not only more accurate structure prediction, but also to include geometric information beyond canonical and Wobble base pairs.

Previous work used conserved structural motifs to select an optimal secondary structure and ease all-atoms 3D reconstruction [20, 21]. In this paper we expend on that IP framework to achieve **the simultaneous prediction of secondary structure with pseudoknots and structural motifs insertion** incorporating ideas from IPknot [14] and using a newly designed local structural modules dataset computed from [22].

In Sec. 2.3 the unified IP equations are presented. We discuss in Sec. 3.9 how a sequence alignment can be used, a feature implemented in our software. We benchmark the secondary structure prediction on all known non-redundant RNA with a determined pseudoknotted structure below 150 nucleotides in Sec. 3.4, and how the predictions fare for the canonical and non-canonical interactions inside the motifs in Sec. 3.5 and 3.6. We then evaluate how a good sequence alignment can improve the prediction in Sec. 3.9 with a set of hand aligned structures by Rfam. We conclude on analysis in Sec. 3.10 by looking specifically at the kink-turn motif that was present in 4 previous structures and how using an alignment or not influences its prediction.

2 Methods

Let ω an RNA sequence, and Ω its secondary structure. A base pair $(i, j) \in \Omega$ must be canonical (G-C or A-U) or Wobble (G-U), and have $j - i > 3$. The Leontis-Westhof (LW) classification of RNA base pairs [16] defines 12 different geometries possible combining two sides between Watson-Crick (W), Hoogsteen (H), Sugar (S) and an orientations cis (c) or trans (t). The canonical and Wobble base pairs are all of type cis Watson-Crick/Watson-Crick (cWW). Generally any combination of nucleotides can form any type of base pair. Stability in the nearest neighbour model is obtained from stacked base pairs [23], forbidding lonely base pairs implies formally that if $(i, j) \in \Omega \Rightarrow (i - 1, j + 1) \in \Omega$ or $(i + 1, j - 1) \in \Omega$.

The secondary structure Ω can be decomposed in an ensemble of pseudoknot-free structures

$\Omega^1, \ldots, \Omega^m$. Ω^q is pseudoknot-free if there is no crossing between any base pairs, formally for all $(i,j),(k,l) \in \Omega^q \Rightarrow i < k < l < j$ or $k < i < j < l$.

The workflow of our framework is as follows and schematized in Fig. 1:

1. Given a sequence ω and a secondary structure Ω simultaneously:
   (a) Decomposed Ω in pseudoknot-free structures, for each compute a constrained version of a classic folding algorithm to obtain a base pairing probability matrix (if no structure is provided, the algorithm runs without constraints once), sum all the matrices.
   (b) Find all possible motifs location using pattern match in the input sequence ω.
2. Solve the IP model to find the optimal combination of base pairs with pseudoknots and motifs given our objective function detail in Sec. 2.3.3.
3. Iterate until:
   (a) Two iterations give the identical solution.
   (b) A threshold in the number of iterations or time is reached.

2.1 Structural motifs

The validation of structure prediction using motifs database is challenging to ensure that the benchmark is not biased. We decided to use the same database as in the first version of RNA-MoIP built in 2012 [20]. Additional measures to avoid overfitting are discussing in Sec. 3.3.

This database is based on the detection of similar networks of interactions among all RNA 3D structures [19]. The entire dataset is built upon 398 unique common subgraphs (or RINs, Recurrent Interaction Networks), that can be divided into 5278 different nucleotide sequences. Those sequences are composed of multiples strands, which can represent motifs like hairpins, interior loops, and multi-loops.

2.2 Base pair probabilities

Different tools as the ones provided by ViennaRNA [5] and RNAstructure [6] can accurately, in a thermodynamic setting, utilize the most recent set of nearest neighbor parameters to compute base pairing probabilities for pseudoknot-free structures. Following the model of IPknot [14], the secondary structure Ω is decomposed in an ensemble of pseudoknot-free structures $\Omega^1, \ldots, \Omega^m$. For each, a base pair probability (BPP) matrix can be computed such that the base pairs in that sub-structure are enforced as hard constraints, and any position in another sub-structure are forbidden to pair. In the IP formulation, the weight of the base pair (i,j) will be the pseudo-probability $p(i,j) = p^1_{i,j} + \cdots + p^m_{i,j}$. Note that during the evaluation of the BPPs the pairing are considered as hard constrained, they must be preserved, there is no such condition in the IP model.

2.3 Integer Programming Model

The integer programming model is quite complex and we reproduce here all equations for sake of completeness. Sec. 2.3.1 describes how the motifs are encoded. The Sec. 2.3.2 lists all the model variables. Then the objective function is detailed in Sec. 2.3.3. The constraints in regards to the base pairs placement in Sec. 2.3.4 and the ones regarding the motifs insertion are in Sec. 2.3.5. **All new equations and modifications to them are in green.**

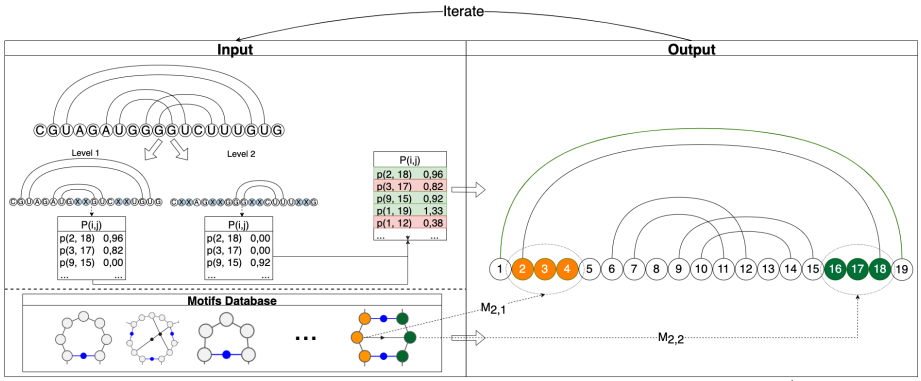


Figure 1: **The RNA-MoIP workflow** Left top: the sequence with a structure (which can be empty). The structure is decomposed in pseudoknot-free sub-structures, for each a constrained BPP is computed, and they are all summed together. Left bottom: a database of structural motifs containing hairpins, interior loops and bulges, and k-way junctions. Right: output an optimal combination between a secondary structure with pseudoknots and motifs inserted in sequence compatible locations. Each motif strand must stack or overlap by 1 position a base pair in the secondary structure.

2.3.1 Input

We denote an RNA sequence as ω and Ω a secondary structure compatible with it. ω_i is the nucleotide at position i and must be in $\{A, C, G, U\}$. The structure can be empty and may contain crossing interactions. We use $n = |\omega|$ as the length of the sequence.

Each motif in the database can be composed of a set of different sequence strands (e.g. an interior loop has two strands). The equation will differentiate between hairpins (1 strand), interior loops and bulges (two strands), and k-way junctions (3 or more strands). Mot^j is the set of motifs with j-strands and each is composed of its list of strands. For any motif $x \in Mot$, the length $|x|$ represents how many nucleotides it contains. Formally, a position x_a^b in a strand can be A, C, G, U or the wild card *, and we have:

$$Mot^j = \{x \mid x := [(x_1^1, \cdots, x_{k_1}^1), \cdots, (x_1^j, \cdots, x_{k_j}^j)]$$
$$\text{and } \exists \text{ a } \textit{match} \text{ of } x \text{ in } \omega\}$$

The strands are ordered in the $5'$ to $3'$ order of the sequence they are extracted from. The model needs to know where the $i-th$ strand of any motif composed of j-parts can be inserted. These can be of different lengths, so the ensemble Seq_i^j will be a set of triplets with the name of the motif and the first and last positions where the $i-th$ strand of that motif can be inserted. The same strand can be placed in many different places. Formally:

$$Seq_i^j = \{(x, a, a + k_i - 1) \mid$$
$$x \in Mot^j \text{ and } x_1^i, \ldots, x_{k_i}^i = \omega_a, \ldots, \omega_{a+k_i-i}\}.$$

2.3.2 Variables

To keep in line with the previous implementation of RNA-MoIP, two type of binary variables are being tracked. First, for the motifs 4 things needs to be kept track of: (1) x the name of the motif (2) if it is the i-th strand of that motif, and (3,4) the interval k, l where it inserted. This will be kept done by the variable $C_{k,l}^{x,i}$ representing the insertion of the i-th component

of the motif x at position (k, l) of the sequence. One such variable exists for every element of every set Seq_i^j. Second, for every pair of position where the BPP is above a certain threshold, the model needs to know if it is instantiated and at which level. Each level will contain a pseudoknot-free structure, and will have to be crossing a base pair in each level below it. The binary variable $D_{u,v}^q$ will be 1 if there is a base pair between ω_u and ω_v at level q. The set $\mathcal{B}$ will contain all pairs of positions (u, v) with a potential base pair.

2.3.3 Objective

Intuitively, we want to maximize the pseudo-base pair probabilities as the amount of information in the sequence. In [20], this is achieved by maximizing the square of the length of the motifs inserted, which will push to insert the smallest amount of largest motifs possible. As in Sec. 2.3.1 the number of nucleotides in motif x is denoted $|x|$. For each potential base pair between positions (u, v) the pseudo-BPPs probability $p(u, v)$ (see Sec. 2.2) is used. A factor of 10 was empirically added. A parameter α will be used to combine the pseudo-BPPs maximisation and motifs insertion. A weight β^q is used to balance the different level of pseudoknots. Following IPknot we select $\beta^1 = 0.5, \beta^2 = 0.25, \beta^3 = 0.125$ and $\beta^4 = 0.0625$. Formally, the objective function is:

$$\max \quad \alpha \sum_{x \in Mot^j} \left(\overbrace{(|x|)}^{\text{Motif length}}{}^{2} \times \sum_{(x,k,l) \in Seq_1^j} \overbrace{C_{k,l}^{x,1}}^{\text{Motif inserted at position (k,l)}} \right) + 10(1-\alpha) \times \sum_{(u,v) \in \mathcal{B}} \sum_{q=1}^{m} \overbrace{D_{u,v}^q}^{\text{Base pair (u,v)}} \; \underbrace{p(u,v)}_{\text{Probability of the base pair (u,v)}} \; \underbrace{\beta^q}_{\text{Weight of level q}} \tag{1}$$

2.3.4 Base pairs constraints

The first three equations will ensure that each position is only in one base pair (Eq. 2), that each level q contains a pseudoknot-free structure (Eq. 3), and that base pairs are added to a level if and only if they cross another base pair in every lower level (Eq. 4). Two additional properties we enforce are that base pairs must be stacked on the same level (Eqs. 5 and 6) and that at least 25% of positions must be in a base pair (Eq. 7). Note that naturally, due to the weights in the objective function, most of the base pairs will concentrate on lower levels.

$$\forall 1 < u < n : \sum_{q}^{m} \sum_{\substack{(\tilde{u},\tilde{v})\in\mathcal{B} \\ \tilde{u}=u \vee \tilde{v}=u}} D^{q}_{\tilde{u},\tilde{v}} \leq 1 \tag{2}$$

$$\forall 1 < q \leq m,\ \forall 1 \leq i < j < k < l \leq n : D^{q}_{i,j} + D^{q}_{k,l} \leq 1 \tag{3}$$

$\forall 1 < q \leq m,\ \forall (u,v) \in \mathcal{B}:$

$$\sum_{i<u<j<v}^{\mathcal{B}} D^{q-1}_{i,j} + \sum_{u<\tilde{i}<v<\tilde{j}}^{\mathcal{B}} D^{q-1}_{\tilde{i},\tilde{j}} \geq D^{q}_{\tilde{u},\tilde{v}}, \tag{4}$$

$\forall 1 < q \leq m,\ \forall 1 < i < n:$

$$\sum_{\substack{(u,v)\in\mathcal{B} \\ u=i-1 \vee u=i+1}} 1 - \sum_{\substack{(u,v)\in\mathcal{B} \\ u=i-1 \vee u=i+1}} D^{q}_{u,v} \geq \sum_{\substack{(u,v)\in\mathcal{B} \\ u=i}} 1 - \sum_{\substack{(u,v)\in\mathcal{B} \\ u=i}} D^{q}_{u,v} \tag{5}$$

$\forall 1 < q \leq m,\ \forall 1 < i < n:$

$$\sum_{\substack{(u,v)\in\mathcal{B} \\ v=i-1 \vee v=i+1}} 1 - \sum_{\substack{(u,v)\in\mathcal{B} \\ v=i-1 \vee v=i+1}} D^{q}_{u,v} \geq \sum_{\substack{(u,v)\in\mathcal{B} \\ v=i}} 1 - \sum_{\substack{(u,v)\in\mathcal{B} \\ v=i}} D^{q}_{u,v} \tag{6}$$

$$2\sum_{q}^{m} \sum_{(u,v)\in\mathcal{B}} D^{q}_{u,v} \geq \omega n \tag{7}$$

2.3.5 Motifs constraints

Following the original formulation of RNA-MoIP, we reproduce here all the equations necessary for the insertion of motifs. While most of them are exactly similar, there is a notable difference. One of the main conditions for the insertion of any strand is that it must be stacked or overlapping the last base pair of a strand. Since the motifs database is defined over the loops of a pseudoknot-free secondary structure, only the base pairs in the first level are considered. We give a brief overview of each equation but more details can be found in [20].

The hairpins insertion that are composed of only one strands are constrained by Eq. 8, which was modified to only consider base pairs in level 1. The insertion of interior loops and bulges must first ensure that strands are placed in acceptable positions (Eq. 9) and that the motif must fill at least 2 unpaired positions, ensuring information is added to the system 10.

$\forall_{(x,k,l)\in Seq^1_1}:$

$$C^{x,1}_{k,l} \leq \sum_{\substack{(u,v)\in\mathcal{B} \\ k-1\leq u\leq k \wedge \\ l\leq v\leq l+1}} (D^{1}_{u,v}) + \sum_{\substack{(\tilde{x},\tilde{k},\tilde{l})\in Seq^2_1 \\ \tilde{l}=k-1}} C^{\tilde{x},1}_{\tilde{k},\tilde{l}} + \sum_{\substack{(\tilde{x},\tilde{k},\tilde{l})\in Seq^2_2 \\ \tilde{k}=l+1}} C^{\tilde{x},2}_{\tilde{k},\tilde{l}} \tag{8}$$

$\forall (u,v) \in \mathcal{B},\ \forall x \in Mot^2:$

$$\begin{aligned} &-n(1-D^{1}_{u,v}) \leq \\ &\sum_{\substack{(x,k,l)\in Seq^2_1 \\ l<u \vee v<k}} C^{x,1}_{k,l} - \sum_{\substack{(x,k,l)\in Seq^2_2 \\ l<u \vee v<k}} C^{x,2}_{k,l} \leq n(1-D^{1}_{u,v}) \end{aligned} \tag{9}$$

$$\forall (x,k,l) \in Seq_1^2, \forall (x,\tilde{k},\tilde{l}) \mid$$
$$\tilde{k} > l \wedge 2 \cdot \sum_{\substack{(u,v)\in\mathcal{B} \\ k\le u\le l \wedge \tilde{k}\le v\le \tilde{l}}} 1 + \sum_{\substack{(u,v)\in\mathcal{B} \\ k\le u\le l \oplus \tilde{k}\le v\le \tilde{l}}} 1 \ge l-k+\tilde{l}-\tilde{k}+1 \in Seq_2^2 :$$
$$C_{k,l}^{x,1} + C_{\tilde{k},\tilde{l}}^{x,2} \le 1 \tag{10}$$

The k-way junctions admissibility of insertion is decided in Eq. 11, ensuring that each strand can be reached without crossing the base pairs in the first level. This is equivalent to Eq. 9 for the interior loops.

$$\forall j \ge 3, \forall (u,v) \in \mathcal{B} : \; -n(1-D_{u,v}^1) \le$$
$$(j-1) \cdot \sum_{\substack{(x,k,l)\in Seq_1^j \\ u\le k\le l\le v}} C_{k,l}^{x,1} - \sum_{\substack{1<i\le j \\ (x,k,l)\in Seq_i^j \\ u\le k\le l\le v}} C_{k,l}^{x,i} \le n(1-D_{u,v}^1) \tag{11}$$

An important feature of RNA structure is that their sequence is ordered, from the $5'$ to the $3'$ end, and that it is not symmetric. In a motif, an order is defined over the strands following that direction. The model constrains where a strand in a motif can be placed given the insertion of the previous (Eq. 12) or next (Eq. 13) strand of the same motif. An important consideration is that at the end there must exist a mutually exclusive decomposition of the strands such that each inserted motif is complete, even if many copies are found (Eq. 14).

$$\forall\, 1 < i \le j,\ \forall (x,k,l) \in Seq_i^j :\ C_{k,l}^{x,i} \le \sum_{\substack{(x,\tilde{k},\tilde{l})\in Seq_{i-1}^j \\ \tilde{l}<k-5}} C_{\tilde{k},\tilde{l}}^{x,i-1} \tag{12}$$

$$\forall\, 1 \le i < j,\ \forall (x,k,l) \in Seq_i^j :\ C_{k,l}^{x,i} \le \sum_{\substack{(x,\tilde{k},\tilde{l})\in Seq_{i+1}^j \\ l+5<\tilde{k}}} C_{\tilde{k},\tilde{l}}^{x,i+1} \tag{13}$$

$$\forall\, j > 1,\ \forall x \in Mot^j,\ \forall 1 < i \le j :$$
$$\sum_{(x,k,l)\in Seq_1^j} C_{k,l}^{x,1} - \sum_{(x,\tilde{k},\tilde{l})\in Seq_i^j} C_{\tilde{k},\tilde{l}}^{x,i} = 0 \tag{14}$$

Finally, to unify both models, it is important to avoid clashes between the motifs and base pairs at the different levels. This is the role of Eq. 15. Since the structural motifs are defined on the secondary structure, we insert them in relation to the base pairs in D^1 and forbid other base pairs to form inside them. We assume that motifs are a cohesive geometric unit.

$$\forall\, 1 < i \le j,\ \forall (x,k,l) \in Seq_i^j :$$
$$\sum_{q=1}^{m} \sum_{\substack{(u,v)\in\mathcal{B} \\ k\le u\le l \vee \\ k\le v\le l}} D_{u,v}^q \le \left(1 - C_{k,l}^{x,i}\right) \cdot \sum_{q=1}^{m} \sum_{\substack{(u,v)\in\mathcal{B} \\ k\le u\le l \vee \\ k\le v\le l}} 1 \tag{15}$$

2.4 Incorporating evolutionary information from sequence alignments

Sequence alignment contain large amount of evolutionary information that can be leveraged for better prediction. When an alignment is provided to RNA-MoIP the execution logic is slightly adapted in two ways to take this new data into account.

First, instead of relying on RNAfold, the base pairings probabilities matrix is calculated taking the alignment into account using RNAalifold [24] which is now part of the ViennaRNA Package.

Second, each motif insertion score is weighted in function of its compatibility with the alignment. A motif can now be inserted if it (a) perfectly matches the input sequence at that position, or (b) is at most at Hamming distance 1 of at least 50% of the alignment at these positions. In the objective function (Eq. 1) the term $C_{k,l}^{x,1}$ representing motif x inserted in position k, l is now weighted by the sum for each strand of the fraction of its match with the alignment.

3 Results

3.1 Implementation

The Integer Programming framework is implemented in Python 3, with an interface to facilitate usage with different solvers. In this study we used the open-source solver Or-Tools [25], giving better performance. Instructions are also provided to use the open source IP solver CBC [26] through the MIP library [27], or the proprietary Gurobi [28] solver. We ran our benchmarks on Ubuntu 21.04 on an Intel Xeon Processor W-2295 with 512GB 8x64GB DDR4 2933. The source code, data, and results, are available at `https://gitlab.info.uqam.ca/cbe/RNAMoIP`.

3.2 Dataset

For benchmarking, all RNA structures between 20 and 150 nucleotides in the PDB [29] were selected, filtering for identical sequences. To avoid molecular redundancies, we kept one structure per non-redundant class as defined by the BGSU RNA Structure Atlas [30] v3.208.

The canonical base pairs in the secondary structure can be deconvoluted in different ways into a main knot-free structure and an ensemble of pseudoknots of increasing complexity [31]. A reference secondary structure from which pseudoknots are defined was determined using RNApdbee [32].

The benchmark dataset is composed of the remaining 101 structures with at least a pseudoknot.

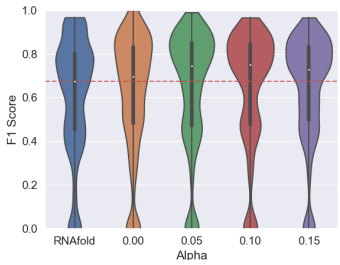
3.3 Solver

Version 2.5.0a5 of ViennaRNA is used to compute the base pair probability matrices. The terminating conditions are set to a maximum of 3 iterations, or two iterations with the same results. A time of $10^4 s$ was allotted for each prediction and sequences. While many optimal solutions can exist for one IP formulation only the first one achieved was used.

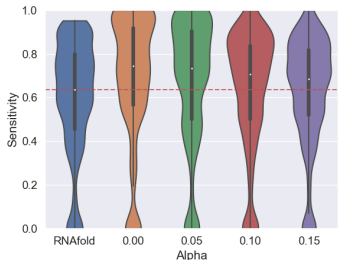
To avoid overfitting, before each sequence prediction all motifs belonging to any structure in the same RNA Structure Atlas v3.208 non-redundant class were removed [30].

3.4 Motifs insertion improve secondary structure prediction

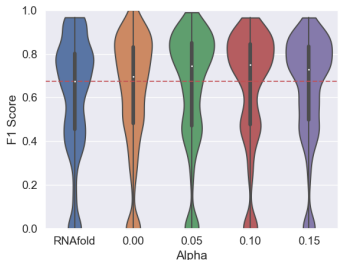
To evaluate the capacity of RNA-MoIP to predict the secondary structure, the set of True Positives (TP) consists of canonical and Wobble base pairs. The Positive Predicted Value (PPV: $\frac{TP}{TP+FP}$), sensitivity (STY: $\frac{TP}{TP+FN}$) and F1 ($\frac{2 \cdot PPV \cdot STY}{PPV+STY}$) are used as metrics and shown in Fig. 2. We compare the results of RNAfold and RNA-MoIP with different values of α. When $\alpha = 0$, no motifs are considered, and the model is equivalent to IPknot. Note that RNAfold STY can not reach 1, since only pseudoknotted structures are in the benchmark set, but RNAfold can not predict crossing interactions. And this is what we observe, where it has lower STY than almost all values of α. On the other hand, RNAfold predictions are in general more sensitive than the IP model, especially when no motifs are used. The average values over all finished models are shown in Table 1. The optimal F1, balancing the amount of base pairs predicted and their sensitivity, is achieved with $\alpha = 0.1$, complementing the base pairs with motifs information.



(a) PPV



(b) STY



(c) F1

Figure 2: **Predicting secondary structure with pseudoknots** Comparing results for RNAfold (can not predict crossing interactions), without motifs insertion ($\alpha = 0$), and for different values of α. When $\alpha > 0$, all base pairs inserted in the same motifs are counted as true positives.

	RNAfold	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$
PPV	0.63	0.57	0.62	0.654	**0.671**
STY	0.598	**0.683**	0.666	0.632	0.625
F1	0.607	0.616	0.638	0.638	**0.641**

Table 1: **Predictions results summary**.

The statistics shown previously are computed over all the secondary structure. Our models allow up to 2 crossing levels between the base pairs yet 95 of the benchmarked structures have only 1 level of crossing interactions. No over-prediction of the pseudoknots level was observed, as shown in Table 2. Pseudoknots are in fact usually under-predicted. When no motifs are inserted, around 15% of structures have a pseudoknot level too low, up to 50% when motifs are added to the model. Nonetheless, the improvement in PPV and of over 10% in F1 measure indicates that although no pseudoknot is predicted the structure is much more accurate.

α	0	0.05	0.1	0.15
PK lvl too low	25	51	62	68
PK lvl correct	76	50	39	33

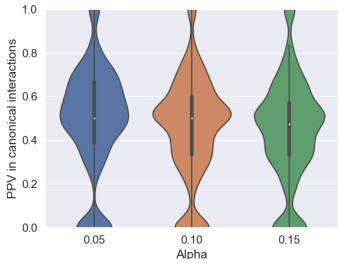
Table 2: **Predicting pseudoknot lvl** As α is increased, we underestimate the lvl of pseudoknots that are present in the structure. The complexity of the IP model increases and makes it more challenging to find a feasible optimal solution in time.

3.5 Recovering canonical and Wobble interactions in the motifs

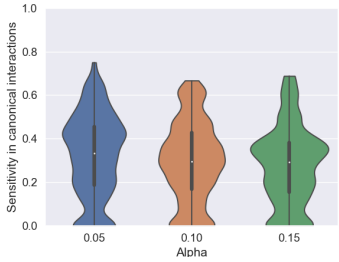
For each motif inserted, we retrieved from the RNA structure atlas all canonical and Wobble interactions at the inserted positions to build our positive examples. Note that these interactions can be crossing inside the loops and do not need to be stacked, therefore they are not necessarily part of the secondary structure. Since a motif sequence in our database can match different sub-structures, the one with the best structural match was used in this and the subsequent section. The same metrics as previously, PPV, STY, and F1 are computed and averaged over all motifs in all structures and shown in Fig. 3. Inside the motifs, an F1 value around 40% is achieved, a slightly lower precision for predicting the canonical and Wobble pairs in the motifs than the pseudoknotted secondary structure.

3.6 Non-canonical interactions remain challenging to predict

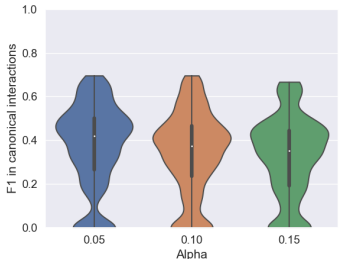
While in the minority and hard to predict, non-canonical interactions have been shown to be necessary for many RNA functions. As in the previous section, for each motif the positive set consists in the ensemble of all non-canonical interactions between inserted positions. The distribution of PPV, STY is shown in Fig. 4 (F1 in supp. mat.) highlighting how the motifs as of now are still limited to predict this finer grain information. While looking dramatic, it can be explained in two different ways. First, their number is really low, as shown in Fig. 5. The Y-axis shows that at these positions rarely more than 4 non-canonical interactions are present per RNA, and rarely more than 2 found in the inserted motifs. Second, this is underestimating the non-canonical interactions since many can not be predicted by our model. On top of those that are at locations where no motifs are predicted, many can be linking motifs together, but those can not be found in the dataset used.



(a) PPV



(b) STY

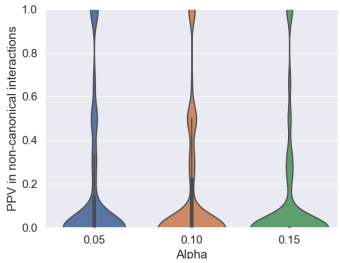


(c) F1

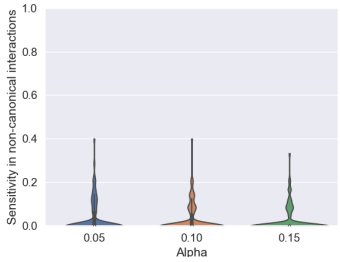
Figure 3: **Predicting canonical and Wobble interactions in motifs** For α values of $0.05, 0.1, 0.15$ that more than half of the canonical and Wobble base pairs in the motifs are correctly predicted, and 40% of them are generally captured.

3.7 Performance

Integer programming is known to be NP-complete, but decades of optimization have allowed to leverage efficient implementations to express complex models. We remind that a maximum of 10^4s was enforced and the 14 sequences without a solution at the time cutoff with $\alpha = 0.1$ were discarded. We show in Fig. 6 the execution time in seconds. More heuristics as the ones developed by IPknot for long sequences could be used, at the cost of a decrease in the accuracy. We expect that the optimal gains would be achieved by optimizing the location where we allow motifs to be inserted.



(a) PPV



(b) STY

Figure 4: **Predicting non-canonical base pairs in motifs**. True positives are the non-canonical base pairs at positions where one motif is inserted in the sequence. They are composing at most 15% of the interactions in the inserted motifs, and are hard to predict.

3.8 Tools Comparison

When benchmarked only on the canonical and Wobble base pairs, while a value of $\alpha = 0.15$ seems to give marginally better results 1 in the secondary structures, it is a bit worse inside the motifs 3. We therefore selected $\alpha = 0.1$ as default parameter value.

We evaluated in Fig. 7 RNA-MoIP with $\alpha = 0.1$ against 6 other tools: RNAfold [5], our implementation of IPknot [14] (i.e. $\alpha = 0$), PKnots [9], HotKnots [10] Knotty [11], and SPOT-RNA [12].

While in the task of predicting base pairs of the secondary structure RNA-MoIP clearly outperforms the first 4 competitors.

When compared with Knotty it produces as good results while providing more information in the form of motifs inserted. While SPOT-RNA seems to outperform all competitors, we remind that it is un-retrainable, and most probably overfitted since it does not perform well on sequences not related to its training set [13].

3.9 Including alignments increases predictability

Our model was extended in Sec. 2.4 to incorporate evolutionary information as sequence alignments. The Rfam database [33] is a repository of curated structured RNA families and has recently started providing hand aligned PDB structures. As of January 2023 they are 40 Rfam families alignments with 201 PDB structures aligned. Out of those, they belong to 23 different families (as defined by RNA Structure Atlas). The representative of each family was selected, or a random for the two families with a structure in an alignment but not its representative.

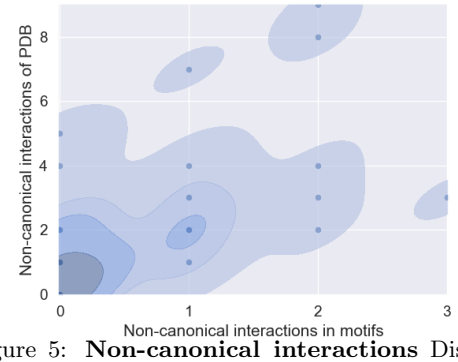


Figure 5: **Non-canonical interactions** Distribution of the number of non-canonical interactions that are observed at motifs inserted locations. On the y-axis the number in the real structure, on the x-axis how many are annotated in the inserted motif.

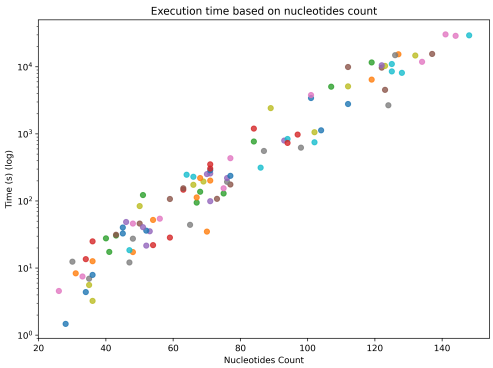


Figure 6: **Execution Time based on nucleotide count of the sequence at** $\alpha = 0.1$. A maximum of 10^4s was allowed, and 14 sequences didn't return a solution in that time.

The name of the 23 structures are provided in the directory with the code in the ISMB folder.

To the best of our knowledge, no other algorithm easily integrates multiple sequence alignments and pseudoknots. On those 23 structure we computed as baseline RNAalifold [5]—which doesn't take into account pseudoknots—and RNA-MoIP as shown in Fig. 8. We compare with RNAfold and RNA-MoIP without the alignments in Fig. 9.

A drastic increase in PPV is observed as much for RNAalifold vs RNAfold as for RNA-MoIP. We see that with a sequence alignment, there isn't a significant improvement brought by the motifs under our scheme. The F1 distributions for RNA-MoIP at $\alpha = 0$ where no motif is taken into account or $\alpha = 0.1$ are almost similar, increasing α slowly decreases the F1 value.

Incorporating motif prediction using our simple scheme does not decrease accuracy while providing additional information, as we will explore in the next section.

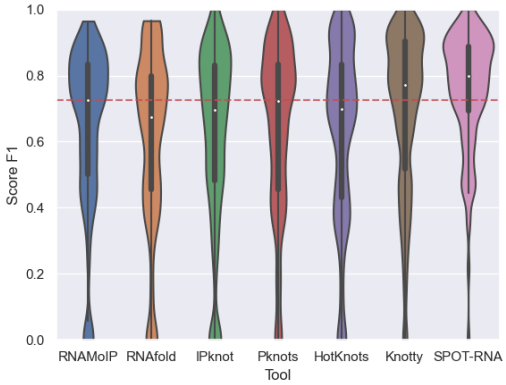


Figure 7: **Tools comparison of F1 scores** RNA-MoIP with parameter $\alpha = 0.1$ is compared with 6 other tools. Knotty does not provide additional geometric information. SPOT-RNA cannot be re-trained and probably learned on the predicted examples or close relatives.

3.10 Kink-turn

A particularly interesting geometrically complex local motif is the kink-turn [34, 35] which is linked to many different biological processes [36]. Since they greatly constrain the structure they are key pieces completely ignored in the secondary structure representation.

The RNA Structure Atlas [18] maintains different classes of kink-turn as all their occurrences in known RNA structures. From the previous set of 23 structures with alignments, 4 have a kink-turn annotated in them all in group `IL_29549.4`, in PDB and chains 3V7E-C, 4AOB-A, 4KQY-A and 5FJC-A. Since all four belong to the same group, they are represented with the same non-canonical diagram.

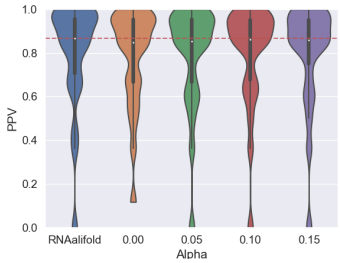
We investigate where motifs are predicted at the position of the kink turn with the alignment in Fig. 10 and without in Fig.11. In both cases, different interior loop are colored as green or blue, and hairpins as orange. When exactly the right base pair was predicted, it is also colored, the other ones inside the predicted motifs are omitted.

We observe immediately that there is always an interior loop matching on each strand of the kink-turn, in green. The prediction without alignment seem to be more cohesive, having always a match on all the position forming a cycle with the triangular pattern. But we can see using the alignment that the match on 5FJC chain A recover 3 out of the 4 non Watson–Crick/Watson–Crick base pairs.

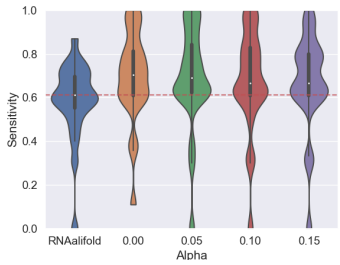
Since the non-canonical interaction do not necessitate an exact match to reflect similar geometries [37] a more in depth study out of the scope of the present work is needed. Nonetheless, RNA-MoIP predicts relevant geometric information that can not be represented by the secondary structure at the kink-turn location.

4 Conclusion

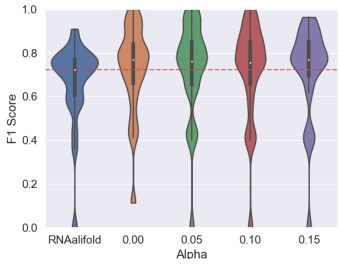
In this work, an integer programming framework allowing simultaneous prediction of the secondary structure with pseudoknots, and insertion of structural motifs, is presented. The implementation in RNA-MoIP is benchmarked over the 101 non-redundant pseudoknotted RNAs with known structure. We show that combining the approach of IPknot to construct pseudoknotted structures based on the base pair probability matrix obtained by the standard thermodynamic model, implemented in the ViennaRNA, with the insertion of known conserved



(a) PPV



(b) STY

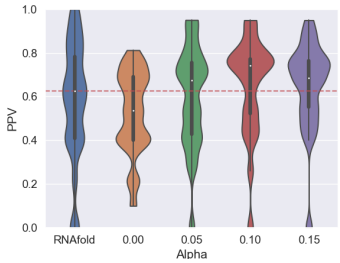


(c) F1

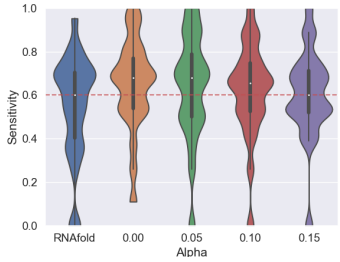
Figure 8: **Predicting secondary structure based on alignment**. Result of the alignment's dataset with the help of the alignment informations.

structural motifs, allows to: (1) increase the accuracy of the prediction of secondary structure with pseudoknots, (2) generates accurate knowledge about canonical and Wobble interactions present inside structural motifs, which might not belong to the secondary structure, (3) improves drastically under a simple scheme to incorporate evolutionary information from multiple sequence alignment as validate with 23 independent structures aligned by Rfam, and (4) predict non-canonical interaction motifs at kink-turn locations.

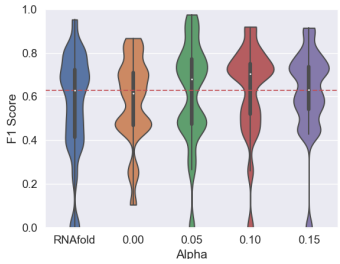
Two main limitations are highlighted by our work. First, motifs can be inserted now based on a perfect sequence match. More advanced probabilistic techniques, as RMDetect [38], JAR3D [39] or BayesPairing [40], would allow to integrate a more rigorous term in the objective function, as match motifs with altered sequence, increasing diversity and therefore the range of predictable structure. Second, the database of motifs only incorporates loops (i.e. hairpin, interior loops, multi-loops).



(a) PPV



(b) STY

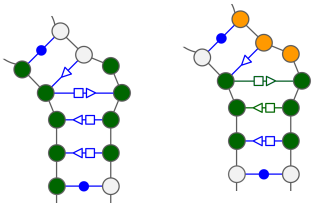


(c) F1

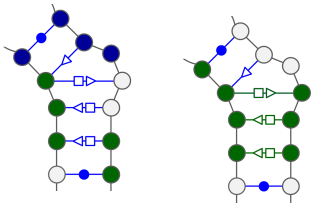
Figure 9: **Predicting secondary structure without alignment**. Result of the alignment's dataset without using the alignment informations.

These approaches can directly use multiple sequence alignment which would not only increase the general base pairs as shown in this work, but would probably show a more significant boost from incorporating the motifs.

Advances in structural molecular biology are pushing against the limitation of the nearest neighbour model. While the biological importance of networks of non-canonical interactions are becoming more and more evident, the capacity to predict them lags far behind. The IP programs remains a promising direction for RNA structure determination due to the flexibility of their formulation allowing to go above the nearest neighbour model. Expending to more complex conserved structures, as groups of interacting and conserved loops containing pseudoknots described in Carnaval [19, 22] would allow to take fully advantage of the IP formulation and extend the notion of pseudoknots prediction to all non-canonical interactions. This flexible formulation will also allow to give specific rules to help incorporate chemical modifications and other features that are absent from the nearest neighbour model.

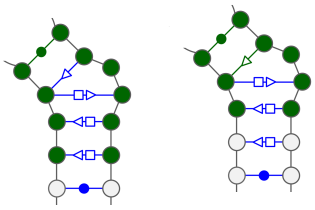


(a) 3V7E chain C (b) 4AOB chain A

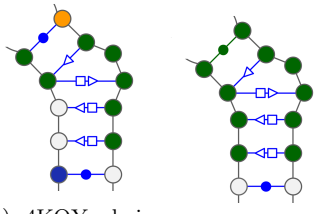


(c) 4KQY chain A (d) 5FJC, A

Figure 10: **Motifs inserted on the kink-turn with alignment** For the 4 PDB with a kink-turn the representation of it's group and colored the position where RNA-MoIP predicted motifs on the kink-turn as the exact base pair.



(a) 3V7E chain C (b) 4AOB chain A



(c) 4KQY chain A (d) 5FJC, A

Figure 11: **Motifs inserted on the kink-turn without alignment** For the 4 PDB with a kink-turn the representation of it's group and colored the position where RNA-MoIP predicted motifs on the kink-turn as the exact base pair.

Acknowledgements

The authors are grateful for discussions about including alignments with Ivo Hofacker and Sebastian Will, and Nancy Ontiveros for aligning PDBs with RFAM families.

Funding

VR is supported by NSERC and FRQS grants.

References

[1] Feng Wang, Travis Zuroske, and Jonathan K Watts. RNA therapeutics on the rise. *Nat Rev Drug Discov*, 19(7):441–442, 2020.

[2] Ai-Ming Yu, Young Hee Choi, and Mei-Juan Tu. RNA drugs and RNA targets for small molecules: principles, progress, and challenges. *Pharmacological reviews*, 72(4):862–898, 2020.

[3] Zhichao Miao, Ryszard W Adamiak, Maciej Antczak, Michał J Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Yi Cheng, Fang-Chieh Chou, Rhiju Das, et al. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8):982–995, 2020.

[4] Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[5] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for molecular biology*, 6(1):1–14, 2011.

[6] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):1–9, 2010.

[7] Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, 28(8):963–976, 2020.

[8] Rune B Lyngsø and Christian NS Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427, 2000.

[9] Elena Rivas and Sean R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068, 1999.

[10] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *Rna*, 11(10):1494–1504, 2005.

[11] Hosna Jabbari, Ian Wark, Carlo Montemagno, and Sebastian Will. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*, 34(22):3849–3856, 2018.

[12] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 10(1):1–13, 2019.

[13] Marcell Szikszai, Michael J Wise, Amitava Datta, Max Ward, and David Mathews. Deep learning models for RNA secondary structure prediction (probably) do not generalise across families. *bioRxiv*, 2022.

[14] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–93, Jul 2011.

[15] Kengo Sato and Yuki Kato. Prediction of RNA secondary structure including pseudoknots for long sequences. *Briefings in Bioinformatics*, 23(1):bbab395, 2022.

[16] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4):499–512, 2001.

[17] Neocles B Leontis, Aurelie Lescoute, and Eric Westhof. The building blocks and motifs of RNA architecture. *Current opinion in structural biology*, 16(3):279–287, 2006.

[18] Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *Rna*, 19(10):1327–1340, 2013.

[19] Vladimir Reinharz, Antoine Soulé, Eric Westhof, Jérôme Waldispühl, and Alain Denise. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. *Nucleic acids research*, 46(8):3841–3851, 2018.

[20] Vladimir Reinharz, François Major, and Jérôme Waldispühl. Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics*, 28(12):i207–i214, 2012.

[21] Jason Yao, Vladimir Reinharz, François Major, and Jérôme Waldispühl. RNA-MoIP: prediction of RNA secondary structure and local 3D motifs from sequence data. *Nucleic acids research*, 45(W1):W440–W444, 2017.

[22] Antoine Soulé, Vladimir Reinharz, Roman Sarrazin-Gendron, Alain Denise, and Jérôme Waldispühl. Finding recurrent RNA structural networks with fast maximal common subgraphs of edge-colored graphs. *PLoS computational biology*, 17(5):e1008990, 2021.

[23] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, 38(suppl_1):D280–D282, 2010.

[24] Stephan H. Bernhart, Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinf.*, 9(1):1–13, December 2008.

[25] OR-Tools. OR-Tools | Google Developers, August 2021. [Online; accessed 18. Oct. 2022].

[26] coin or. Cbc, April 2022. [Online; accessed 26. Apr. 2022].

[27] Federal University of Ouro Preto Departament of Computing, ICEB. Python MIP Documentation — Python-MIP documentation, Jan 2021. [Online; accessed 15. Apr. 2022].

[28] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, v9.5.0, 2022.

[29] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–42, Jan 2000.

[30] BGSU RNA group. Representative Sets of RNA 3D Structures , Apr 2022. [Online; accessed 15. Apr. 2022].

[31] Sandra Smit, Kristian Rother, Jaap Heringa, and Rob Knight. From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, 14(3):410–416, 2008.

[32] Tomasz Zok, Maciej Antczak, Michal Zurkowski, Mariusz Popenda, Jacek Blazewicz, Ryszard W Adamiak, and Marta Szachniuk. RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Research*, 46(W1):W30–W35, 04 2018.

[33] Ioanna Kalvari, Eric P. Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R. Eddy, Robert D. Finn, Alex Bateman, and Anton I. Petrov. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, 49(D1):D192–D200, January 2021.

[34] Daniel J Klein, T Martin Schmeing, Peter B Moore, and Thomas A Steitz. The kink-turn: a new RNA secondary structure motif. *The EMBO journal*, 20(15):4214–4221, 2001.

[35] Shigeyoshi Matsumura, Yoshiya Ikawa, and Tan Inoue. Biochemical characterization of the kink-turn RNA motif. *Nucleic acids research*, 31(19):5544–5551, 2003.

[36] Lin Huang and David MJ Lilley. The kink-turn in the structural biology of RNA. *Quarterly reviews of Biophysics*, 51, 2018.

[37] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of RNA base pairs. *Nucleic acids research*, 37(7):2294–2312, 2009.

[38] José Almeida Cruz and Eric Westhof. Sequence-based identification of 3d structural modules in RNA with RMDetect. *Nature methods*, 8(6):513–519, 2011.

[39] James Roll, Craig L Zirbel, Blake Sweeney, Anton I Petrov, and Neocles Leontis. JAR3D Webserver: Scoring and aligning RNA loop sequences to known 3D motifs. *Nucleic acids research*, 44(W1):W320–W327, 2016.

[40] Roman Sarrazin-Gendron, Hua-Ting Yao, Vladimir Reinharz, Carlos G Oliver, Yann Ponty, and Jérôme Waldispühl. Stochastic sampling of structural contexts improves the scalability and accuracy of RNA 3D module identification. In *International Conference on Research in Computational Molecular Biology*, pages 186–201. Springer, 2020.