

# An integer programming framework for simultaneous prediction of RNA structure with pseudoknots and insertion of local 3D motifs

Gabriel Loyer  
Vladimir Reinharz

Université du Québec à Montréal

August 26, 2022













# Problematic

- ▶ The prediction of RNA structure canonical base pairs from a single sequence, **especially pseudoknotted ones**, remains challenging.
- ▶ Structural motifs in the loops are essential for the final shape of the molecule, **enabling its multiple functions**.

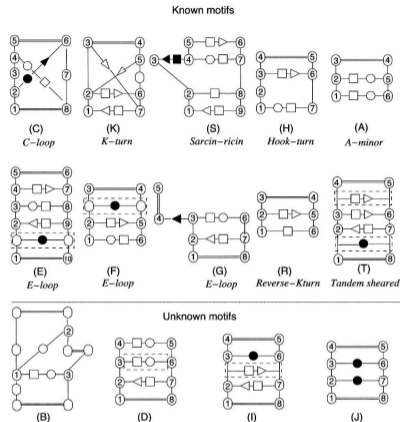
Proposition:

- ▶ Use motifs representations to match known functional structure into the sequence.
- ▶ Combine with a decomposition approach to maximise base pairings probabilities, including pseudoknots.

# RNA Motifs

Non-Canonical Interactions			
	cWW		cHH
	tWW		tHH
	cWH		cHS
	tWH		tHS
	cWS		cSS
	tWS		tSS

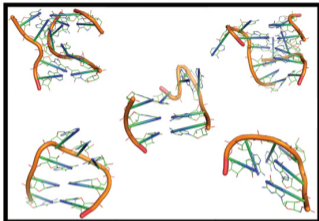
(a) Type of non canonical interactions in motifs.



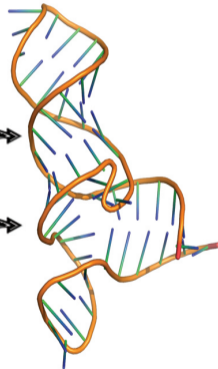
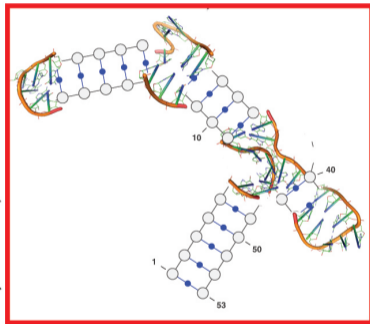
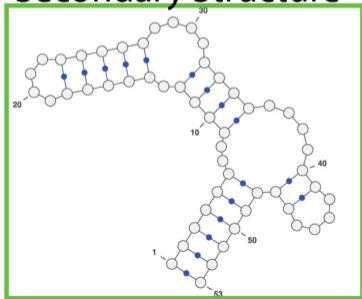
(b) Examples of motifs present in the database compiled. Djelloul, M. and Denise, A. (2008) Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14, 2489–2497

# RNA Motifs over Integer Programming (RNA-MoIP)

## Motifs Database



## Secondary Structure



# Using IP

## DISCLAIMER:

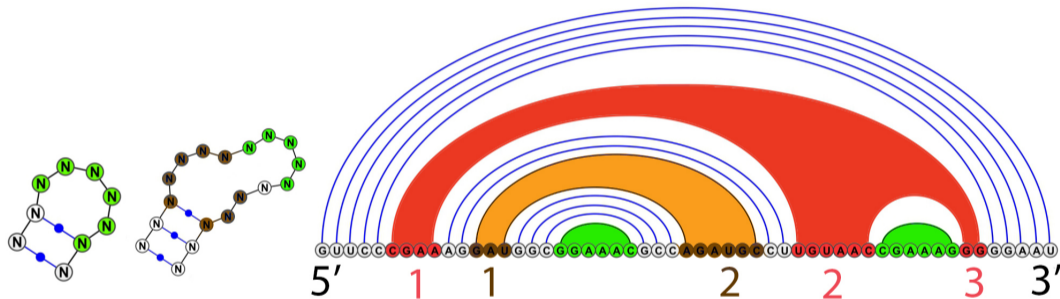
*Integer programming is more of an art than a science (David Avis)*

- ▶ System of Linear Inequations
- ▶ Variables must be integer
- ▶ A linear function to minimize or maximize (the objective function)

## Data and Variables

- ▶ Motif  $M^x := [(x_1^1, \dots, x_{k_1}^1), \dots, (x_1^j, \dots, x_{k_j}^j)]$   
 $|M^x| = \text{length of motif } x$
- ▶  $D_{u,v}$  indicates if the base pair  $(u, v)$  is removed in the secondary structure
- ▶  $C_{k,l}^{x,j}$  indicates the insertion of the  $j$ -th component of the motif  $x$  between positions  $k$  and  $l$

# Constraints



- ▶ Components stacked to a base pair
- ▶ No lonely base pairs
- ▶ No crossing arc
- ▶ Components ordered

# Objective Function

Penalty for removing a base pair

$$10 * \sum_{(u,v) \in B} D_{u,v} - \sum_{x \in Mot^j} \left( (|M^x|)^2 \cdot \underbrace{\sum_{(x,k,l) \in Seq_1^j} C_{k,l}^{x,1}}_{\text{Motif Insertion}} \right)$$

- ▶ Penalty of 10 for every based pair removed
- ▶ Bonus of the  $|length|^2$  for every motif inserted



# Limitations

There were still some limitations to this method:

1. Pseudoknotted interactions were not supported.
2. An initial secondary structure was to be provided to help guide insertions.

# Structure Decomposition

1. **Decomposition based on IPknot:** Each pseudoknotted structure can be decomposed into multiple structures that are pseudoknot free.

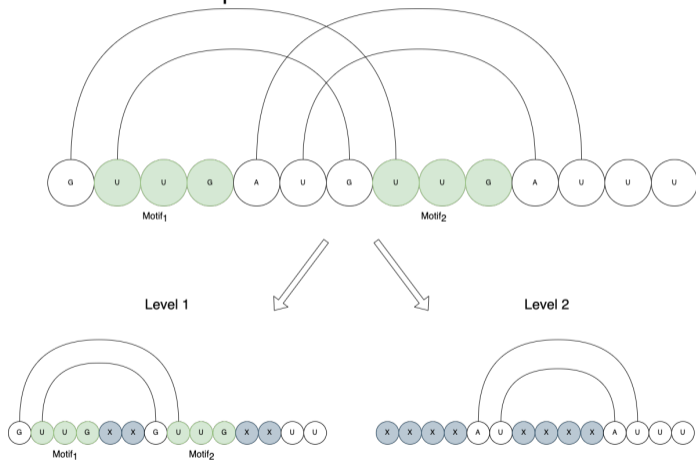


Figure: Example of decomposition for a pseudoknotted structure into 2 pseudoknot-free structures.

## Structure Decomposition

- Calculate probabilities matrices.** For each sub-structures, we can use a thermodynamic approach (in our case, RNAfold) to calculate the probabilities matrix for all possible pairings.

$i \setminus j$	0	1	2	...	n
0	-	0.05	0.01	...	0.8
1	0.05	-	0.02	...	0.1
2	0.01	0.02	-	...	0.03
...	...	...	...	...	...
n	0.8	0.1	0.03	...	-

Table: Example of probability matrix.

# Structure Decomposition

3. **Combine the matrices.** Sum all the matrices together. Only pairings with probability above  $\theta$  are considered (fixed at  $10^{-3}$  in our experiment). The resulting matrix is then multiply by a factor of 10.
4. **Insert Motifs.** Look for possible motif insertions. Note that the motifs can only be inserted in the first level of substructure.

# Using IP

- ▶  $D_{u,v}$  become  $D_{u,v}^p$ , indicating if the base pair  $(u, v)$  at the substructure level  $p$  is removed in the secondary structure.
- ▶ Constraints to respect structure decomposition.
- ▶ We want to maximise **the sum of probabilities of the pairings inserted.**

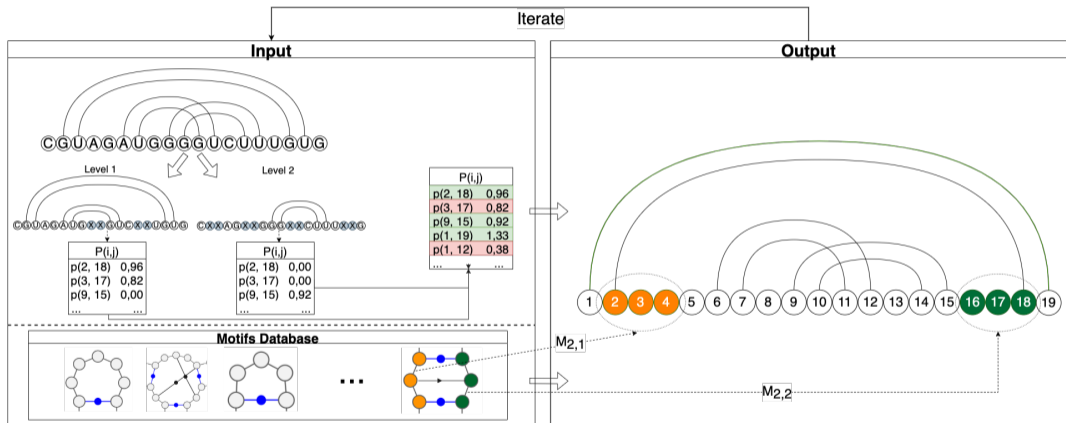
## IP Objective

The objective function is the **motif insertion objective** with the addition of **maximisation of the sum of probabilities of the base pairings inserted**.

$$\max \quad \alpha \sum_{x \in \text{Mot}^j} \left( (|M^x|)^2 \cdot \sum_{(x,k,l) \in \text{Seq}_1^j} C_{k,l}^{x,1} \right) \quad (1)$$

$$+(1 - \alpha)10 \sum_{(u,v) \in \mathcal{B}} \sum_{q=1}^m (1 - D_{u,v}^q) p(u,v) \beta^q. \quad (2)$$

# New Workflow



# Benchmark

- ▶ Benchmark over the all non-redundant RNAs below 150 nucleotides. Representant class 3 Å, with one structure per non-redundant class as defined by the BGSU RNA Structure Atlas v3.208.
- ▶ On each test, we remove motifs from the pdb of the sequence tested.
- ▶ To evaluate, we use different methods comparing the pairings returned.
  - ▶ **True Positive (TP)**: Pairing present in the predicted and in the known structure.
  - ▶ **False Positive (FP)**: Pairing present in the predicted and not the known structure (overpredict).
  - ▶ **False Negative (FN)**: Pairing missing in the predicted and in the known structure (underpredict).



# Benchmark

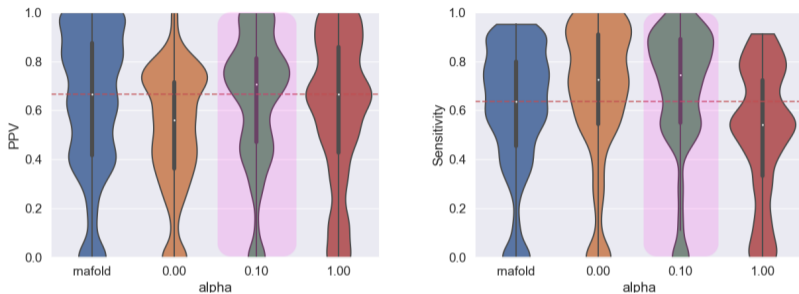
Using those metrics, we were able to evaluate the results with pseudoknot count, maximum pseudoknot level and the following formulas:

$$PPV = \frac{TP}{TP+FP}$$

$$STY = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \cdot PPV \cdot STY}{STY + PPV}$$

## Results - Pairings



(a) PPV - How many that we predicted are of the known structure

(b) STY - How many of the known structure do we predict

**Figure: Predicting secondary structure with pseudoknots** Comparing results for RNAfold (can not predict crossing interactions), with only bases pairings probabilities ( $\alpha = 0$ ), with only motifs insertions ( $\alpha = 1$ ), and for an hybrid approach ( $\alpha = 0.1$ ). When  $\alpha > 0$ , all base pairs inserted in the same motifs are counted as true positives.

## Results - Pairings

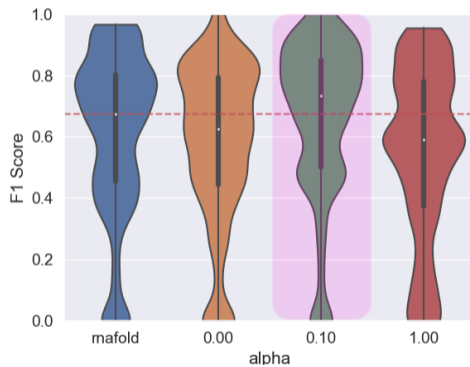


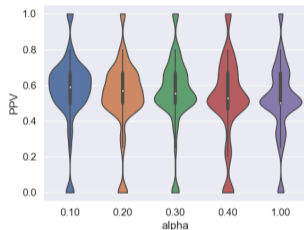
Figure: F1 Score - Correlation between PPV and STY

## Results - Pseudoknot

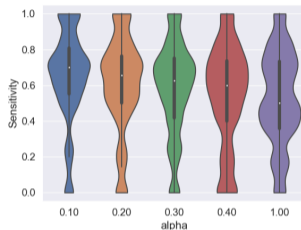
$\alpha$	0	0.1	0.2	0.3	0.4	1
PK lvl too low	16	45	50	57	51	64
PK lvl correct	85	42	22	15	18	1

**Table: Predicting pseudoknot lvl** As  $\alpha$  is increased we underestimate the lvl of pseudoknots in the structure. The complexity of the IP model increases and its more challenging to find an optimal solution in time (in  $10^4$  seconds).

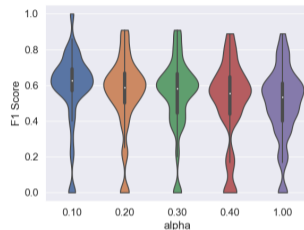
## Result - Canonical Pairs In Motifs



(a) PPV



(b) STY



(c) F1

**Figure: Predicting canonical and Wobble interactions in motifs** For  $\alpha$  values of 0.1, 0.2, 0.3, 0.4, and 1 that more than half of the canonical and Wobble base pairs in the motifs are correctly predicted, and over 65% of them are generally captured.

## Result - Non Canonical Pairs In Motifs

However, non-canonical interactions are still hard to predict:

$\alpha$	0.1	0.2	0.3	0.4	1
PPV	0.14	0.134	0.132	0.163	<b>0.174</b>
STY	0.17	0.185	0.154	0.173	<b>0.187</b>
F1	0.143	0.147	0.131	0.152	<b>0.168</b>

**Table:** Percentage of correctly well predicted non-canonical interactions.

# Overall

The results shows that RNA-MoIP:

1. **Improves** the ratio of well predicted interactions in the secondary structure
2. Predicts well **canonical and Wobble pairs** at the location where motifs are inserted.
3. Non-canonical interactions remains challenging to predict.

# New Database

- ▶ Recurent subgraph in all pdbs
- ▶ Preliminar test show similar results, despite smaller database size (1287 unique sequences vs 4596 motifs)
- ▶ Still need to validate canonical and non-canonical interactions in motif.
- ▶ More at Vladimir Reinharz talk on Thursday !

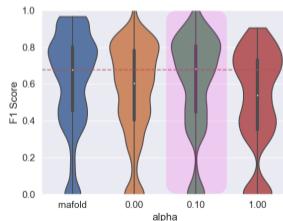
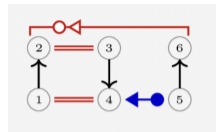


Figure: F1 Score - Correlation between PPV and STY



# Alignments

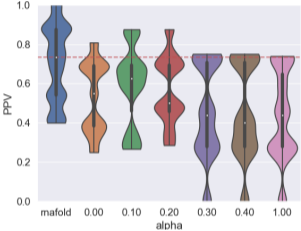
But wait... Was it a session about secondary structure prediction from alignment ?

# Alignments

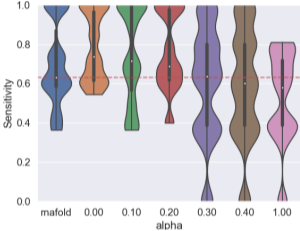
- ▶ Using RFAM Alignments (Thanks to Nancy!)
- ▶ Use ViennaRNA Package folding (RNAalifold)
- ▶ Predicting for the sequence with the help of the alignments
- ▶ Only allow motif to be inserted if it exactly fit in at least 50% percent of the sequences of the alignment.
  - ▶ Only compare with subsequences in alignments not too distance for the target subsequence.
- ▶ Benchmark on 11 chains (5 pseudoknots)



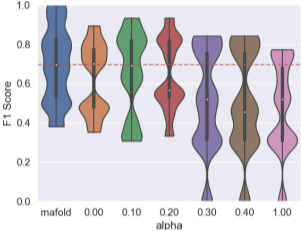
# Without Alignments - Results



(a) PPV



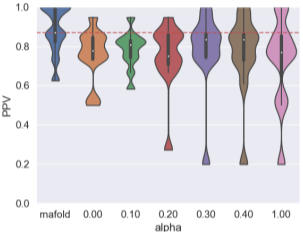
(b) STY



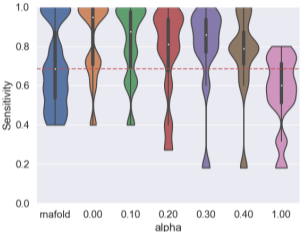
(c) F1

Figure: Result of chains without alignments

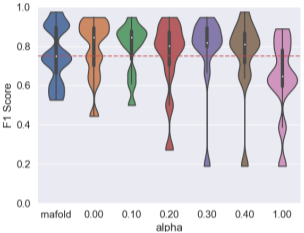
# With Alignments - Results



(a) PPV



(b) STY



(c) F1

Figure: Result of chains with alignments

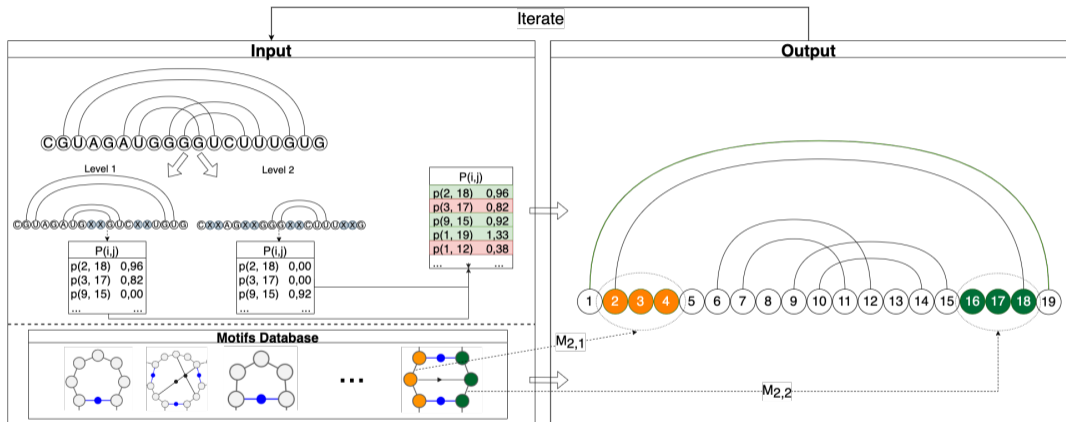
# Futur Challenges

1. Better scheme for potential location of motifs (e.g. BayesPairing) to improve flexibility in insertion points.
2. Motifs combining many loops through long range interactions (canonical or not)

# Acknowledgments

- ▶ **Supervisor:**  
Vladimir Reinharz
- ▶ Roman Sarrazin-Gendron

# RNA-MoIP



► Available: [gitlab.info.uqam.ca/cbe/RNAMoIP](https://gitlab.info.uqam.ca/cbe/RNAMoIP)