

INTERMEDIATE WORK REPORT:

Louis Becquey, January 23, 2019

*A biobjective IP framework that uses detected possible RNA modules
to predict RNA secondary structures*



under the supervision of:

Fariza Tahi, HDR

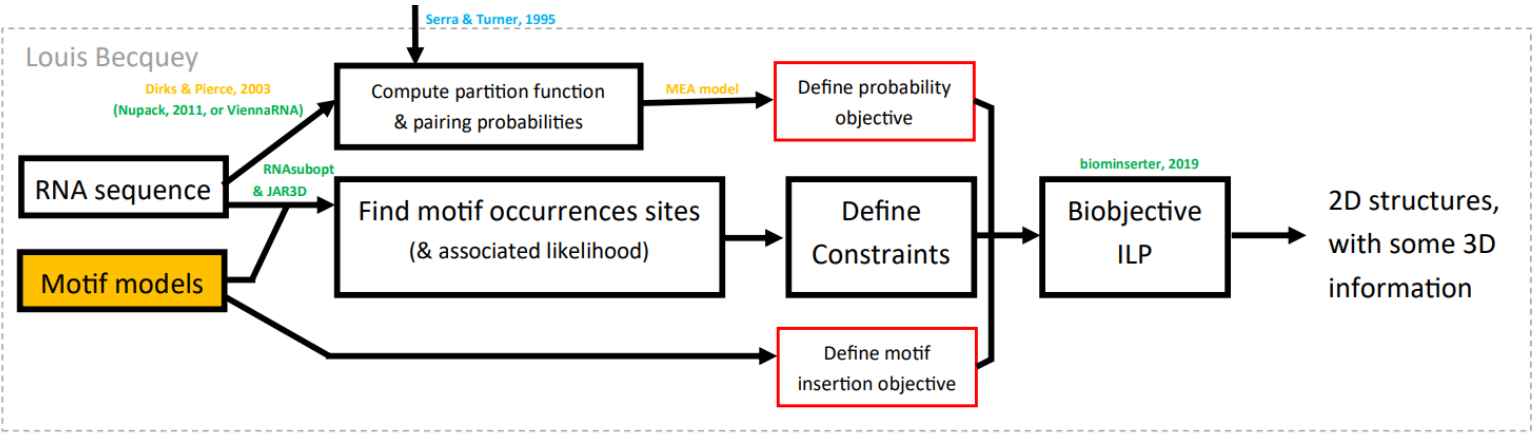
Eric Angel, HDR

AROBAS, IBISC, Paris-Saclay University



université
PARIS-SACLAY

Graphical Abstract



Contents

| | | |
|-----|--|---|
| 1 | Motivation | 2 |
| 2 | Computation of basepair probabilities | 2 |
| 3 | Detection of potential modules in the sequence input | 2 |
| 4 | Formulation of the IP problem | 3 |
| 4.1 | Variables | 3 |
| 4.2 | Objectives | 4 |
| 4.3 | 10 Constraints to bind them all | 4 |
| 5 | Remaning questions | 5 |

1 Motivation

We now have quite consistent RNA module databases (RNA3Dmodule [1], the RNA Motif Atlas [2], or CaRNAval [3]). We call a RNA module the combined description of these points:

- A particular base-pairing pattern of canonical & wobble pairing (bi-dimensional information),
- A particular organisation of non-canonical contacts in space (tri-dimensional information).
- A sequence or consensus sequence that we know to adopt that particular base-pairing organisation (sequence information), or at least a probabilistic model to predict if a given sequence will fold according to the module.

Then, a RNA module is direct data about how a sequence can fold in space. Here, we are interested in using them to predict the bi or tri dimensional structure of an RNA sequence from scratch. Then, the algorithm we plan to use is the following:

- Find all possible occurrences of known RNA modules in the query sequence by using the sequence probabilistic models of the modules against the query bases,
- Define constraints on the secondary structure imposed by motives if they would be included
- Find a secondary structure that satisfies as much as possible both the expected accuracy of the structure and a criterion of module inclusion, by solving a bi-objective IP problem, using the previous constraints.

2 Computation of basepair probabilities

To compute the probability of a structure to exist in the equilibrium thermodynamical ensemble, we need to estimate the basepair probabilities. This is usually done by a dynamic programming algorithm using an energy parameter model for base stackings in stems and energies of loops. As we are supposed to allow the existence of pseudoknots, we chose the Dirks & Pierce algorithm [4] and its C++ implementation called Nupack.

Discussion of that choice

The Dirks and Pierce algorithm has a complexity of $O(N^5)$ instead of the usual $O(N^3)$ algorithm, and its implementation in libNupack is even not parallel. We may want to use a faster one (like ViennaRNA's one), even if it does not officially supports pseudoknots. Actually, to form a simple class of pseudoknots, an RNA needs to have a high likelihood to form each of the two independant loops in a non-pseudoknotted form. Then, using a probability matrix computed with an algorithm that does not model pseudoknot may result in similar structures. Our bioobjective program would establish the pseudoknot after all as forming more stems stabilizes the structure and as we do not forbid the pseudoknot.

3 Detection of potential modules in the sequence input

Choice of a module model, choice of a sequence probabilistic model

Depending on the authors and databases, different approaches have been published. A very complete model has been developed by the BGSU team [5] : For each recurrent module of their "3D Motif Atlas", they build a probabilistic model for sequence variability based on a hybrid Stochastic Context-Free Grammar/Markov Random Field (SCFG/MRF) method. Concerning the other databases, a probabilistic model using a bayesian Network can be designed for an RNA module [6] to estimate the probability that the sequence would fold into the module.

As BGSU's model is already implemented in a software called JAR3D that perfectly suits our need, we decided to use BGSU's model of modules and their database. JAR3D finds, given the sequence of an RNA loop with no additional information, the likelihood that it folds in every known 3D module of the RNA Motif Atlas [2].

For each module group of the database, JAR3D estimates the probability that a sequence folds into the 3D pattern, given instances of the module in real RNA 3D structures. The instances might differ in sequence, length of bulged insertions, and other details, but share a common graph of 3D non-canonical interactions. It takes into account crossed contacts, base-ribose and base-phosphate interactions, and triple base-pairs. This is the most complete approach to use to my knowledge.

Detection of the loops in an RNA sequence

As JAR3D scores the sequence of RNA loops, we first need to predict the 2D loops positions in the RNA sequence.

For now, we have been using RNAsubopt from the ViennaRNA package [7] to compute the 10 most probable 2D structures of the RNA and extract the loops that occur in those structures. This computation should be avoided in the future : It requires the computation of the RNA's partition function and basepair probabilities again, which is something we already computed with Nupack. Unfortunately, to avoid computing it twice, we should reimplement ourselves the algorithms instead of using the actual implementations. We have programs that output the basepair matrix but no structures, and programs that output structures but not the basepair matrix, and we don't have a program that can do both.

4 Formulation of the IP problem

This formulation is an improved merge between the IPknot [8] formulation to fold RNAs with pseudoknots (but without "levels") and RNAMoIP's one [9] to include modules. The bi-objective algorithm is inspired from BiokoP [10].

4.1 Variables

Let n be the number of nucleotides in the query RNA sequence s .

Let M be the set of modules that could be inserted in s (a loop from s has a good JAR3D score against one of the modules).

Let x be a module of M , $\|x\|$ be the number of distinct components of x , and $p(x)$ the associated score of insertion given by JAR3D for that motif inserted at a particular position.

Let $P_{x,i}$ be the position in s where we can insert the i th component of module x .

As the same module model can be inserted several times in s , several different x modules in M may refer to the same theoretical module, but inserted at different positions.

Let $k_{x,i}$ be the size in nucleotides of that i th component of x .

Let y_v^u be the **decision boolean variable** indicating that $s[u]$ and $s[v]$ form a canonical base pairing. According to the standard loop model, we always have $v > u + 3$.

Let C_i^x be the **decision boolean variable** indicating that we do insert the i th component of module x at position $P_{x,i}$.

Note that a base pair y_v^u is possible if and only if $v > u + 3$, and that we do not need to use two variables y_v^u and y_{vu} for the same pair. Then, we have $\sum_{i=4}^n (n - i)$ decision variables ($\approx \frac{1}{2}n^2$ decision variables) of the form y_v^u . Regarding the C_i^x , if we have an average insertion of ν motives by RNA sequence, the motives having in average μ components, components that can be inserted in average at π different positions in s , then we need to add, in average, $\nu \times \mu \times \pi$ decision variables C_i^x .

Then, we expect having around $\frac{1}{2}n^2 + \nu\mu\pi$ decision variables.

4.2 Objectives

We have two objectives : Find a structure with correct expected accuracy, and find a structure which includes (large) known modules. Let X be the vector of all our decision variables, we define the following objective functions to maximize:

$$\begin{aligned}
 f_{1A}(X) &= \sum_{x \in M} \left[(\|x\|)^2 \times \sum_{p \in P_{x,1}} C_p^{x,1} \right] \\
 f_{1B}(X) &= \sum_{x \in M} \left[p(x) \times \sum_{p \in P_{x,1}} C_p^{x,1} \right] \\
 f_{1C}(X) &= \sum_{x \in M} \left[\frac{\|x\|}{\log_2(\sum_{i=1}^{\|x\|} k_{x,i})} \times p(x) \times \sum_{p \in P_{x,1}} C_p^{x,1} \right] \\
 f_2(X) &= \sum_{u < v} p_{uv} \times y_v^u \times I[p_{uv} > \theta], \quad p_{uv} = \sum_{\sigma \in S(s)} y_v^u \cdot p(\sigma|s)
 \end{aligned}$$

The different f_1 objectives are supposed to maximize the number of inserted motives in s , weighted by their number of components (squared for 1A), or the JAR3D score (B and C). In 1C, a penalty is added on the number of nucleotides involved in the looped zone (sum of $k_{x,i}$) to avoid long unpaired zones. We don't know yet which one will give the better results. f_2 is supposed to maximise the expected accuracy of the secondary structure. p_{uv} are the base pairing probabilities.

Note that f_{1A} is taken from RNA MoIP [9], to compare performance.

4.3 10 Constraints to bind them all

Constraint to ensure there only is 0 or 1 canonical pairing by nucleotide

$$\sum_{v < u} y_u^v + \sum_{v > u} y_v^u \leq 1 \quad \forall u \in \llbracket 1, n \rrbracket \quad (1)$$

Constraints to forbid lonely base pairs

$$\sum_{v=u}^n y_v^{u-1} - \sum_{v=u+1}^n y_v^u + \sum_{v=u+2}^n y_v^{u+1} \geq 0 \quad \forall u \in \llbracket 1, n \rrbracket \quad (2)$$

$$\sum_{u=1}^{v-2} y_{v-1}^u - \sum_{u=1}^{v-1} y_v^u + \sum_{u=1}^v y_{v+1}^u \geq 0 \quad \forall v \in \llbracket 1, n \rrbracket \quad (3)$$

These conditions ensure that if a base pair exists with $s[i]$, one of the adjacent bases is paired too. Equation 2 is useful if $s[u]$ is paired with $s[v > u]$ (a nucleotide later in the sequence), and equation 3 if $s[v]$ is paired with $s[u < v]$ (a nucleotide earlier in the sequence).

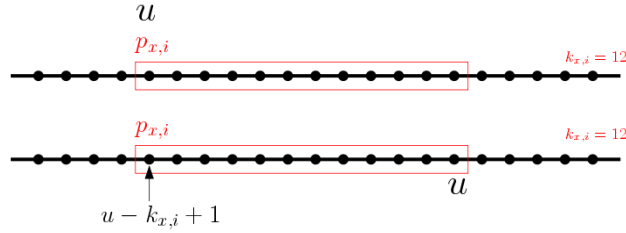
Constraint to forbid pairings inside a module component

$$k_{x,i} C_i^x + \sum_{u=P_{x,i}+1}^{P_{x,i}+k_{x,i}-2} \left[\sum_{v>u} y_v^u + \sum_{v<u} y_u^v \right] \leq k_{x,i} \quad \forall x \in M, i \in \llbracket 1, \|x\| \rrbracket \quad (4)$$

Constraint to forbid component to overlap

$$\sum_{x \in M} \sum_{i=1}^{\|x\|} \sum_{u=P_{x,i}}^{P_{x,i}+k_{x,i}-1} C_i^x \leq 1 \quad \forall u \in \llbracket 1, n \rrbracket \quad (5)$$

Then, whatever the nucleotide u , it can be part of a module component only once. u belongs to a component x, i if and only if $u - k_{x,i} + 1 \leq p_{x,i} \leq u$.



Constraints to respect the structure of large motives ($\forall x \in \{x \in M \mid \|x\| \geq 2\}$)

The first two constraints ensure that a component is inserted if and only if the next and the previous one are also inserted somewhere. They also force a minimal distance of 3 nucleotides between any two consecutive components of the same module.

$$C_i^x \leq \sum_{\substack{p' \in P_{x,i+1} \\ p' > p + k_{x,i} + 2}} C_{p'}^{x,i+1} \quad \forall p \in P_{x,i} \quad \forall x \in \{x \in M \mid \|x\| \geq 2\}, i \in \llbracket 1, \|x\| \rrbracket \quad (6)$$

$$C_i^x \leq \sum_{\substack{p' \in P_{x,i-1} \\ p' < p - k_{x,i-1} - 2}} C_{p'}^{x,i-1} \quad \forall p \in P_{x,i} \quad \forall x \in \{x \in M \mid \|x\| \geq 2\}, i \in \llbracket 1, \|x\| \rrbracket \quad (7)$$

We add another one to ensure that all components of any module are inserted the same number of times in s :

$$\sum_{p \in P_{x,1}} C_p^{x,1} - \frac{1}{\|x\|} \sum_{i=1}^{\|x\|} \sum_{p \in P_{x,i}} C_i^x = 0 \quad x \in M \quad (8)$$

And finally, we force base pairs between the end of a component and the beginning of the next one:

$$C_i^x \leq \sum_{p' \in P_{x,i-1}} y_p^{p' + k_{x,i} - 1} \quad \forall x \in \{x \in M \mid \|x\| \geq 2\}, i \in \llbracket 1, \|x\| \rrbracket \quad (9)$$

$$C_i^x \leq \sum_{p' \in P_{x,i+1}} y_{p'}^{p + k_{x,i} - 1} \quad \forall x \in \{x \in M \mid \|x\| \geq 2\}, i \in \llbracket 1, \|x\| \rrbracket \quad (10)$$

5 Remaning questions

- I am not sure how to compute the base-pair probabilities for the MEA model (objective f_2). If you already have a set of secondary structures $S(s)$, why do you need this program ?
- In objective f_1 , weighting the module by the number of components (squared) shows no reason to be the best idea. Some exploration is needed. An additional weight given by the pattern-matching step might be an idea, reflecting the probability to see the module here in s .

References

- [1] Mahassine Djelloul and Alain Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14(12):2489–2497, January 2008.
- [2] 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, January 2013.
- [3] 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.

- [4] Robert M. Dirks and Niles A. Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of Computational Chemistry*, 25(10):1295–1304, 2004.
- [5] Craig L. Zirbel, James Roll, Blake A. Sweeney, Anton I. Petrov, Meg Pirrung, and Neocles B. Leontis. Identifying novel sequence variants of RNA 3d motifs. *Nucleic Acids Research*, 43(15):7504–7520, September 2015.
- [6] José Almeida Cruz and Eric Westhof. Sequence-based identification of 3d structural modules in rna with rmdetect. *Nature methods*, 8(6):513, 2011.
- [7] 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*, 6:26, November 2011.
- [8] 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–i93, July 2011.
- [9] 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, June 2012.
- [10] Audrey Legendre, Eric Angel, and Fariza Tahi. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. *BMC Bioinformatics*, 19:13, January 2018.