INTERMEDIATE WORK REPORT:

Louis Becquey, October 22, 2018

An IP framework that uses detected possible RNA motifs to predict RNA secondary structures

under the supervision of:
Fariza Tahi, HDR
Eric Angel, HDR
AROBAS, IBISC, Paris-Saclay University

ibisc   université PARIS-SACLAY

Contents

Motivation

We now have quite consistent RNA motif databases (RNA3Dmotif [1], the RNA Motif Atlas [2]).
We call a RNA motif the combined description of these points:

- A sequence or consensus sequence that we know to adopt a particular base-pairing organisation (sequence information),
- A particular base-pairing pattern in this sequence (or consensus sequence) considering canonical & wobble pairing (bi-dimensional information),
- A particular organisation of non-canonical contacts in space (tri-dimensional information).

Then, a RNA motif 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 occurences of known RNA motifs in the query sequence by pattern matching of the motifs consensus sequences against the query bases,
- Define constraints on the secondary structure imposed by motives if included
- Find a secondary structure that satisfies both the most the expected accuracy (MEA model) and criterion of motif inclusion, by solving a bi-objective IP problem, using the previous constraints.

Pattern matching to detect motif sites

A very complete model has been developped by the BGSU team [3], which perfectly suits this need: JAR3D finds, given a RNA sequence with no additional information, the sites that are likely to fold in known 3D motives of the RNA Motif Atlas [2].

The RNA 3D Motif Atlas is an updated database of substructures exracted from a non-redundant set of 3D RNA structures. They are grouped into "motifs", which are a given pattern of 3D non-canonical interactions. Each motif contains one or several instances observed in real RNA structures. The instances might differ in sequence, length of bulged insertions, and other details, but share a common graph of 3D non-canonical interactions.

For each motif group of the database, JAR3D estimates the probability that a sequence folds into the 3D pattern, given the known members of the motif group. It takes into account crossed contacts, base-ribose and base-phosphate interactions, and triple base-pairs.

This is the most scientifically correct approach to use to my knowledge. Unfortunately, the code is in MatLab and seems hard to use on a single computer (JAR3D is basically a webserver).

Formulation of the IP problem

This formulation is an improved merge between the IPknot [4] formulation (without "levels") and RNAMoIP's one [5].

Variables

Let n be the number of nucleotides in the query RNA sequence s.
Let M be the set of motifs that could be inserted in s (a match exist between s and the sequence of the motifs' components).
Let x be a motif of M, x having j^x distinct components.
Let $P_{x,i}$ be the set of positions in s where we can insert the ith component of motif x.
Let $k_{x,i}$ be the size in nucleotides of that ith component of x.
Let y_v^u be the **decision boolean variable** indicating that $s[u]$ and $s[v]$ form a canonical base pairing. Then, we always have $u \neq v$.
Let $C_p^{x,i}$ be the **decision boolean variable** indicating that we do insert the ith component of motif x at position p.

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 $C_p^{x,i}$, 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 $C_p^{x,i}$.

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

Objectives

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

$$f_1(X) = \sum_{x \in M} \left[(j^x)^2 \times \sum_{p \in P_{x,1}} C_p^{x,1} \right]$$

$$f_2(X) = \sum_{u<v} p_{uv} \times y_v^u \times I[p_{uv} > \theta], \qquad p_{uv} = \sum_{\sigma \in S(s)} y_v^u . p(\sigma|s)$$

f_1 is supposed to maximize the number of inserted motives in s, weighted by their number of components. f_2 is supposed to maximise the expected accuracy of the secondary structure (MEA model). p_{uv} are the base pairing probabilities that can be estimated from a set $S(s)$ of secondary structures of s.

10 Constraints to bind them all

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

$$\sum_{v>u} y_v^u + \sum_{v<u} y_u^v \leq 1 \qquad \forall u \in [\![1, n]\!] \tag{1}$$

Constraints to forbid lonely base pairs

$$1 + \sum_{v=u}^{n} y_v^{u-1} - \sum_{v=u+1}^{n} y_v^u + \sum_{v=u+2}^{n} y_v^{u+1} \geq 1 \qquad \forall u \in [\![1, n]\!] \tag{2}$$

$$1 + \sum_{u=1}^{v-2} y_{v-1}^u - \sum_{u=1}^{v-1} y_v^u + \sum_{u=1}^{v} y_{v+1}^u \geq 1 \qquad \forall v \in [\![1, n]\!] \tag{3}$$

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).

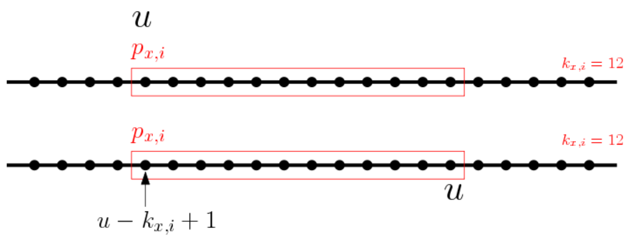
Constraint to forbid pairings inside a motif component

$$k_{x,i}\, C_p^{x,i} + \sum_{u=p}^{p+k_{x,i}-1} \left[\sum_{v>u} y_v^u + \sum_{v<u} y_u^v \right] \leq k_{x,i} \qquad \forall p \in P_{x,i} \qquad \forall x \in M, i \in [\![1, j^x]\!] \tag{4}$$

Constraint to forbid component to overlap

$$\sum_{x \in M} \sum_{i=1}^{j^x} \sum_{\substack{p \in P_{x,i} \\ u-k_{x,i}+1 \leq p \leq u}}^{u} C_p^{x,i} \leq 1 \qquad \forall u \in [\![1, n]\!] \tag{5}$$

Then, whatever the nucleotide u, it can be part of a motif 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 ($\forall x \in \{x \in M | j^x \geq 2\}$)

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 motif.

$$C_p^{x,i} \leq \sum_{\substack{p' \in P_{x,i+1} \\ p' > p + k_{x,i} + 2}} C_{p'}^{x,i+1} \qquad \forall p \in P_{x,i} \qquad \forall x \in \{x \in M | j^x \geq 2\}, i \in [\![1, j^x[\![\tag{6}$$

$$C_p^{x,i} \leq \sum_{\substack{p' \in P_{x,i-1} \\ p' < p - k_{x,i-1} - 2}} C_{p'}^{x,i-1} \qquad \forall p \in P_{x,i} \qquad \forall x \in \{x \in M | j^x \geq 2\}, i \in]\!]1, j^x]\!] \tag{7}$$

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

$$\sum_{p \in P_{x,1}} C_p^{x,1} - \frac{1}{j^x} \sum_{i=1}^{j^x} \sum_{p \in P_{x,i}} C_p^{x,i} = 0 \qquad x \in M \tag{8}$$

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

$$C_p^{x,i} \leq \sum_{p' \in P_{x,i-1}} y_p^{p' + k_{x,i} - 1} \qquad \forall x \in \{x \in M | j^x \geq 2\}, i \in]\!]1, j^x]\!] \tag{9}$$

$$C_p^{x,i} \leq \sum_{p' \in P_{x,i+1}} y_{p'}^{p + k_{x,i} - 1} \qquad \forall x \in \{x \in M | j^x \geq 2\}, i \in [\![1, j^x[\![\tag{10}$$

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 motif 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 motif 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] 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.

[4] 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.

[5] 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.