Clustering strings with mutations using an expectation-maximization algorithm In the context of RNA structure prediction

Afaf Saaidi, Yann Ponty, Mireille Regnier

Clustering strings with mutations using an expectation-maximization algorithm

In the context of RNA structure prediction

Afaf **Saaidi** [1,2] Yann Ponty [2] Mireille Régnier [2]

[1]School of Mathematics, Georgia Institue of Technology , [2]Laboratory of Computer Sciences LIX, Ecole Polytechnique, France

Outline

In comparative analysis, an RNA **structure** (a set of **base pairs** and **unpaired** nucleotides) is predicted from a set of RNA **variants** (similar sequences) under the assumption of the conservation of the structure during evolution.
The combination of RNA variants with Experimental data informing about the **local** (nucleotide) **structure** may lead to more accurate structure prediction.
The experimental protocol consists of mutating nucleotides likely to be 'unpaired'.
A simultaneous reading of RNA variants sequences that underwent the experimental mutation protocol lead to the following issue:
How to cluster 'mutated' substrings of similar parent strings such that each substring is correctly assigned to its parent string?
We developed an **E**xpectation **M**aximization algorithm that uses Mutational profiles (mutation distributions) to assign the substrings to their strings of origin.

RNA structure

- RNA is key to understand many biological processes (As in viral RNA).
- RNA maintains a stable **functional structure** during its **evolution**.
- Computational methods allow to have accurate 2D structure predictions ($PPV \approx 75\%$), less accurate predictions for long RNA.
- + Experimental probing data informing about **local** (nucleotide) structure improve predictions.

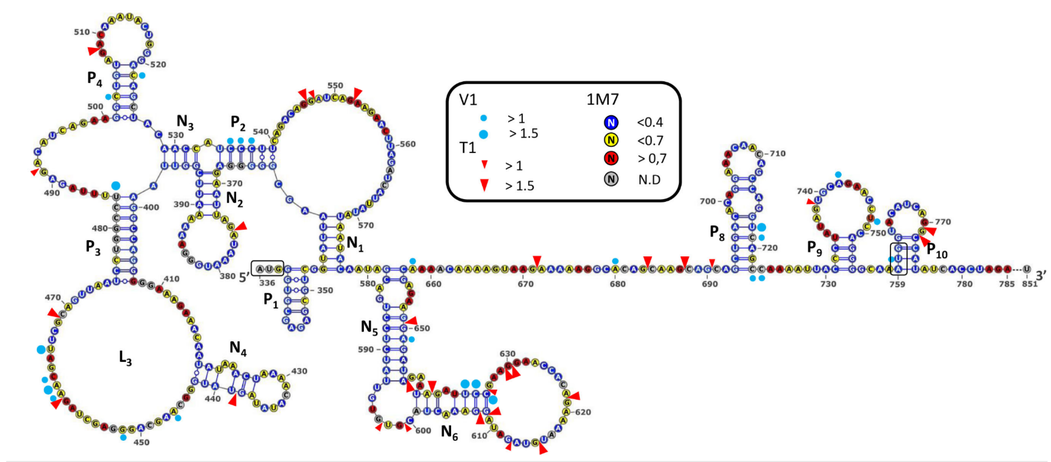

Figure 1: RNA secondary structure model of the HIV1 Gag-IRES with the projection of Experimental probing data [1]

Mutational profiles

A set of 3 **similar** strings with symbols in the Alphabet $\Sigma = \{A, C, G, U\}$

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
WT	C	G	A	C	G	C	U	C	U	U
V_1	C	G	A	C	**U**	C	**A**	C	U	U
V_2	**U**	**A**	A	C	G	C	U	C	U	U

● Mutation of **'unpaired'** Nucleotide

The assignment problem

Given the substring r_1 **CAAC**:

Variant V	Locus (r_1, V)	Distance(r_1, V)
WT	p_1	$\underline{1}$
V_1	p_1	$\underline{1}$
V_2	p_1	$\underline{1}$

Q.: Since r_1 is showing the same nucleotide distance to all the variants, what is its string of origin ?
$\rightarrow$ **Mutational Profiles** with the **Expectation Maximization** algorithm.

The objective

Our goal is to resolve the assignment problem: given substrings derived from a set of similar sequences and where structural mutations happen, what are the original strings of each substring?

Key idea of using the Expectation Maximization algorithm

- **Circular dependency** between the mutational profiles and the assignments of substrings [2]
  $\rightarrow$ Performs a **joint inference** of the **mutational profile** and **the assignment**

$$L(\theta; x, z) = \prod_{i=1}^{r} \prod_{j=1}^{V} \left[\frac{f(x_i; M_j)}{V}\right]^{1_{z_i=j}}$$

- θ: **Mutation probabilities** $\theta = \{M_j\}_{j=1}^{V}$
- M: **Mutational profile** $Mj : [1, n] \times \{m, \bar{m}\} \rightarrow [0, 1]$ for v_j
- **Density function**:

$$f(x_i; M_j) = \prod_{k \in [s_i, e_i]} M_j(k, x_{i,k-s_i})$$

- **Assignment probability**:

$$T_{j,i}^{(t)} := P(Z_i = j \mid X_i = x_i; \theta^{(t)}) = \frac{f\left(x_i; M_j^{(t)}\right)}{\sum_{j'=1}^{V} f\left(x_i; M_{j'}^{(t)}\right)}$$

Steps on EM for the substrings assignment

1. **Initialization**: For each substring j choose the initial estimates $M_j(k, c)^{(0)}$, $c \in N$, and compute the initial loglikelihood.
2. **E step**: Compute the density function $f(x_i; M_j^{(t)})$.
3. **M step**: Update $T_{j,i}$ then, compute the new estimates $M_j(k, c)^{(t+1)}$, $c \in N = \{\mathrm{m}, \bar{\mathrm{m}}\}$.
4. Iterate until convergence.

Results on simulated substrings with mutation

Set	Iteration	ET(s)	EM-assignment correct	EM-assignment incorrect	TMAP (MapQ ≤ 1) correct	TMAP (MapQ ≤ 1) incorrect
1	1		**$\underline{292}$**	2144	$\underline{239}$	148
	2		295	2141		
	117		1444	992		
	1000	5713.5	**1538**	898		
2	1		**$\underline{983}$**	1243	$\underline{215}$	2
	2		1096	1130		
	295		1122	1104		
	1000	5537.8	**1138**	1088		

References

[1] J. Deforges, S. de Breyne, M. Ameur, N. Ulryck, N. Chamond, A. Saaidi, Y. Ponty, T. Ohlmann, and B. Sargueil.
Two ribosome recruitment sites direct multiple translation events within HIV1 Gag.
Nucleic Acids Research, 2017.

[2] Afaf Saaidi.
Multi-dimensional probing for rna secondary structure(s) prediction.
Ph.D. thesis, Chap.6, 2018.

Acknowledgment

- Ph.D. grant (2015-2018), **F**ondation pour la **R**echerche **M**édicale, France.
- Post-doc NIH grant (2019-2022).