



**HAL**  
open science

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

Afaf Saaidi, Yann Ponty, Mireille Regnier

► **To cite this version:**

Afaf Saaidi, Yann Ponty, Mireille Regnier. Clustering strings with mutations using an expectation-maximization algorithm In the context of RNA structure prediction. 34th Clemson Mini-Conference on Discrete Mathematics and Algorithms, Oct 2019, Clemson, United States. hal-02332313

**HAL Id: hal-02332313**

**<https://inria.hal.science/hal-02332313>**

Submitted on 24 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

### 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 **Expectation Maximization** 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.

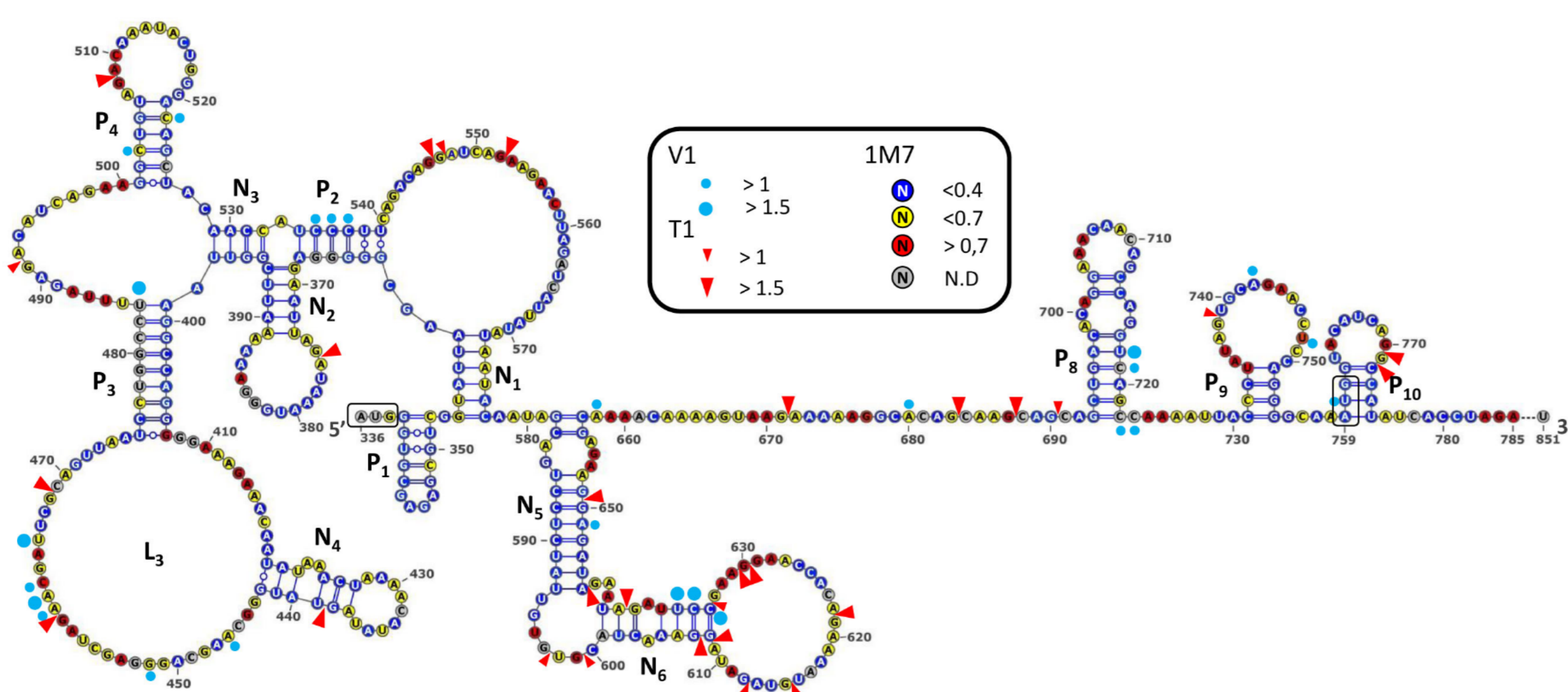
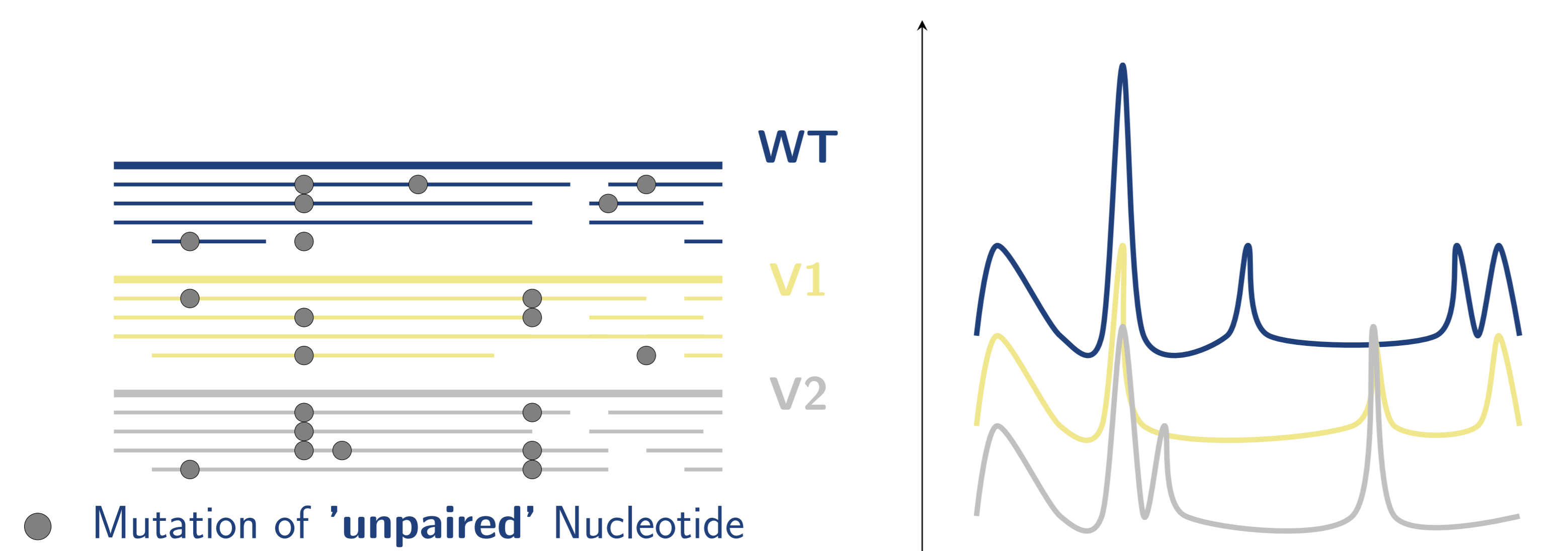


Figure 1: RNA secondary structure model of the HIV1 Gag-IRES with the projection of Experimental probing data [1]

### Mutational profiles

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

	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>	P <sub>9</sub>	P <sub>10</sub>
WT	C	G	A	C	G	C	U	C	U	U
V <sub>1</sub>	C	G	A	C	U	C	A	C	U	U
V <sub>2</sub>	U	A	A	C	G	C	U	C	U	U



### 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 <sub>1</sub>	$p_1$	$\underline{1}$
V <sub>2</sub>	$p_1$	$\underline{1}$

**Q.:** Since  $r_1$  is showing the same nucleotide distance to all the variants, what is its string of origin ?

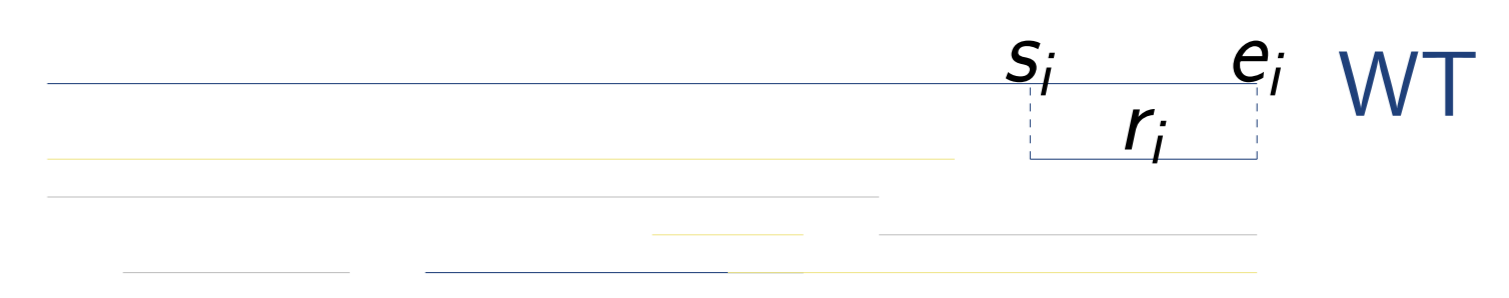
→ **Mutational Profiles** with the **Expectation Maximization** algorithm.

### 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]  
→ Performs a **joint inference** of the **mutational profile** and the **assignment**



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

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

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

### 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  $M_j(k, c)^{(t+1)}$ ,  $c \in N = \{m, \bar{m}\}$ .
4. Iterate until convergence.

### Results on simulated substrings with mutation

Set	Iteration	ET(s)	EM-assignment		TMAP (MapQ $\leq 1$ )	
			correct	incorrect	correct	incorrect
1	1		<b>292</b>	2144	<b>239</b>	148
	2		295	2141		
	117		1444	992		
	1000	5713.5	<b>1538</b>	898		
2	1		<b>983</b>	1243	<b>215</b>	2
	2		1096	1130		
	295		1122	1104		
	1000	5537.8	<b>1138</b>	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), **Fondation pour la Recherche Médicale**, France.
- ▶ Post-doc NIH grant (2019-2022).