# Automatic exploration of the natural variability of RNA non-canonical geometric patterns with a parameterized sampling technique

Théo Boury, Vladimir Reinharz, Yann Ponty

# Automatic exploration of the natural variability of RNA non-canonical geometric patterns with a parameterized sampling technique

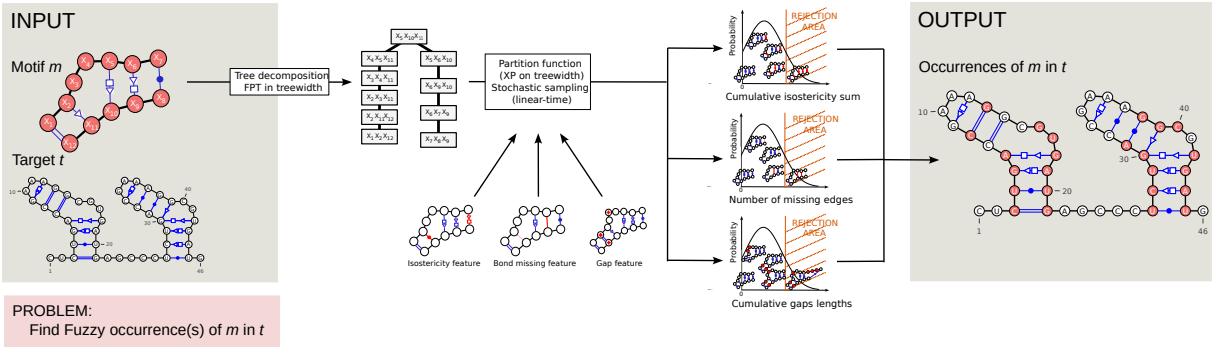
Théo BOURY[1,2,3], Vladimir REINHARZ[2] and Yann PONTY[1]

[1] Laboratoire de l'X (CNRS/LIX; UMR 7161), Institut Polytechnique de Paris, 91120, Palaiseau, France
[2] Department of Computer Science, Université du Québec à Montréal, H2X 3J8, Montréal, Canada
[3] Computer Science Department, Ecole Normale Supérieure de Lyon, 69007, Lyon, France

Corresponding author: theo.boury@lix.polytechnique.fr

***Reference paper:*** **Boury *et al.* (2023) Automatic exploration of the natural variability of RNA non-canonical geometric patterns with a parameterized sampling technique. *Workshop on Algorithms in Bioinformatics (WABI 2023)*.** https://doi.org/10.4230/LIPIcs.WABI.2023.20

**Keywords** Subgraph Isomorphism, 3D RNA, Parameterized Complexity, Kink-Turn family




## Abstract

The essential regulatory and catalytic roles played by RNAs in cellular processes can largely be attributed to the highly versatile nature of their structures. RNA structure is highly modular, particularly in the context of non-coding RNAs. Consequently, identifying structural motifs has become critical to understand their functions. Efforts have been made to cluster and classify RNA modules, attempting identify homologous occurrences and understand their evolution [1].

In this work, we address the mining of available 3D RNA data (PDB), to generate a catalog such as CaRNAval [2]. A common approach is to use discrete representations (graphs) to represent the secondary structure, possibly augmented with non-canonical interactions, and utilize a pragmatic solution to the NP-hard Subgraph Isomorphism problem. Such 2D methods typically manage to detect exact occurrence, but fail to report near-hits, overlooking the natural variability of RNA motifs. Moreover, their complexities are typically exponential.

Our FuzzTree method relies on a "2.5D" graphs representation: Edges represent one of the 12 possible (non-canonical) bonds [3], along with the 3D coordinates of the center of mass of each nucleotide, allowing access to the geometric proximity of nucleotides. This information allows FuzzTree to capture the variability of the motifs within RNAs. Finding and sampling approximate occurrences is done in bounded time and space, the method's complexity being only exponential on the treewidth.

## Method

FuzzTree samples occurrences of the target motif into a given RNA. All occurrences are Boltzmann-weighted based on a global notion of distance, so that exact occurrences are favored but the method also allows "small variations". To quantify the distance of a putative occurrence to the target motif, we introduced three criteria: bond-type compatibility, presence/absence of bonds, and length of successive gaps. Contributions of metrics were calibrated based on evolutionary and geometric models. Their relative weights can be set by the user to address specific use-cases (*eg* remote homology).

Behind the scene, our method uses dynamic programming to compute a partition function, followed by stochastic backtrack. To order the computation, the input graph is transformed into a tree decomposition,*ie* a "tree-like" structure whose nodes, also called bags, are sets of (related) nodes from the motif. The treewidth of a decomposition is the maximum size of a bag. Such a decomposition enables a bottom-up computation, using dynamic programming, of the partition function in time only exponential on the treewidth, which is minimized while computing the tree decomposition.

## Results

FuzzTree computes the partition function in XP complexity, namely in $\mathcal{O}(k\, n^{(tw+1)})$ time and $\mathcal{O}(n^{(tw+1)})$ space, for $tw$ the treewidth of the input motif over $n$ nucleotides, and $k$ the number of nucleotides in the motif. The treewidth parameterization is practically relevant as RNA motifs graphs have small treewidths (typically 2 or 3 in our dataset).

We have validated our method on a dataset composed of Kink-Turn families. By requesting a single well-chosen Kink-Turn motif, we were able to retrieve 82% of known Kink-Turns. Our method also found 2 occurrences, unannotated as Kink-Turns until now. Further analysis on our side concerning angles and composition, suggests them as close or related to Kink-Turns.

## Availability and Implementation

Our implementation relies on the `Infrared` [4] to automatically build a suitable tree decomposition. Our implementation runs in command line, can be fully customized to prioritize certain homology metrics, and is freely available at https://github.com/theoboury/FuzzTree.


## Funding information

Reinharz, Vladimir: NSERC RGPIN-2020-05795, FRQS CBJ1


## References


[1] Petrov AI, Zirbel CL, Leontis NB. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA. 2013 Oct;19(10):1327-40. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3854523/.

[2] Reinharz V, Soulé A, Westhof E, Waldispühl J, Denise A. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. Nucleic Acids Research. 2018 May;46(8):3841-51. Available from: https://doi.org/10.1093/nar/gky197.

[3] Stombaugh J, Zirbel CL, Westhof E, Leontis NB. Frequency and isostericity of RNA base pairs. Nucleic Acids Research. 2009 Apr;37(7):2294-312.

[4] Yao HT, Marchand B, Berkemer SJ, Ponty Y, Will S. Infrared: a declarative tree decomposition-powered framework for bioinformatics. Algorithms for Molecular Biology. 2024 Mar;19(1):13. Available from: https://doi.org/10.1186/s13015-024-00258-2.