# A Graph-Based Similarity Approach to Classify Recurrent Complex Motifs from Their Context in RNA Structures

Coline Gianfrotta, Vladimir Reinharz, Dominique Barth, Alain Denise

# A Graph-Based Similarity Approach to Classify Recurrent Complex Motifs from Their Context in RNA Structures

**Coline Gianfrotta**
Université de Versailles Saint-Quentin-en-Yvelines, Université Paris-Saclay, DAVID lab, France
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France

**Vladimir Reinharz**
Department of Computer Science, Université du Québec à Montréal, Québec, Canada

**Dominique Barth**
Université de Versailles Saint-Quentin-en-Yvelines, Université Paris-Saclay, DAVID lab, France

**Alain Denise**
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
Université Paris-Saclay, CNRS, I2BC, 91400, Orsay, France

## Abstract

This article proposes to use an RNA graph similarity metric, based on the MCES resolution problem, to compare the occurrences of specific complex motifs in RNA graphs, according to their context represented as subgraph. We rely on a new modeling by graphs of these contexts, at two different levels of granularity, and obtain a classification of these graphs, which is consistent with the RNA 3D structure.

RNA many non-translational functions, as a ribozyme, riboswitch, or ribosome, require complex structures. Those are composed of a rigid skeleton, a set of canonical interactions called the secondary structure. Decades of experimental and theoretical work have produced precise thermodynamic parameters and efficient algorithms to predict, from sequence, the secondary structure of RNA molecules. On top of the skeleton, the nucleotides form an intricate network of interactions that are not captured by present thermodynamic models. This network has been shown to be composed of modular motifs, that are linked to function, and have been leveraged for better prediction and design. A peculiar subclass of complex structural motifs are those connecting RNA regions far away in the secondary structure. They are crucial to predict since they determine the global shape of the molecule, therefore important for the function.

In this paper, we show by using our graph approach that the context is important for the formation of conserved complex structural motifs. We furthermore show that a natural classification of structural variants of the motifs emerges from their context. We explore the cases of three known motif families and we exhibit their experimentally emerging classification.

**2012 ACM Subject Classification** Applied computing → Molecular structural biology

**Keywords and phrases** Graph similarity, clustering, RNA 3D folding, RNA motif

**Digital Object Identifier** 10.4230/LIPIcs.SEA.2021.19

## 1 Introduction

RNA molecules are some of the major actors of the cell: many families of so-called non-coding RNAs intervene, along with proteins, in all major cellular processes. An RNA molecule is composed of a sequence of nucleotides (A, C, G, U) which folds in space into a three-dimensional structure. The function of an RNA molecule is strongly related to its three-dimensional structure. This is why many works since the 1970s have been dedicated

to predict the structure of any RNA molecule from its sequence. The folding depends on interactions between the nucleotides. Strong interactions, called canonical interactions, first form what is called helices, stacks of canonical base pairs. They connect loops, and form the skeleton of the structure. Those loops are composed of weaker interactions, called non-canonical interactions and give the molecule its final structure [14].

It has been observed that specific loop geometries are conserved and found through various RNAs with different functions, with varying sequence [12, 17]. This conservation has been leveraged by graph and other geometric methods to predict structure from sequence [18, 25]. Yet all those methods only focus on interactions networks within one loop, which have been extensively studied [21, 5]. While specific complex joining loops together are well known, as the A-minor [13], only recent algorithmic progress, using a graph representation of the RNA, have allowed to extend this automatic classification to combinations of loops connected between themselves through additional non-canonical interactions [23]. A major challenge in the field is the prediction of the location of those interconnected pairs of loops, a crucial determinant for the structure, and therefore the function of the RNA.

To tackle this challenge, we propose that the structural context of a motif [10] such as a A-minor in a molecule can be used as a discriminant for peculiar complex geometries, as those joining pairs of loops. It is a matter of determining whether two structurally similar contexts induce identical geometry and function. Considering a modeling of molecules by graphs [8] or hypergraphs, several definitions and similarity approaches between molecules have already been studied [22], mainly due to the principle stating that structurally similar molecules are expected to display similar properties [26, 9, 16, 24]. To measure the similarity of structures of molecules, one main approach considers the resolution of the problem of finding a Maximum Common Edge Subgraph [22] (MCES) between two graphs. This NP-complete problem is initially seen as a generalization of graph isomorphism, with different metrics evaluating the size of this subgraph compared to those of the two graphs to be compared, in particular some specific to a molecular context [26, 6, 1, 2].

When consider solving the MCES problem to measure the structural similarity of molecular graphs, two limitations could occur. First, the required computation time is exponential with respect to the number of vertices of the two graphs, which is a major limitation when considering comparing one molecule with all molecules in a database. Second, considering molecular graphs could provide a similarity measure not sufficiently focused on structural similarity, especially if two molecular structures are similar, but the associated graphs differ slightly in number and nature of vertices and links. This is why we introduce here a new graph representation of the molecular structures of RNA at a level of granularity lower than that of the nucleotides, allowing in particular a reduction in the size of the graphs to be processed. We then solve MCES problem on these graphs, based on specific subgraph isomorphism definition, to study the targeted structural similarity.

We show that the similarity in our new graph representation correlates with the geometric distance between the 3D models, while reducing by 75% the computation time. We validate our approach by applying it to three specific known and complex RNA structural motifs. We observe that the clustering induced by the similarity measure segregates well the different structural contexts. This study shows that the structural context matters for those complex motifs and could be leveraged for the prediction of their location.

# 2 Representation of the Context of RNA Structural Motifs

## 2.1 Prior Definitions

Our study will focus on a peculiar kind of interactions between nucleotides that contribute to the 3D shape of an RNA molecule: the canonical and non canonical interactions. These interactions belong to 12 pairing families, according to their geometry, defined in the Leontis–Westhof nomenclature [11]. Three faces on each nucleotide can interact with another nucleotide. The pairing family depends on the face of the two interacting nucleotides (Watson-Crick (W), Hoogsteen (H), or Sugar (S)) and on the orientation of the interaction (*cis* (C) or *trans* (T)). In the Leontis–Westhof nomenclature, these families are represented by a three-letter code indicating the orientation of the interaction and the faces of the two nucleotides (for example, CSS for cis Sugar-Sugar), or by a symbol (see examples in Figure 1 and all the symbols in appendix A.1 Table 3). The canonical interactions belong to the CWW family.

We will now give some prior definitions, useful for the construction of our representation. After defining an RNA graph, we will describe particular RNA graphs that we will focus on in this work. We will then define what an occurrence of one of this particular RNA graphs is. To finish, we will define a particular subgraph of RNA graph we will use in Section 2.2 to represent structural contexts.

▶ **Definition 1.** RNA graph

An **RNA graph** is a connected mixed graph $G = (V, A, E)$, with $A$ a set of directed edges, also called arcs, and $E$ a set of undirected edges. This graph represents all or part of an RNA tertiary structure. Vertices of $V$ correspond to nucleotides, edges of $A$ to the bonds of the primary sequence, and edges of $E$ to canonical and non-canonical interactions between nucleotides.

- The set $A$ of directed edges constitutes one or several path(s) forming the primary sequence of molecules, oriented from the 5' end to the 3' end.
- For each edge $[x, y] \in E$, we define a type $\boldsymbol{t([x, y])}$. This type corresponds to the pairing family to which the undirected edge belongs, according to the Leontis-Westhof nomenclature [11]. In particular, undirected edges corresponding to canonical bonds are annotated as such (CAN). Not all non canonical bonds are symmetrical, which is why $t([x, y])$ can be different from $t([y, x])$.
- For each vertex $x \in V$, we also define a type $\boldsymbol{\tau(x)}$ according to its direct neighborhood. This type will be taken into account in the search for graph isomorphisms (see Section 3.1).
  - $\tau(x) = 0$ if $x$ has no incident edge (belonging to $E$)
  - $\tau(x) = C$ if $x$ has just one incident edge $[x, y] \in E$ and if $t([x, y]) = CAN$.
  - $\tau(x) = N$ if $x$ has at least one incident edge $[x, y'] \in E$ such as $t([x, y']) \neq CAN$ and no incident edge $[x, y] \in E$ such as $t[x, y] = CAN$.
  - $\tau(x) = M$ if $x$ has an incident edge $[x, y] \in E$ such as $t([x, y]) = CAN$ and at least another incident edge $[x, y'] \in E$ such as $t([x, y']) \neq CAN$.
- For each vertex $x \in V$ such as $\tau(x) = C$, we define the ***canonical neighbor*** of $x$ as the neighbor $y \in V$ of $x$ such as $[x, y] \in E$ and $t([x, y]) = CAN$. By definition of vertex types $\tau$, this neighbor exists and it is unique because a nucleotide cannot form more than one canonical bond.

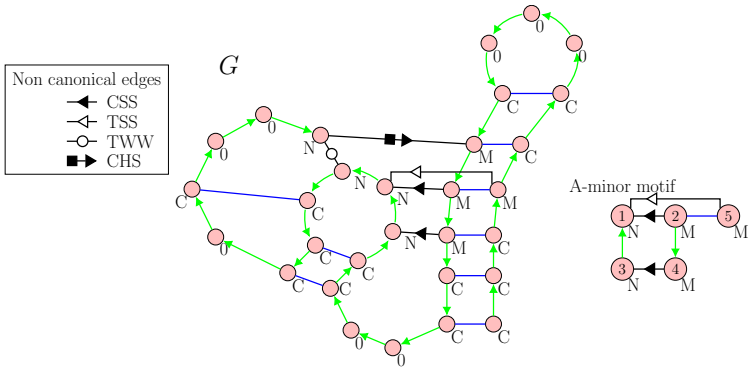



**Figure 1** Two examples of RNA graphs : a typical graph $G$ and the A-minor motif. The arcs are in green, the undirected edges of the canonical type are in blue and the undirected edges of the other types are in black, annotated by the Leontis–Westhof nomenclature [11]. Each vertex is annotated by its type.

Examples of RNA graphs are presented in Figure 1.

Note that, since we focus on the *structural* context only in this study, we do not consider the sequence (i.e. the types of nucleotides) in the RNA graph.

This work is focusing on particular RNA graphs, we called ***motifs***, that represent substructures frequently found in RNA tertiary structures as explained in the introduction (see Figure 1 for the example of A-minor motif).

▶ **Definition 2.** Motif occurrence
Given an RNA graph $G$, **a motif occurrence** is a partial subgraph of $G$, denoted as $O = (V^O, E^O, A^O)$, which is isomorphic to a motif $M = (V_M, A_M, E_M)$, with respect to the types of edges and vertices.

A motif occurrence is then a subgraph induced by the arcs and the edges of the motif $M$. The vertices of $V^O$ will be noted like the vertices of $V_M$ in $M$, for ease of writing. For example, for an A-minor motif, vertices of $V^O$ will be noted 1,2,3,4,5 (Figure 1). In the motif occurrence $O$, the types of the vertices of $V^O$ become specific to each vertex (for each $x \in V^O, \tau(x) = x$). An example of A-minor occurrence in an RNA graph $G$ is shown in Figure 2a.

▶ **Definition 3.** Specific subgraph of an RNA graph and a motif occurrence
Given an RNA graph $G$ and a motif occurrence $O$, the graph $\mathbf{G^{O,A^-} = (V, A \setminus A^O)}$ is the spanning subgraph of $G$ having no edge of $E$ and having all the arcs of $A$ except those of the motif occurrence $O$.

The graph $G^{O,A^-}$ is composed of chains, each of which containing at least one vertex of the motif occurrence $O$. There is an example of a graph $G^{O,A^-}$ in Figure 2b. This subgraph will help us to define the structural context of a motif occurrence (see Section 2.2).

## 2.2 Definition of a k-extension

As seen in introduction, we are interested in comparing the structural contexts of motif occurrences in RNA graphs. This context consists in a special subgraph which contains and surrounds the motif. As the bounds in the primary sequence play an important role in

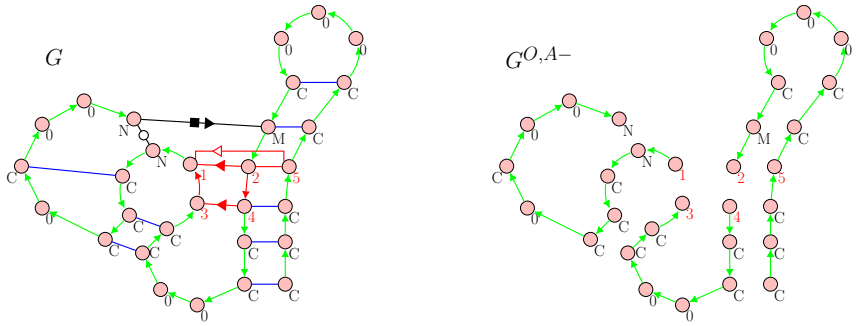



**Figure 2** In (a), the RNA graph $G$ with an A-minor occurrence in red, and in (b), the subgraph $G^{O,A-}$ of $G$. Every vertex is annotated by its type.

the tertiary structure, the graph $G^{O,A^-}$, which contains only this kind of bounds, plays a fundamental role in the definition of the context of a motif. We give below the definition of the structural context, that we call **k-extension of a motif occurrence**.

▶ **Definition 4.** $k$-extension of a motif occurrence $O$
Given a motif occurrence $O$, a subset $S$ of its vertices and an integer $k$, the **k-extension of a motif occurrence O according to S** is the subgraph $G_O = (V_O, A_O, E_O)$ of an RNA graph $G$, induced by three sets of vertices (which may be non-disjoint):

- the set $V^O$ of the vertices of the occurrence O (see definition 2)
- the set of vertices $V_k^O$ being at a distance strictly lower than $k$ in the graph $G^{O,A^-}$ (see definition 3), from one of the vertices of $S$.
- the set of vertices $V_k^{O+}$ neighbors by an edge of a vertex of $V_k^O$ in $G$.

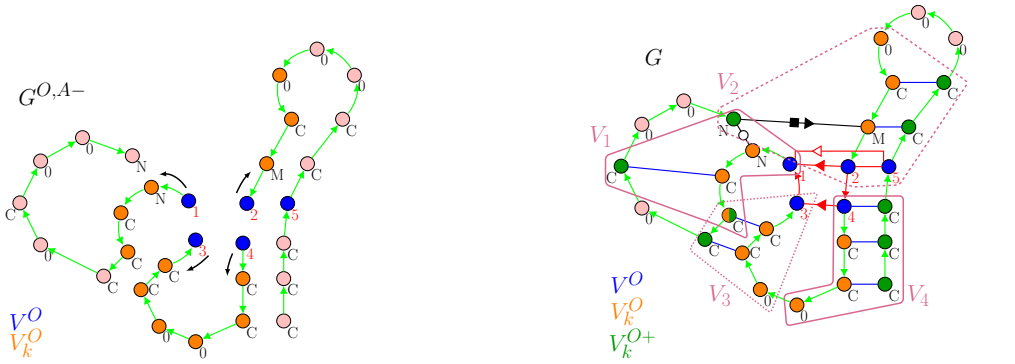
The set $S$ is the subset of vertices of the motif occurrence, from which we want to extend it. This choice will be explained in Section 5. For example, in the A-minor motif, we consider the first four vertices, i.e. the subset $\{1, 2, 3, 4\}$ (see the black arrows in Figure 3a).

In the example of RNA graph in Figure 3, the set of vertices $V^O$ is represented in blue, the set $V_k^O$ in orange and the set $V_k^{O+}$ in green (Figure 3b).

The vertices of $V_O$ are grouped into several subsets (not necessarily disjoint). In $G^{O,A-}$, we consider each path having for extremity one of the vertices of the motif occurrence $O$. The vertices of $V_k^O$ belonging to each of these paths constitute a subset of vertices. The vertices of $V_k^{O+}$ belong to the same subset(s) as their neighbor(s) in $V_k^O$. When we extend around an A-minor motif, we have in general four subsets of vertices, as shown in Figure 3b (framed in purple).

## 2.3 Definition of a Contracted k-extension

RNA structures are subject to modifications, due to evolution. Slight local changes in structures, like adding or deleting one nucleotide in a loop or an helix, may not change noticeably the 3D structure of the molecule, and thus may not change its function. This is why we present below a contracted representation of the context, allowing to represent similar but different contexts in an almost identical way. As will be seen in the Results section (Section 5), this new representation not only allows to better take the evolution into account, but also significantly decreases the computation time when comparing motif contexts.




**(a)** Graph $G^{O,A-}$ in which the vertices of $V^O$ are annotated in blue and the vertices of $V_k^O$ are annotated in orange for $k$ = 4.

**(b)** RNA Graph G in which the vertices of $V^O$ are annotated in blue, the vertices of $V_k^O$ in orange and the vertices of $V_k^{O+}$ in green for $k$ = 4.

**Figure 3** Construction of a k-extension for k=4. The vertices belonging to the k-extension are colored in blue, orange and green in both graphs $G^{O,A}$ (a) and $G$ (b). In (b) the four subsets of vertices ($V_1$, $V_2$, $V_3$ and $V_4$) are framed in purple.

We define a second graph $\tilde{G}_O$, derived from $G_O$, in which some edges and some vertices are contracted.

To do so, we have to define first the notion of *contractable path*, that will determine the vertices and the edges to contract. We define it in the graph $G_O^{O,A-}$ which is the spanning subgraph of the k-extension $G_O$ that contains no edge of $G_O$ and all the arcs of $G_O$ except those of the motif occurrence $O$ (see definition 3).

▶ **Definition 5.** Contractable paths in $G_O^{O,A-}$
A **contractable path** is a maximal path, in the graph $G_O^{O,A-}$ in which:
- the vertices are all of type C or all of type 0 and all belong to the same subset $V_k^O$ or all to the same subset $V_k^{O+}$
- and if the vertices are all of type C, the canonical neighbors of these vertices (see definition 1) also induce a contractable path in $G_O^{O,A-}$.

These paths connect vertices that are not involve in any edges (type 0) or vertices that are only involve in canonical edges (type C). It allows us to represent secondary structure elements, such as helices and loops, as blocs. Examples are presented in Figure 4.

▶ **Definition 6.** Contracted k-extension of a motif occurrence $O$
A **contracted k-extension**, denoted as $\tilde{G}_O$, is a graph derived from a k-extension $G_O$, in which the vertices of each contractable path in $G_O^{O,A-}$ are contracted in one single vertex. If these vertices are of type C, their canonical neighbors also induce a contractable path in $G_O^{O,A-}$ (according to the definition 5), and will thus be contracted. In this case, an edge of canonical type connects the two contracted vertices.

Each contracted vertex in $\tilde{G}_O$ is of the same type $\tau$ as the vertices from which it is derived in $G_O$. The number of vertices of $G_O$ grouped in $\tilde{G}_O$ in one single vertex $x \in V_O$ is noted $p(x)$. An example of graph $\tilde{G}_O$ is presented in Figure 4.

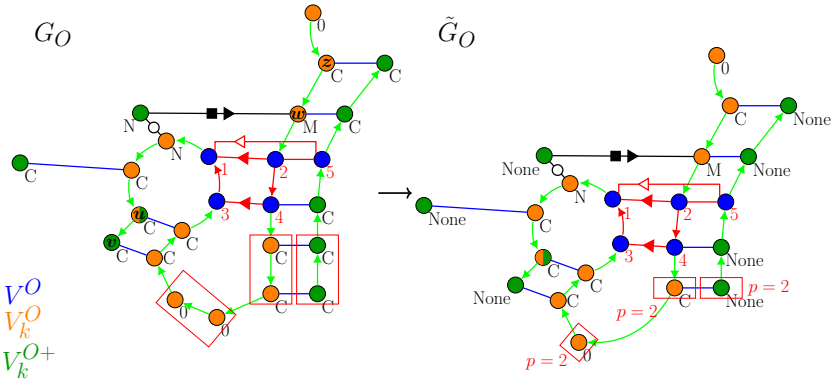



**Figure 4** On the left, the graph $G_O$ with 3 contractable paths circled in red. The path between the vertices $u$ and $v$ is not a contractable path because the vertex $u$ belongs to the subset $V_O^k$ and the vertex $v$ does not. In the same way, there is no contractable path between $w$ and $z$ because they are not of the same type (M for $w$ and C for $z$). Because of that, their canonical neighbors do not induce a contractable path either. On the right, the graph $\tilde{G}_O$ obtained by contraction, with the contracted vertices, framed in red and annotated by their weight $p$. The types of the vertices of $V_k^{O+}$ become *None*.

As defined in definition 5, the contractable paths are maximal paths, i.e. a set of contractable vertices cannot belong to a larger set of contractable vertices. Thus, the resulting graph $\tilde{G}_O$ is unique. Moreover, the same vertex cannot belong to two different contractable paths. Consequently, the graph $\tilde{G}_O$ does not depend on the order of treatment of the contractable paths.

In this model, the vertices of $V_k^{O+}$ in $\tilde{G}_O$ take the type *None* to differentiate them from the vertices of $V_k^O$.

The notations we defined in this section are summarized in the Table 1.

**Table 1** Summary of the graphs we define.

| | |
|---|---|
| $G$ | RNA graph |
| $O$ | motif occurrence in $G$ (subgraph of $G$, isomorphic to a motif) |
| $G^{O,A-}$ | spanning subgraph of $G$ containing no edge of $G$ and all the arcs of $G$ (except those belonging to the motif occurrence $O$) |
| $G_O$ | k-extension of a motif occurrence $O$ in $G$ (subgraph of $G$) |
| $G_O^{O,A-}$ | spanning subgraph of $G_O$ containing no edge of $G_O$ and all the arcs of $G_O$ (except those belonging to the motif occurrence $O$) |
| $\tilde{G}_O$ | contracted k-extension of a motif occurrence $O$ in $G$ (obtained from the contraction of vertices, arcs and edges in $G_O$) |

## 3 Similarity between Contracted k-extensions

We aim to compare the structural contexts of motif occurrences in RNA structures. For this purpose, we compare the contracted k-extensions of motif occurrences (noted $\tilde{G_O}$ in Section 2.3), in order to obtain, for each pair of contracted k-extensions, a common subgraph that maximizes a similarity metric we will define below in Section 3.2.

## 3.1 Maximum Common Subgraph : Variant of the MCES Problem

We will start by defining the maximum common subgraph on which we will calculate a similarity metric. To do so, we rely on the MCES problem.

The MCES problem aims to find a subgraph, common to any two graphs $G$ and $H$, maximizing the number of edges.

In our study, we search for a common subgraph as such, with supplementary constraints on the vertices and the edges, that we detail in the next paragraph.

Let $\tilde{G}_{O_1} = (\tilde{V}_{O_1}, \tilde{E}_{O_1}, \tilde{A}_{O_1})$ and $\tilde{G}_{O_2} = (\tilde{V}_{O_2}, \tilde{E}_{O_2}, \tilde{A}_{O_2})$ be two graphs of contracted k-extensions obtained from two motif occurrences $O_1$ and $O_2$ (Section 2.3).

We define $\tilde{G'}_{O_1} = (\tilde{V'}_{O_1}, \tilde{E'}_{O_1}, \tilde{A'}_{O_1})$ a subgraph of $\tilde{G}_{O_1}$ such that $\tilde{G'}_{O_1}$ contains the vertices of the motif occurrence $O_1$, and $\tilde{G'}_{O_2} = (\tilde{V'}_{O_2}, \tilde{E'}_{O_2}, \tilde{A'}_{O_2})$ a subgraph of $\tilde{G}_{O_2}$ such that $\tilde{G'}_{O_2}$ contains the vertices of the motif occurrence $O_2$.

We seek to find a subgraph $\tilde{G'}_{O_1}$, isomorphic to $\tilde{G'}_{O_2}$, and such that :

- each vertex $u \in \tilde{V'}_{O_1}$ is mapped to a vertex $v \in \tilde{V'}_{O_2}$ of the same type and belonging to a same subset of vertices (see 2.2),
- and such that each edge $[u_1, u_2] \in \tilde{E'}_{O_1}$ is mapped to an edge $[v_1, v_2] \in \tilde{E'}_{O_2}$ of the same type

Moreover, the subgraph $\tilde{G'}_{O_1}$ is not necessarily connected, but for all pairs of vertices $\{u, v\} \in \tilde{V'}^2_{O_1}$ in $\tilde{G'}_{O_1}$, if there is a path containing only arcs in $\tilde{G}_{O_1}$ between $u$ and $v$, there has to be a path containing only arcs in $\tilde{G}_{O_2}$ between the vertex mapped with $u$ in $\tilde{G'}_{O_1}$ and the vertex mapped with $v$ in $\tilde{G'}_{O_2}$. It means that the subgraph $\tilde{G'}_{O_1}$ must take into account the order of the vertices in these paths in the contracted k-extensions $\tilde{G}_{O_1}$ and $\tilde{G}_{O_2}$.

The subgraph $\tilde{G'}_{O_1}$ is thus a common subgraph to $\tilde{G}_{O_1}$ and $\tilde{G}_{O_2}$.

It has been shown that the decision problem associated with the calculation of a MCES between any two graphs is NP-complete [7]. Algorithms have been developed, able to solve the MCES problem for small instances, in particular for graphs representing molecules, such as the RASCAL algorithm [22]. This algorithm is an exact resolution of the problem. To find the MCES between two graphs G and H, it constructs the modular graph product P between the line graphs of G and H, and searches for a maximum clique in this graph P with a branch and bound method. We relied on this method to obtain the maximum common subgraph between our contracted k-extensions. We also developed a heuristic that builds the best common subgraph step by step, starting with the vertices with the highest degree.

## 3.2 Definition of the Similarity Metric to Maximize : the Contextual Graph Similarity

We will now explain how to evaluate the common subgraph we found, by defining a similarity measure, we call *contextual graph similarity*.

Although we have based ourselves on the RASCAL algorithm, the metric we want to maximize is slightly different. In the RASCAL algorithm, the similarity measure computes the number of edges and vertices in the common subgraph relative to the number of edges and vertices in the two initial graphs. Our contextual graph similarity takes into account only the number of edges, and not the number of vertices, in a common subgraph between two contracted k-extensions, because the interactions within an RNA molecule contribute the most to its tertiary structure. We do not consider arcs either, because we want to focus on the importance of canonical and non canonical interactions in RNA 3D structures.

▶ **Definition 7.** Contextual graph similarity
The **contextual graph similarity** between the two contracted k-extensions $\tilde{G}_{O_1}$ and $\tilde{G}_{O_2}$ is calculated as follows :

$$sim(\tilde{G'}_{O_1}, \tilde{G}_{O_1}, \tilde{G}_{O_2}) = \frac{\sum_{[u,v]\in \tilde{E'}_{O_1} \setminus E^{O_1}} min(p(u), p(u'))}{max(\sum_{[u,v]\in \tilde{E}_{O_1} \setminus E^{O_1}} p(u), \sum_{[u,v]\in \tilde{E}_{O_2} \setminus E^{O_2}} p(u))}$$

with $u' \in \tilde{V'}_{O_2}$ the vertex in $\tilde{G'}_{O_2}$ (subgraph of $\tilde{G}_{O_2}$ isomorphic to $\tilde{G'}_{O_1}$, see Section 3.1), that is mapped with $u \in \tilde{V'}_{O_1}$ in $\tilde{G'}_{O_1}$

We count the proportion of edges in the common subgraph $\tilde{G'}_{O_1}$ compared to the maximum number of edges between $\tilde{G}_{O_1}$ and $\tilde{G}_{O_2}$. We do not take into account the edges of $E^{O_1}$, i.e. the edges of the occurrence $O_1$ of the motif (or $O_2$ as the occurrences are isomorphic). Indeed, by definition, these edges are present in all contracted k-extensions.

The vertices incident to the same edge have necessarily the same weight, noted $p$ (see Section 2.3). Each edge in $\tilde{G'}_{O_1}$ is weighted by the minimum weight of its incident vertices in $\tilde{G'}_{O_1}$ and their mapped vertices in $\tilde{G'}_{O_2}$. Each edge in $\tilde{G}_{O_1}$ or $\tilde{G}_{O_2}$ is weighted by the weight of its incident vertices. This number corresponds to the number of nucleotides that the vertex represents (see Section 2.3). In this definition of the metric,the weights of the vertices are thus taken into account, which means that small differences in the structure will be counted. However, thanks to the contracted graphs, it is possible to parameterize the metric to take into account the weights of the vertices in a less restrictive way.

To illustrate the behaviour of our metric, examples of common subgraphs with high and low contextual similarities are shown in Figure 5.

We can note that we are interested in the maximum contextual graph similarity value between two contracted k-extensions, that can be obtained from several different common subgraphs.

## 4 Classification of k-extensions and Search for a Maximum Common Graph to a Class

We seek to establish a classification of contracted k-extensions of motif occurrences (noted $\tilde{G_O}$ in Section 2.3).

For this purpose, we define a graph $G_s = (V_s, E_s, \omega)$, called **similarity graph**, in which each vertex represents a contracted k-extension of motif occurrence and there is an edge between all pairs of vertices, weighted by the contextual similarity value. This weighting is noted by the function $\omega : E_s \to [0, 1]$. In this similarity graph, we remove the edges weighted by a value inferior to a threshold $s$. This threshold $s$ is set so that contracted k-extensions with contextual similarity less than $s$ are considered as not similar.

Then we define a **classification** as a set of subsets of vertices $W = \{V_{s1}, V_{s2}, ..., V_{sn}\}$, such that $V_s = \bigcup_{i=1}^{n} V_{si}$. The subsets of vertices in $W$ are not necessarily two by two disjoint, which means that a vertex can belong to two different classes. Our classification is therefore a coverage of $G_s$ and not a partition. It allows us to take into account the case where one contracted k-extension is close to two other contracted k-extensions, which are, for their part, very different.

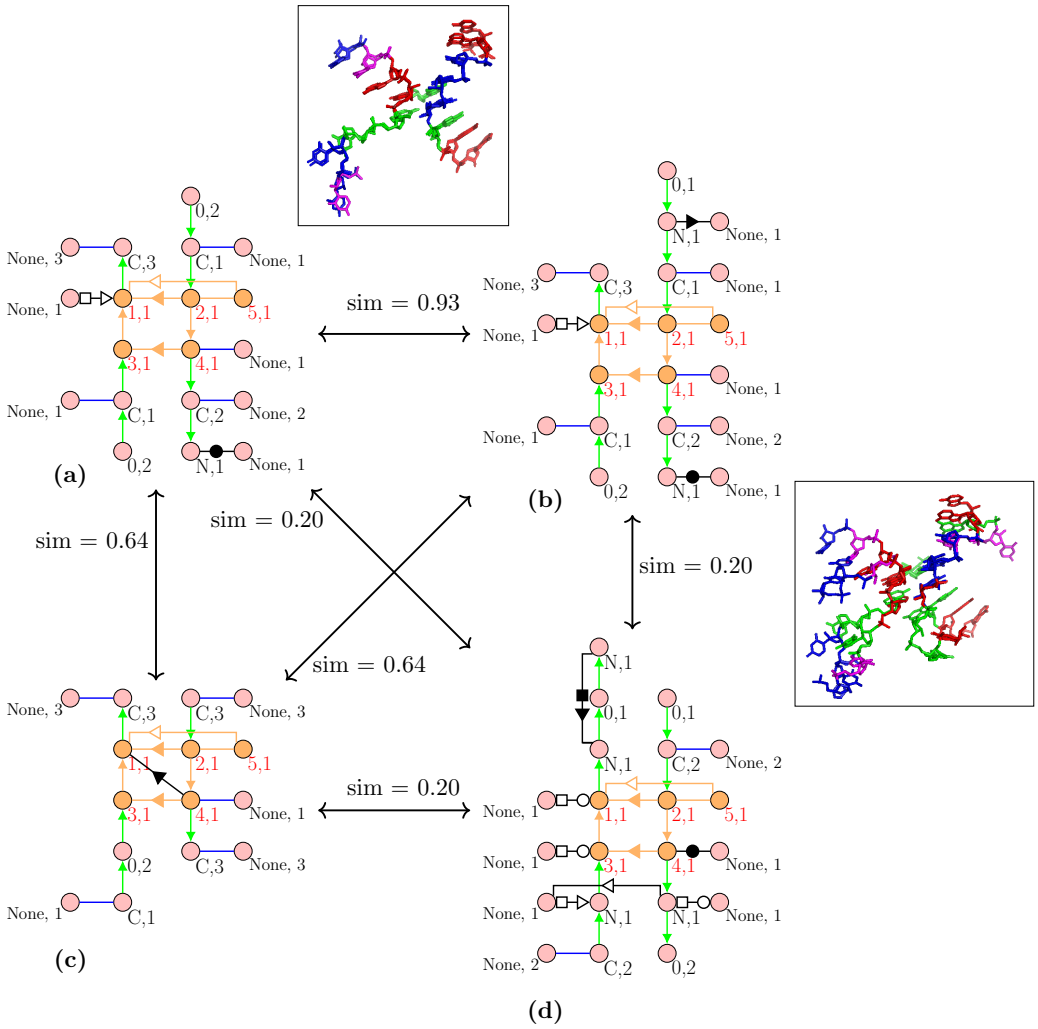
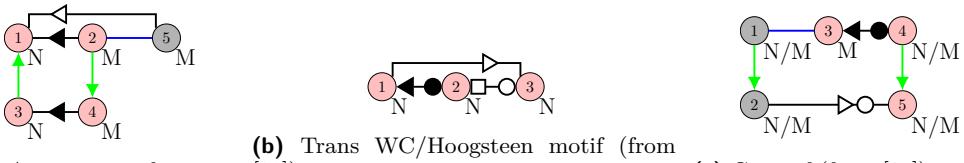



**Figure 5** Contextual graph similarities between four contracted 4-extensions. As seen before, arcs are represented in green, canonical edges in blue and non canonical edges in black, with the symbols of the Leontis–Westhof nomenclature. Each node is annotated by a doublet (type, weight). The 3D structure alignment of the 4-extensions (a) and (b) (resp. (b) and (c)) is presented at the top (resp. on the right). In the 3D structures, each type of nucleotides (A,C,G,U) is colored with the same colour. The two contracted 4-extensions above are the most similar (similarity of 0.93), and their corresponding 3D structures are very close too, as shown in the alignment. The 4-extension (c) has the smallest number of edges. However, it is still relatively similar to the 4-extensions (a) and (b) (similarity of 0.64). On the contrary, the 4-extension (d) is very different from the three others (similarity of 0.20), because it has many non canonical edges (represented in black) that do not appear in the other 4-extensions. The 3D alignment between (b) and (d) also highlights the differences.

We evaluate our classification according to cluster density and average similarity within clusters. Those two criteria allow us to obtain classes of similar contracted k-extensions, and so where the motif occurrences corresponding share close structural contexts. We thus apply a clustering method, developed in [20], that seeks to maximize those two criteria and also, that authorizes to obtain a coverage of the similarity graph and not a partition.




**(a)** A-minor motif. **(b)** Trans WC/Hoogsteen motif (from [23]). **(c)** G motif (from [23]).

**Figure 6** The three motifs we studied. The pink nodes constitute the subset of nodes from which we extend the motif to obtain the k-extensions (see Section 2.2).

We then characterize each of our classes by a representative. To do that, for each class of size $n$, we consider a maximum common subgraph to every contracted k-extensions of the class, defined in the same way as the maximum common subgraph for two graphs (Section 3.1), but for $n$ graphs. The quality of a class is notably linked to the size of this maximum common subgraph. The larger the size of the common subgraph, the more similar the contracted k-extensions of the class will be.

In Results section, we will analyze this classification in order to evaluate its relevance in a biological point of view.

## 5 Experimental Results

This section illustrates the relevance of our approach on three complex RNA motifs (Figure 6) : The A-minor motif, the Trans WC/Hoogsteen motif and a third motif which we call the G motif. These motifs are among those connecting RNA regions far away in the secondary structure. The A-minor motif frequently occurs in the RNA 3D structures, and has been proved to be involved in crucial cellular mechanisms [13]. The other two motifs come from the database of recurrent 3D motifs CaRNAval [23].

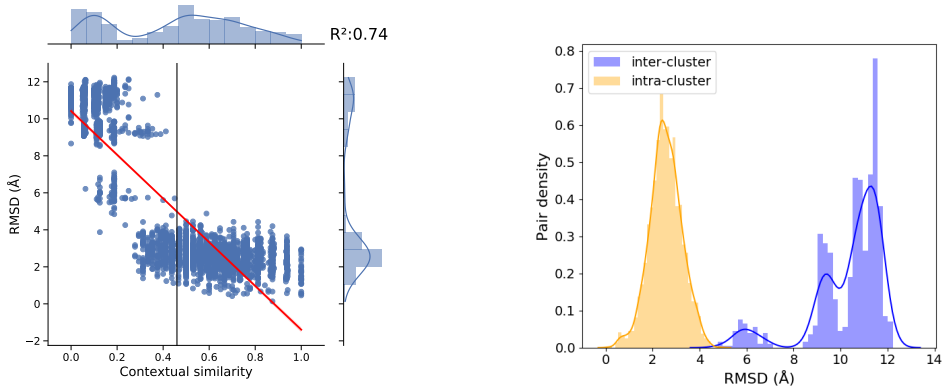
These three motifs are not predictable by current computational methods, to the best of our knowledge. That is why we choose to apply our method on those motifs. The fundamental question is: in terms of graphs, can the context of a motif help us to predict its presence in the molecule? The experiments shown in this section are intended to make progress towards answering this question. We first show that there is a clear correlation between our graph similarity and the geometrical similarity in 3D structures. Then we show that the automatic classification of RNA motif occurrences, based on our graph similarity, is consistent with RNA 3D structures. And finally, we show some advantages of the contracted graph representation, notably in terms of running time.

We applied our method on a dataset of non-redundant occurrences of those motifs from the PDB: 89 occurrences of the Trans WC/H motif, 391 of the A-minor motif, and 159 of the G motif. To choose the vertices from which we extend the motif (the subset $S$ in Definition 4), we considered vertices that are involved in non canonical edges and only one of the two incident vertices to a canonical edge (see the pink vertices in Figure 6).

We chose an extension size $k$ of 4 because this size gives us the most discriminating results.

### 5.1 Correlation between Graph Similarity and 3D Similarity

Firstly, we compare our contextual similarity measure to a measure of similarity on the 3D structures. To do so, we consider, for each contracted k-extension in our dataset, the 3D structure of the RNA graph induced by the contracted k-extension. We then use the




**(a)** Distribution of the RMSD values according to the contextual graph similarity values. The linear regression line of the distribution is presented in red, and the correlation coefficient $R^2$ is indicated. The histograms in the margin of the diagram show the distribution of values of contextual similarity (above) and RMSD (on the right).

**(b)** Distribution of the inter-cluster and intra-cluster RMSD values, with a clustering obtained with a contextual similarity threshold of 0.46 (black line in (a)).

**Figure 7** Two representations of the distribution of RMSD values in relation to the contextual graph similarity values, for the Trans WC/H motif occurrences.

RMSD (Root Mean Square Deviation) [3] as a quantitative measure of similarity between 3D structures. We align each pair of 3D structures nucleotide by nucleotide and calculate the RMSD by considering each nucleotide by its carbon 3', as it is usually done ([15], [4]). Then we compare the RMSD values to our contextual graph similarity values. The lower the RMSD, the more similar the two considered structures are. On the contrary, a contextual graph similarity value near to 0 (resp. 1) indicates that the k-extensions are very different (resp. very similar).

We present the distribution of the RMSD values according to the contextual graph similarity values, for the occurrences of the Trans WC/H motif (Figure 7a). The correlation coefficient $R^2$ associated with this distribution is very high (0.74). This is confirmed by the diagram where two main sets of dots can be observed, corresponding to the occurrences with a very high RMSD (superior to 7.5Å) and a very low contextual similarity (inferior to 0.25), and to the occurrences with a low RMSD (inferior to 4Å) and with a contextual similarity superior to 0.3.

For the other two motifs (results shown in appendix A.2.1, Figure 9), the correlation coefficient is equal to 0.33 for the A-minor motif and to 0.56 for the G motif. Those two motifs have thus a correlation coefficient, not as high as the Trans WC/H motif. They seem to be less dependent on their local environment.

## 5.2 Motif Classification

We also classified the contracted k-extensions, according to the contextual graph similarity. We aim to determine whether grouping motif occurrences by similar environment leads to different classes, and whether these classes are consistent with the RMSD. To do so, we used the clustering method detailed in Section 4. This method requires to choose a similarity threshold, below which the contracted k-extensions cannot be placed in the same cluster.

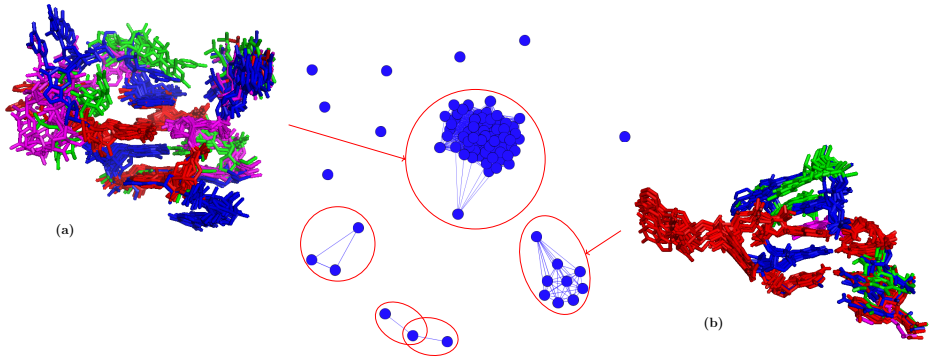



**Figure 8** Clustering and 3D alignments. In the middle, the similarity graph of the Trans WC/H motif occurrences. A node corresponds to an occurrence and there is an edge between two nodes if the contextual graph similarity between the contracted k-extensions is greater than 0.46. The clustering is indicated in red circles. On both sides, the 3D alignment of the contracted k-extensions of the two clusters. There is one color for each type of nucleotide. In (a), is presented the alignment of the 3D structures corresponding to a subset of contracted k-extensions of the largest cluster, and in (b), the alignment of the 3D structures corresponding to all of the contracted k-extensions of the smaller cluster.

We chose this threshold in an effort to have the better consistency between the contextual similarity values and the RMSD values. It means that the pairs of contracted k-extensions that have a contextual similarity value above (resp. below) the threshold must have similar (resp. not similar) 3D structures according to the RMSD. For the Trans WC/H motif, we choose a threshold of 0.46 because all the pairs of contracted k-extensions with a contextual similarity value above this threshold, correspond to 3D structures with a RMSD inferior to 5Å (Figure 7a). On the other hand, with this threshold, we lose some pairs with an RMSD value inferior to 3Å.

However, the Figure 7b shows a very clear consistency between the contextual similarity values and the RMSD values. Indeed, the RMSD of pairs of contracted k-extensions within a cluster does not generally exceed a value of 4.5Å and the RMSD of pairs of contracted k-extensions of two different clusters is generally superior to 4.5Å. This result thus also shows the relevance of our contextual graph similarity measure in relation to the RMSD.

Similar results hold for the two other motifs (see appendix A.2.1 Figure 10). The thresholds we have to choose to have a better consistency with RMSD values are higher (0.75 for the A-minor motif, and 0.65 for the G motif), and the maximum RMSD within clusters is higher for these two motifs, in particular for the A-minor motif where the RMSD values within clusters can reach 7Å for a few pairs of contracted k-extensions (appendix A.2.1 Figure 10).

We are now interested in the relevance of the classification itself. The classification we obtained for the Trans WC/H motif is composed of a large cluster of more than 60 occurrences, and four smaller clusters of respectively, 9, 3, 2 and 2 occurrences. The similarity graph (see Section 4) associated with this threshold is presented in Figure 8. It is a sufficiently dense graph for the classification to make sense, and to justify the use of a clustering method.

The cluster with 9 occurrences is composed of occurrences (and their contexts) sharing very close 3D structures (Figure 8, alignment (b)), and the maximum common subgraph (defined in Section 5.2) of the contracted k-extensions of this cluster is quite large. It is indeed composed of 6 edges (including 4 non canonical edges) which corresponds to one third

of the number of edges in the smallest contracted k-extension of the cluster. These motif occurrences are found in RNA molecules of the same family, which explains their high 3D similarity. The largest cluster is composed of occurrences (and their contexts) sharing less close 3D structure. Indeed, the 3D alignment for a subset of occurrences of this cluster in Figure 8 (alignment (a)) is quite good for the right parts of the structure, but differences appear for the top left part. These motif occurrences are found in RNA molecules of different but close families. The classification obtained with the two other motifs, available in appendix A.2.2 Figure 11, also groups together motif occurrences sharing close 3D structures. These results show that the classification based on the contextual graph similarity measure, is able to group together motif occurrences in relevant clusters that share very similar environments in 3D.

## 5.3 Advantage of the Contracted Representation

The contracted graph representation presents two main advantages. The first one is the running time: the time needed to execute the search for a maximum common subgraph is largely reduced for the contracted k-extensions, even though it stays exponential with the exact method. The results obtained with the exact method, on the three motifs, are presented in Table 2. For the A-minor motif occurrences, for example, which corresponds to our larger dataset (391 occurrences), the contracted representation makes it possible to divide the execution time by 4.

Perhaps more importantly, contrarily to other approaches based also on graph isomorphism (e.g. [19, 23]), our metrics allows us to consider slight changes in the number of vertices and edges in the graphs as identical (see Section 2.3). This allows to group together contexts which are different at the graph level, but very similar in terms of 3D structure.

# 6 Conclusion and Perspectives

In this study, we wanted to determine if the structural context of complex motifs in RNA structures can give useful information about the 3D structure of this context and thus help to predict the presence and the position of the motif in RNA 3D structures.

To find out, we represented the structural contexts of motif occurrences by specific graphs, at two granularity levels, and developed a method, based on solving a MCES problem, to compare them using a dedicated similarity metric. The MCES problem is used in many approaches looking for similarities between molecules [26, 6, 1, 2], but here we have some additional constraints on the graphs and a different metric. The granularity of the graphs we defined allows us to consider two structural contexts as similar even if slight differences occur

**Table 2** Execution time of the search for maximum common subgraph for each pair of k-extensions (contracted or non contracted) for the three datasets, on a PC Intel Core i5-7440HQ 4x2.80GHzCPU.

| Motif | Execution time (in hours) for contracted k-extensions | Execution time (in hours) for non contracted k-extensions |
|---|---|---|
| A-minor motif (391 occurrences) | 4 | 16 |
| G motif (159 occurrences) | 0,8 | 2,2 |
| Trans W/H motif (89 occurrences) | 0,22 | 0,45 |

in terms of nucleotides and bonds, since they have little impact in the 3D configurations.

Our results show that there is a clear correlation between our contextual graph similarity measure and the RMSD, which is a measure of similarity on the 3D structures (Section 5.1). Moreover, the reduced size of our graphs compared to graphs representing each nucleotide and each interaction separately, allows a considerable saving in computing time, especially when searching in databases. We also established an automatic and exhaustive classification of the three motifs we studied (Section 5.2). This classification is consistent with the 3D structures, which means that it groups together motif occurrences that share both close structural contexts and close local 3D structures.

Regarding perspectives, we now have to study further the motif classifications. Notably, it will be worth considering the A-minor motif, which is ubiquitous in RNA structures and for which there is no available prediction method. We believe that a method which combines both our graph approach and sequence considerations could lead to useful results. Many other motifs have also to be studied, notably from the CaRNAval database [23].

From a more theoretical point of view, we plan to refine our similarity measure by devising weights for different classes of modifications in the RNA graphs. For example, nucleotide insertions and deletions could give different costs according to these parameters. Then, computing parameter values would need a thorough study of motifs in databases.

## References


1 Faisal N. Abu-Khzam, Nagiza F. Samatova, Mohamad A. Rizk, and Michael A. Langston. The maximum common subgraph problem: Faster solutions via vertex cover. In *IEEE/ACS International Conference on Computer Systems and Applications*, pages 367–373, 2007. doi:10.1109/AICCSA.2007.370907.

2 Tatsuya Akutsu and Hiroshi Nagamochi. Comparison and enumeration of chemical graphs. *Computational and Structural Biotechnology Journal*, 5, 2013. doi:10.5936/csbj.201302004.

3 Rafael Brüschweiler. Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins*, 50(1):26–34, 2003. doi:10.1002/prot.10250.

4 Emidio Capriotti and Marc A. Marti-Renom. RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16):i112–i118, 2008. doi:10.1093/bioinformatics/btn288.

5 Grzegorz Chojnowski, Tomasz Waleń, and Janusz M. Bujnicki. RNA Bricks—a database of RNA 3D motifs and their interactions. *Nucleic acids research*, 42(D1):D123–D131, 2014.

6 Hanna Eckert and Jürgen Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, 12(5):225–233, 2007. doi:10.1016/j.drudis.2007.01.011.

7 Michael R. Garey and David S. Johnson. *Computers and intractability*, volume 29. WH Freeman New York, 2002.

8 Johann Gasteiger. *Handbook of Chemoinformatics: From Data to Knowledge*. Wiley, 1 edition, 2003. doi:10.1002/9783527618279.

9 Mark A. Johnson and Gerald M. Maggiora. *Concepts and applications of molecular similarity*. The American Chemical Society, 1988.

10 Neocles B. Leontis, Aurélie Lescoute, and Eric Westhof. The building blocks and motifs of RNA architecture. *Current Opinion in Structural Biology*, 16(3):279–287, 2006. doi:10.1016/j.sbi.2006.05.009.

11 Neocles B. Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, April 2001.

12 Neocles B. Leontis and Eric Westhof. Analysis of RNA motifs. *Current opinion in structural biology*, 13(3):300–308, 2003.

13 Aurélie Lescoute and Eric Westhof. The A-minor motifs in the decoding recognition process. *Biochimie*, 88(8):993–999, 2006. doi:10.1016/j.biochi.2006.05.018.

14 Aurélie Lescoute and Eric Westhof. The interaction networks of structured RNAs. *Nucleic acids research*, 34(22):6587–6604, 2006.

15 Marcin Magnus, Kalli Kappel, Rhiju Das, and Janusz M. Bujnicki. RNA 3D structure prediction guided by independent folding of homologous sequences. *BMC Bioinformatics*, 20(1):512, 2019. doi:10.1186/s12859-0193120-y.

16 Stefi Nouleho Ilemo, Dominique Barth, Oliver David, Franck Quessette, Marc-Antoine Weisser, and Dimitri Watel. Improving graphs of cycles approach to structural similarity of molecules. *PLOS ONE*, 14(12):1–25, 2019. doi:10.1371/journal.pone.0226680.

17 Carlos Oliver, Vincent Mallet, Roman Sarrazin-Gendron, Vladimir Reinharz, William L. Hamilton, Nicolas Moitessier, and Jérôme Waldispühl. Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic acids research*, 48(14):7690–7699, 2020.

18 Marc Parisien and François Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–5, 2008. doi:10.1038/nature06684.

19 Samuela Pasquali, Hin H. Gan, and Tamar Schlick. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucleic Acids Research*, 33(4):1384–1398, 2005. doi:10.1093/nar/gki267.

20 Airel Pérez-Suàrez, José F. Martínez-Trinidad, Jésus A. Carrasco-Ochoa, and José E. Medina-Pagola. An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition*, 46(11):3040–3055, 2013. doi:10.1016/j.patcog.2013.03.022.

21 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, 2013.

22 John Raymond, Eleanor Gardiner, and Peter Willett. RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs. *Computer Journal*, 45:631–644, April 2002.

23 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. doi:10.1093/nar/gky197.

24 Roger Sayle, John May, Noel O'Boyle, J. Andrew Grant, Stefan Senger, and Darren V.S. Green. Chemical similarity based on graph edit distance:efficient implementation and the challenges of evaluation. In *7th Joint Sheffield Conference on Chemoinformatics*, 2015.

25 Jason Yao, Vladimir Reinharz, François Major, and Jérôme Waldispühl. RNA-MoIP: prediction of RNA secondary structure and local 3D motifs from sequence data. *Nucleic acids research*, 45(W1):W440–W444, 2017.

26 Laura A. Zager and George C. Verghese. Graph similarity scoring and matching. *Applied Mathematics Letters*, 21(45):86–94, 2008. doi:10.1016/j.aml.2007.01.006.

# A Appendix

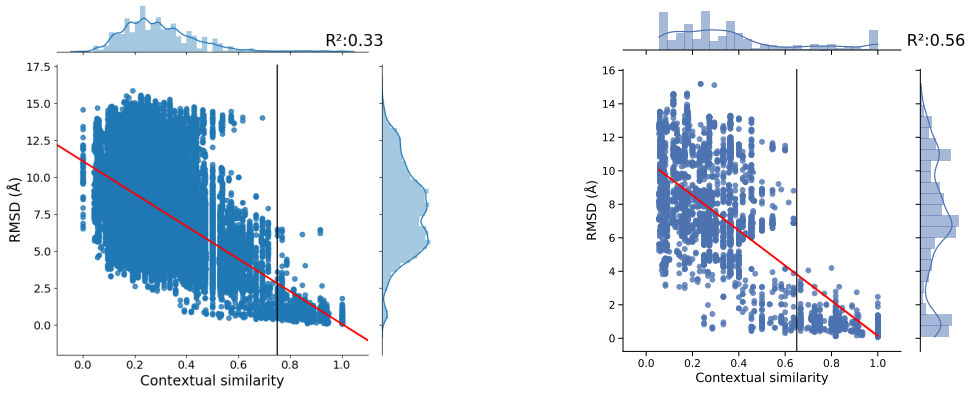
## A.1 Representation of the Context

**Table 3** Symbols of the Leontis–Westhof nomenclature for the non canonical interactions.

| Orientation | Interacting Edges | Symbol |
|---|---|---|
| *Cis* | Watson–Crick / Watson–Crick (cWW) | -●- |
| *Trans* | Watson–Crick / Watson–Crick (tWW) | -○- |
| *Cis* | Watson–Crick / Hoogsteen (cWH) | ●-■ |
| *Trans* | Watson–Crick / Hoogsteen (tWH) | ○-□ |
| *Cis* | Watson–Crick / Sugar Edge (cWS) | ●- ▶ |
| *Trans* | Watson–Crick / Sugar Edge (tWS) | ○-▷ |
| *Cis* | Hoogsteen / Hoogsteen (cHH) | -■- |
| *Trans* | Hoogsteen / Hoogsteen (tHH) | -□- |
| *Cis* | Hoogsteen / Sugar Edge (cHS) | ■- ▶ |
| *Trans* | Hoogsteen / Sugar Edge (tHS) | □- ▷ |
| *Cis* | Sugar Edge / Sugar Edge (cSS) | -▶- |
| *Trans* | Sugar Edge / Sugar Edge (tSS) | -▷- |

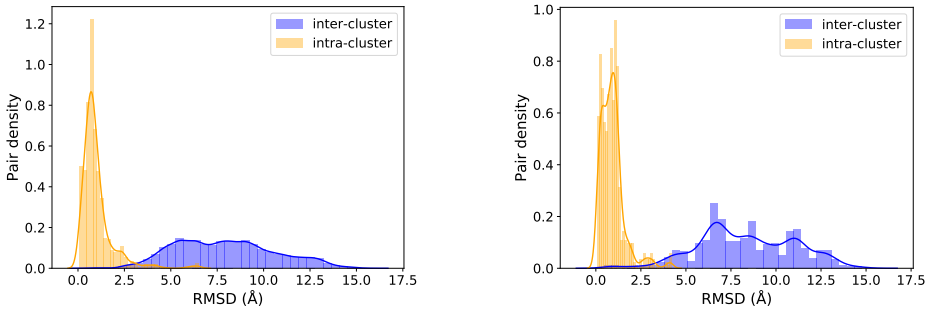
## A.2 Experimental Results

### A.2.1 Correlation between Graph Similarity and 3D Similarity




**(a)** A-minor motif. **(b)** G motif.

**Figure 9** Distribution of the RMSD values according to the contextual graph similarity values for the A-minor and the G motif. The linear regression line of the distribution is in red, and the correlation coefficient $R^2$ is indicated. The univariate distributions of RMSD and contextual graph similarity are presented in the margin of the diagram (above for the contextual graph similarity and on the right for the RMSD). The chosen contextual similarity threshold for the clustering for each motif is in black.

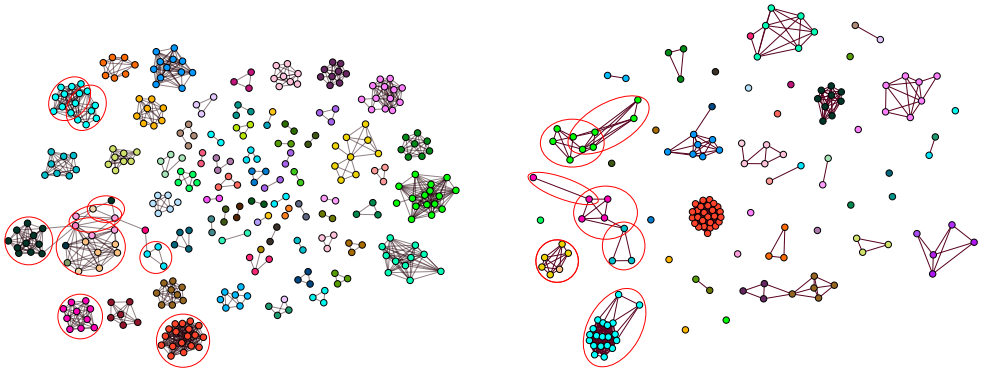



**(a)** A-minor motif (contextual similarity threshold = 0.75).

**(b)** G motif (contextual similarity threshold = 0.65).

**Figure 10** Distribution of the RMSD values intercluster and intracluster, for the two other motifs, with a clustering obtained with the contextual graph similarity thresholds indicated for each case.

## A.2.2 Motif Classification



**(a)** A-minor motif (contextual similarity threshold = 0.75).

**(b)** G motif (contextual similarity threshold = 0.65).

**Figure 11** Similarity graphs for the two other motifs, with the similarity threshold indicated in each case. A node corresponds to a motif occurrence, and there is an edge between two nodes if the contextual graph similarity is greater than the indicated threshold. Examples of clusters are circled in red in both graphs. Nodes of the same color correspond to motif occurrences found in the same RNA family.