Loop-loop Interaction Metrics on RNA Secondary Structures with Pseudoknots

Michela Quadrini and Emanuela Merelli
University of Camerino, via Madonna delle Carceri, Camerino, Italy

Keywords: Structure Comparison, RNA Secondary Structures, Interaction Distance, Pseudoknotted Structure, Loop.

Abstract: Many methods have been proposed in the literature to face the problem of RNA secondary structures comparison. From a biological point of view, most of these methods are satisfactory for the comparison of pseudoknot free secondary structures, whereas the problem of pseudoknotted motifs comparison has not been solved yet. In this paper, we propose *loop-loop interaction metrics*, a new measure able to compute the distance of two pseudoknotted secondary structures by comparing loops and their interactions. The new measure is defined for RNA molecules whose structural and biological information is represented as algebraic expressions of hairpin loops, so that each RNA secondary structure can be represented as a *word*, which describes the interactions among loops and uniquely defines the *intersection set*, the set of pairs of loops that cross. Hence, the interaction metrics is defined as the symmetric set difference applied to the intersection sets of molecules. To illustrate how to apply the proposed methodology, we compare two RNA molecules, PKB66 and PKB10, extracted from Pseudobase++ database. To test the validity of the measure, we evaluated the evolutionary conservation of the pseudoknot domain of Vertebrate Telomerase RNA.

1 INTRODUCTION

Ribonucleic acid (RNA) is a linear polymer of nucleotides arranged in a sequence referred to as a *primary structure*. This sequence is made of four different types of nucleotides, known as Adenine (**A**), Guanine (**G**), Cytosine (**C**) and Uracil (**U**). Such nucleotides are linked together by phosphodiester bonds in a way that the orientation can be established according to the polarity $5'$ to $3'$ of the molecule. Neutralization of the molecule determines the initial event of the folding process, which generates *complex three-dimensional shapes* (Dill, 1990), (Ferré-D'Amaré and Doudna, 1999). During such process each nucleotide can interact at most with one other nucleotide establishing a hydrogen bond. In this work, the phosphodiester bond between two consecutive nucleotides is referred to as a *strong interaction*, while the relations dynamically created during the folding process are called *weak interactions*. Both interactions are chemical bonds: the latter, in contrast to the former, are weak bonds that can be easily broken, and their formation is subject to restrictions. In fact, each nucleotide can form a base pair by interacting with another one performing the Watson-Crick base pairs (**G-C** and **A-U**) and wobble base pairs (**G-U**). In $2-$dimensions, the folding process can perform many RNA secondary structures; it depends on the free energy of RNA configurations. The RNA secondary structure is composed of five basic structural elements namely *hairpins*, *bulges*, *internal loops*, *multi-loops* and *helixes* (or *stacks*). Each structural element is generated when at least one base pair is performed. Thus, each of them is characterized by strong and weak interactions. We can observe that each structure element performs a *loop*, therefore secondary structures are composed of loops. If no interaction among loops is present, the secondary structure is *pseudoknot free*, as illustrated in Figure 1 (A), otherwise it is *pseudoknotted*, as depicted in Figure 1 (B).

Pseudoknots are tertiary structures that occur widely in RNA and they play a multitude of roles in the cell (Staple and Butcher, 2005), including the catalysis of various ribozymes (Rastogi et al., 1996), and the alteration of gene expression by inducing ribosomal frameshifting in many viruses (Shen and Jr, 1995). The biological functions of an RNA molecule depend on its structure (Laskowski and Thornton, 2008). The presumption is that to a preserved function corresponds a preserved configuration. In other words, the molecule cannot sustain substantial changes to its secondary and tertiary struc-

Quadrini, M. and Merelli, E.
Loop-loop Interaction Metrics on RNA Secondary Structures with Pseudoknots.
DOI: 10.5220/0006610700290037
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 3: BIOINFORMATICS*, pages 29-37
ISBN: 978-989-758-280-6

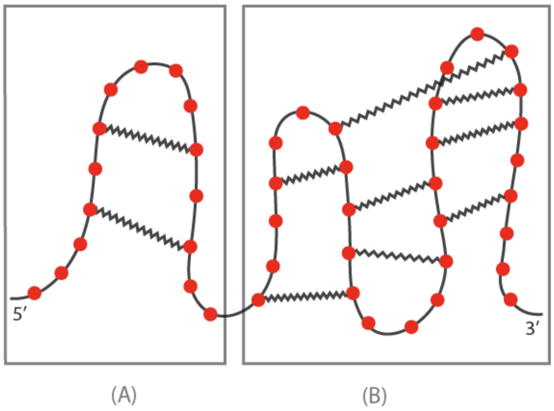

Figure 1: RNA secondary structure.

ture to preserve a particular function. Therefore, the structure comparison is used in the classification of RNA molecules, the prediction of the folding process and the measurement of the evolution stability. The comparison of RNA secondary structures is one of the main basic computational problems regarding the study of RNAs. In literature, many approaches have been proposed for facing this problem. One of them consists in the ordered trees comparison, but it works only for RNA pseudoknot free structures, since only this type of structure can be mapped into an ordered tree. The method for ordered trees comparison can be generally classified into two categories: *tree edition* and *tree alignment* (Herrbach et al., 2010). Both are based on the edit operations on nodes, i.e., node substitution, node insertion, and node delation. For each operation, a cost is associated. Thus, given two trees, through a sequence of edit operations, one changes into the other associating a cost which is given by the sum of the cost of each operation. In contrast to sequences, the alignment and edition model are not equivalent for trees. In fact, tree edition consists in constructing a common supertree, while tree alignment permits to find the common subtree. Which model is favourable depends on the biological problem of interest. It is trivial to observe that the edition problem is useful to identify the conserved structures during the folding process, while the alignment is suitable for clustering RNA molecules purely on the structure level. The problem of pseudoknotted motifs comparison has not been solved yet. Only few algorithms have been developed for studying specific cases of pseudoknots: the progress in this field has been hindered by the complexity of the problem. From an algorithmic perspective, the problem of comparing RNA structures is usually formalized as the comparison of arc-annotated sequences featuring crossing interactions. An arc-annotated sequence is a sequence over a given alphabet, together with additional structural information specified by arcs connecting pairs of positions. The problem of computing a distance between two arc-annotated sequences was introduced in (Evans, 1999) with a model that used only three edit operations either on single nucleotides or base pairs: it has been proved by (Blin and Touzet, 2006) that such a problem is NP-hard. Thus, a new representation of RNA secondary structures and a new approach for their comparison are necessary.

In this paper, we define a new measure, *loop-loop interaction metrics*, able to compute the distance of two pseudoknotted secondary structures in terms of interactions among loops. In particular, we use an algebraical representation of RNA secondary structures, both pseudoknot free and pseudoknotted, recently introduced by (Quadrini et al., 2017), that allows us to represent each RNA secondary structure as an algebraic composition of *hairpins*. In our model, the hairpin is the basic loop of such representation. Firstly, starting from such algebraic expression, we design an appropriate procedure to obtain the abstract algebraic expression of the structure, which allows us to define a proper set of functions for associating a word to each RNA secondary structure. The word permits the identification of interactions among loops and to define a unique set, *intersection set*. Such set is composed of all the pairs of loops that cross together. Finally, *interaction metrics* is defined as the symmetric set difference applied to the sets which identifies the crossing among loops. For illustrating in detail an application of our approach, it is applied over two RNA molecules, PKB66 and PKB10, extracted from Pseudobase++ database. To test the measure, we evaluated the evolutionary conservation of the pseudoknot domain of Vertebrate Telomerase RNA. The most featured of this structure is the evolutionary conservation of four structural domains: the pseudoknot domain, the CR4-CR5 domain, the Box H/ACA domain and the CR7 domain (Chen et al., 2000).

The paper is organized as follows. In Section 2, we present related works regarding the RNA secondary structure comparison. The measure of RNA secondary structures with pseudoknots, that we propose, is introduced in Section 3, which in turn is organized into three subsections. In the first subsection, we report an algebraic expression of RNA secondary structures in terms of hairpins. In the second, starting from the defined algebraic expression, we introduce an appropriate procedure to obtain the abstract algebraic expression of the structure. Moreover, a set of functions able to associate a unique word to each abstract algebraic expression is also defined in this subsection. In the last subsection, the measure is described and an example of its application is shown. The results and some critical considerations are discussed in Section 4. The paper closes with some conclusions and future work in Section 5.

2 RELATED WORKS

The structure of a molecule provides a framework for its biological functions (Laskowski and Thornton, 2008). Thus, the knowledge of structures is very important and the ability to compare them is useful in the study of the function and evolution of RNA. In the literature, there are several approaches to represent RNA secondary structures which consist of formalizing them in terms of base pair sets, trees, graphs or diagram representations. As a consequence, several approaches have been proposed for RNA secondary structure comparisons and corresponding similarity measurements. The simplest comparison metric is the base pair distance (Ding et al., 2005), which gives us the number of different base pairs between two structures. Other approaches are also possible, such as the symmetric set difference, the Hausdorff distance, and the mountain metric (Moulton et al., 2000).

For comparing structures using tree representation, a classical approach is to first define a set of basic and atomic operations, called edit operations, that allow to change a structure into another. The methods for ordered tree comparisons can be generally classified into two categories: tree edition and tree alignment (Herrbach et al., 2010). In terms of alignment, a wide amount of algorithms based on tree comparisons have been designed (Shapiro, 1988), (Le et al., 1989), (Corpet and Michot, 1994). In addition, several tree edit distance metrics have been developed (Shapiro and Zhang, 1990), (Moulton et al., 2000), (Dulucq and Tichit, 2003). However, these approaches are not able to take into account the pseudoknotted RNA secondary structures. Mohl *et al.* (Möhl et al., 2010) developed a type system for decompositions. The main idea is that the scheme of a folding algorithm can be transformed into a dynamic-programming algorithm for the alignment. Rastegari and Condon in their work (Rastegari and Condon, 2007) proposed a meta algorithm, which starts by determining the class of each structure, and then selects a suitable dynamic programming algorithm. Song *et al.* (Song et al., 2015) introduced a method for aligning two known RNA secondary structures with pseudoknots based on the partition function to calculate the scores of the alignments between bases or base pairs of the two RNAs with a dynamic programming algorithm. Moreover, Evans (Evans, 2011) in her work proposed a polynomial time algorithm for finding common RNA substructures that include pseudoknots.

3 MATERIAL AND METHODS

The new measure of RNA secondary structures with pseudoknots, that we propose, permits us to compare this kind of RNA structures in terms of interaction among loops. To define it, we use the algebraic expression, introduced by (Quadrini et al., 2017). Such algebraic expression is obtained from an appropriate operator able to model interactions among loops and the relative translation into a multiple context-free grammar. These two concepts are reintroduced in Section 3.1. For more details, the interested readers can refer to (Quadrini et al., 2017). Starting from this algebraic expression, we obtain its abstract algebraic expression through the definition of an appropriate procedure in Section 3.2. Moreover, in the same section, we also introduce a set of functions able to associate a unique word to each abstract algebraic expression. This word permits us to design another procedure to identify interactions among loops and to define a set, where elements are pairs that represent two crossing loops. In Section 3.3, the new measure is introduced.

3.1 Algebraic Expression for RNA Secondary Structures

Each RNA secondary structure is composed of loops, which can be formalized by the operator $\bowtie_k$. The operator maps two arc diagrams into another one, modeling each interaction among loops. It depends on a non-negative integer parameter, k, which indicates that the resulting structure is obtained by attaching the second arc diagram on the $k-$th nucleotides of the first one. The operator is well-defined if each nucleotide of the resulting structure performs at most one weak interaction. This restriction is due to the nature of RNA molecules. In other words, the situation illustrated in Figure 2 has to be excluded.

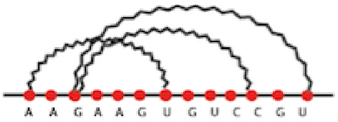

Figure 2: Not permitted structure.

It is also well-defined if the two structures do not share nucleotides, i.e., the first arc diagram is followed by the second one. In other words, the two structures are concatenated, as shown in Figure 3. Formally, it is obtained when k is equal to 0.

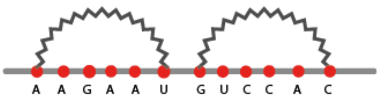

Figure 3: Concatenation of loops.

Algebraically, $\langle a_1^s, a_N^s \rangle[\ \alpha\ ]$ represents an RNA secondary structure. More specifically, α is the sequence of nucleotides (backbone) enclosed by the pseudoweak interaction, a fictitious weak interaction, between the first nucleotide, a_1, and the last one, a_N, identified by pair $\langle a_1^s, a_N^s \rangle$. See Figure 4 (A) for an illustration. Note that the molecule in Figure 4 (B) is a special case of an RNA secondary structure, referred to as a *pseudoloop* in this paper. It is an RNA secondary structure without head and tail. Algebraically, each nucleotide that performs a weak interaction with another one is represented by symbol $\sharp$, while the unpaired nucleotides are indicated by ε.

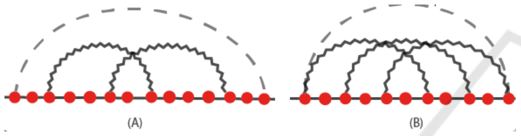

Figure 4: An example of secondary structure (A) and pseudoloop (B).

An example of the crossing operator application is illustrated in Figure 5. The second arc diagram is attached to the fifth nucleotide of the first arc diagram.

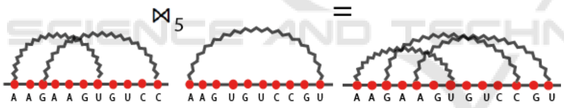

Figure 5: Example of a crossing operator application.

Formally, let S_1 and S_2 be two structures, where $S_1 = (a_1^s, a_N^s)\langle a_2^s \ldots a_{N-1}^s \rangle$ and $S_2 = (b_1^s, b_M^s)\langle b_2^s \ldots b_{M-1}^s \rangle$, the resulting structure, $S_1 \bowtie_k S_2$, is well defined if

$$\frac{k = 0,\ \ s \in \{\varepsilon, \sharp\}}{S_1 \bowtie_k S_2 \rightarrow (a_1^s, b_M^s)\langle\ a_2^s \ldots a_{N-1}^s a_N^s b_1^s \ldots b_{M-1}^s \rangle}$$

$$\frac{\begin{array}{c} k \leq N, s \in \{\varepsilon, \sharp\}, ((b_1 = a_k) \wedge BC), \\ ((b_2 = a_{k+1}) \wedge BC), \ldots, ((b_{N-k} = a_N) \wedge BC) \end{array}}{S_1 \bowtie_k S_2 \rightarrow (a_1^s, b_M^s)\langle\ a_2^s \ldots b_1^s \ldots b_{N-k}^s b_{N-k+1}^s \ldots b_{M-1}^s \rangle}$$

where BC expresses the biological constraint, i.e. each nucleotide performs at most one weak interaction, and it is formalized as follows:

$$BC : (s = \varepsilon, (\bar{s} = \varepsilon \vee \bar{s} = \sharp)) \vee (s = \sharp, \bar{s} = \varepsilon)\ .$$

This operator is translated into a *Multiple Context-Free Grammar* (MCFG), introduced in (Seki et al., 1991). This choice is due to the inadequacy of a Context-Free Grammar to describe the crossing dependence of pseudoknots; it can be proved by applying Ogdens Lemma (Harrison, 1978). Thus, a more expressive grammar is required.

Let $\Sigma_{RNA} = \{A, U, G, C\}$ be the alphabet of RNA nucleotides, and let $\Sigma_{\overline{RNA}} = \{(A,U), (U,A), (G,C), (C,G), (G,U), (U,G)\}$ be the alphabet of weak interactions, whose elements represent Watson-Crick or wobble base pairs. The first entry of each pair is the first nucleotide of the hydrogen bond, whereas the second one represents the corresponding complementary base pair. In other words, the nucleotides are identified by left, $\pi_1(a_1, a_2) = a_1$, and right, $\pi_2(a_1, a_2) = a_2$, which are canonical projection functions of the ordered pair. The grammar utilised is $G_{RNA} = (V_N, V_T, R, S, F)$, where $V_N = \{S, P, L\}$, $V_T = \Sigma_{RNA} \cup \Sigma_{\overline{RNA}} \cup \{[\ ,\ ]\}$, $F = \{f_{(\bowtie,k)}\}$ is the set of partial functions, and the set of productions R is defined as follows:

$$\begin{array}{llll} S & ::= & \alpha P \alpha & \textit{RNA secondary structure} \\ P & ::= & f_{(\bowtie,0)}[\![P\alpha, L]\!] & \textit{Concatenation} \\ & | & f_{(\bowtie,k)}[\![P, L]\!] & \textit{Nesting or Crossing} \\ & | & L & \textit{Hairpin} \\ L & ::= & x[\alpha^+] & \end{array}$$

where $x \in \Sigma_{\overline{RNA}}$, $\alpha \in \Sigma_{RNA}^*$ and

$$f_{(\bowtie,k)}[\![S, L]\!] = \begin{cases} S \bowtie_k L & \text{if } \bowtie_k \text{ is defined;} \\ \textit{undefined} & \text{otherwise.} \end{cases}$$

Such multiple context-free grammar G_{RNA} generates uniquely all RNA secondary structures; as a consequence, each secondary structure can be uniquely decomposed in terms of a particular loop, i.e., hairpin. The start symbol, S, represents any RNA secondary structure. The first production of the grammar formalizes the concatenation between an RNA pseudoloop P followed by a sequence of nucleotides α, eventually empty, and a loop L, whereas the second one represents both the crossing and the nesting between a pseudoloop P and a loop L. Finally, production $P \rightarrow L$ generates a loop. Each loop L is a hairpin, $L \rightarrow x[\alpha^+]$, i.e., a Watson-Crick or a wobble base pair encloses a sequence of unpaired nucleotides, α^+. For illustring an example, we take into account the structure PKB66 obtained from Pseudobase++ database (Taufer et al., 2008) illustrated in Figure 6.

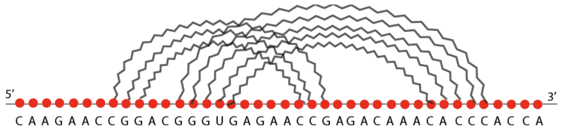
5′
3′
C A A G A A C C G G A C G G G U G A G A A C C G A G A C A A A C A C C C A C C A

Figure 6: The diagram of PKB66 molecule extracted from Pseudobase++ database (Taufer et al., 2008).

It is a pseudoknot of SELEX-isolated inhibitor of HIV-1 reverse transcriptase (Burke et al., 1996). The head and the tail of the structure are $\alpha_1 = CAAGAAC$ and $\alpha_{10} = ACCA$, respectively. The initial pseudoloop involves nucleotides from the 8-th to the 36-th. The pseudoloop is composed of crossings among weak interactions. Such crossings will be formalized making explicit hairpins. The order of choice of hairpins is well determined and such a choice depends on the complementary nucleotides of base pairs. In particular, the hairpin of the pseudoloop having the leftmost complementary nucleotides is selected. Thus, the first selected hairpin is $x_8[\alpha_9]$ where $x_8 = (G,C)$ and $\alpha_9 = GGUGAGAACCGAGACAAACACC$. In this way, the reduced pseudoloop involves nucleotides from the 8-th to the 35-th. In the following step, the hairpin $x_7[\alpha_8]$ has been explicited, where $x_7 = (G,C)$ and $\alpha_8 = GUGAGAACCGAGACAAACAC$. Moreover, each time that a hairpin is added it is necessary to formalize in which nucleotide of the relative pseudoloop the hairpin is attached. Thus, the algebraic expression of the structure is

$$S = \alpha_1 x_1[\alpha_2] \bowtie_2 x_2[\alpha_3] \bowtie_3 x_3[\alpha_4] \bowtie_6 x_4[\alpha_5] \bowtie_7 x_5[\alpha_6] \bowtie_8 x_6[\alpha_7] \bowtie_9 x_7[\alpha_8] \bowtie_{10} x_8[\alpha_9]\alpha_{10} \qquad (1)$$

where

$$\begin{array}{ll}
\alpha_1 = CAAGAAC & \\
\alpha_2 = GGACGGGUGAGAACC & x_1 = (C,G) \\
\alpha_3 = GACGGGUGAGAAC & x_2 = (G,C) \\
\alpha_4 = ACGGGUGAGAA & x_3 = (G,C) \\
\alpha_5 = AGAACCGAGACAAA & x_4 = (G,C) \\
\alpha_6 = GAGAACCGAGACAAAC & x_5 = (G,C) \\
\alpha_7 = UGAGAACCGAGACAAACA & x_6 = (G,C) \\
\alpha_8 = GUGAGAACCGAGACAAACAC & x_7 = (U,A) \\
\alpha_9 = GGUGAGAACCGAGACAAACACC & x_8 = (G,C) \\
\alpha_{10} = ACCA &
\end{array}$$

3.2 From the Algebraic Structure to the Intersection Set

The grammar, introduced in Section 3.1, permits the association of a unique algebraic expression for each RNA secondary structure in terms of hairpins. Such an algebraic expression contains the structural and biological information of the molecule. For each algebraic expression, it is possible to associate an abstract expression obtained by the first one by removing the nucleotides and introducing the position of the weak interaction into the structure. More specifically, each weak interaction divides the backbone into three parts, as illustrated in Figure 7, which are enumerated from left to right starting from 0.
For each algebraic expression

$$S = \alpha x[\alpha^+] \bowtie_k x[\alpha^+] \bowtie_k \cdots \bowtie_k x[\alpha^+]\alpha$$

by applying the procedure of Abstract Algebraic Expression, the abstract algebraic expression is obtained. In other words, such procedure takes in input the algebraic expression of an RNA molecule obtained from the multiple context free grammar and returns another algebraic expression,

$$S' = L \bowtie_t L \bowtie_t \cdots \bowtie_t L$$

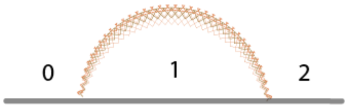

Figure 7: Backbone divided by an arc.

Note that t is a non-negative integer that represents the part of the backbone which the successive loop is attached to. Thus, the operator $\bowtie_t$ is different from the initial crossing operator: the initial one depends on nucleotides, whereas the second one depends on the part of the backbone. We decided to maintain the same symbol in order to not overload the notation.

```
Data: Algebraic Expression of RNA
      Secondary Structure
Result: Abstract Algebraic Expression
$N$ is the number of loops;
Let $\alpha_1$ be the length of $L_1$ ;
Let $d$ the length $P_1 - \alpha_1$ ;
for $i = 2$ to $N - 1$ do
    Compute $P_i$ ;
    $s = 0$ ;
    while $s \leq i$ do
        if $k_{i-1} = 0$ then
            $t = 2(i-1)$ ;
        else if $k_{i-1} < P_{1+s}$ then
            if $k_{i-1} > d$ then
                for $j = 1$ to $i - 1$ do
                    if $k_{i-1} \leq k_j$ then
                        $t = j + s$ ;
                    end
                end
            end
        else
            $s = s + 1$ ;
        end
    end
end
```

Algorithm 1: Abstract Algebraic Expression.

We take into account RNA molecule PKB66 introduced before and illustrated in Figure 6. Starting from its algebraic expression (1) and applying the procedure of Abstract Algebraic Expression, the relative abstract algebraic expression of the considered molecule is obtained. It is

$$S' = L_1 \bowtie_0 L_2 \bowtie_0 L_3 \bowtie_3 L_4 \bowtie_3 L_5 \bowtie_3 L_6 \bowtie_3 L_7 \bowtie_3 L_8 \tag{2}$$

Let S_A be the set of abstract algebraic expressions. Let $\mathcal{E} : S_A \to W_S$ be a rewriting rule that associates to each abstract expression another expression. Each loop is indicated by its starting , x_i, and final, $\overline{x_i}$ points, and a $\wedge_k$ is associated to each $\bowtie_k$. Note that the non-negative integer parameter k is the same for both expressions.

$$\mathcal{E}[\![L_i \bowtie_k S]\!] = \begin{cases} \mathcal{E}[\![L_i]\!]\mathcal{E}[\![\bowtie_k]\!]\mathcal{E}[\![S]\!] & \text{if } S = L_j \vee \\ & S = L_j \bowtie_k S \\ \perp & \text{otherwise.} \end{cases}$$

$$\mathcal{E}[\![L_i]\!] = x_i\, \overline{x_i} \quad i \in \mathbb{N}$$
$$\mathcal{E}[\![\bowtie_k]\!] = \wedge_k \quad k \in \mathbb{N}$$

Let $\mathcal{F} : W_S \to w$ be a rewriting rule that for each element of W_S associates a word that identifies uniquely the structure in terms of initial and final points of loops.

$$\mathcal{F}[\![\omega \wedge_k x_j\overline{x_i}]\!] =$$

$$\begin{cases} w_1 \ldots w_{k-1} x_j w_{k+1} \ldots w_N \overline{x_i} & \text{if } length(\omega) > k \\ \perp & \text{otherwise.} \end{cases}$$

$$\mathcal{F}[\![\omega \wedge_k x_j\overline{x_i} \wedge_{k'} x_s\overline{x_s}]\!] =$$

$$\begin{cases} \mathcal{F}[\![\omega' \wedge_{k'} x_s\overline{x_s}]\!] & \text{if } \omega' = \mathcal{F}[\![\omega \wedge_k x_j\overline{x_i}]\!] \\ \perp & \text{otherwise.} \end{cases}$$

For illustrating an application of the previous rewriting rules, we again consider the molecule PKB66. Applying the rewriting rule $\mathcal{E}$ to the Abstract Algebraic Expression 2, the following term is obtained

$$\omega_A = x_1\overline{x_1} \wedge_0 x_2\overline{x_2} \wedge_0 x_3\overline{x_3} \wedge_3 x_4\overline{x_4} \wedge_3$$
$$x_5\overline{x_5} \wedge_3 x_6\overline{x_6} \wedge_3 x_7\overline{x_7} \wedge_3 x_8\overline{x_8} \tag{3}$$

Applying the rewriting rule $\mathcal{F}$ to previous term, we have

$$w = x_1x_2x_3x_8x_7x_6x_5x_4\overline{x_1x_2x_3x_4x_5x_6x_7x_8} \tag{4}$$

For each word by applying the following procedure, the intersection Loop Set, the Intersection set is obtained. Such set is composed of all the pairs of loops that cross together.

The intersection set of the considered structure, obtained applying the previous algorithm, illustrated in Figure 6 is

$$\begin{aligned} V = \; & \{(L_1,L_4),(L_1,L_5),(L_1,L_6),(L_1,L_7),(L_1,L_8), \\ & (L_2,L_4),(L_2,L_5),(L_2,L_6),(L_2,L_7),(L_2,L_8), \\ & (L_3,L_4),(L_3,L_5),(L_3,L_6),(L_3,L_7),(L_3,L_8)\} \end{aligned}$$

```
Data: $w$, word associated to RNA secondary
      strucuture
Result: Intersection set associated to the
        structure
$N$ is the number of loops;
for $i = 1$ to $N$ do
    Select $x_i$ and $\overline{x_i}$ ;
    $w_i$ is the subword from $x_i$ to $\overline{x_i}$ ;
    $M_i$ is the length of $w_i$ ;
    $V = V \cup \{L_i\}$ ;
    for $j = 1$ to $M_i$ do
        Select $w_i[j] = a_j$ ;
        if $a_j = x_k$ and $\overline{x_k}$ is an element of $w_i$
          then
        else
            $V = V \cup \{L_k\}$ ;
        end
    end
end
```

Algorithm 2: Intersection Loops Set.

3.3 A Measure for Comparing RNA Secondary Structure

Each RNA secondary structure can be represented as an algebraic composition of hairpins, considered as basic loops. The new measure, that we propose, is based on the interactions among loops. Let S_1 and S_2 be two RNA secondary structures with pseudoknots. Let V_1 and V_2 be the respective intersection sets obtained applying the methodology introduced in Section 3.2. Each element of the two sets represents an interaction between two loops. For example, if the pair (L_1, L_2) is an element of V_1, it means that L_1 and L_2 are two loops of structure S_1 and they cross each other. Thus, two structures can be compared taking advantage of the set theory. Many methods have been proposed in literature. In this case, the *symmetric set difference* is a good first approach to evaluate the difference of structures.

Definition 1: The *interaction metrics* d_I is the cardinality of the *symmetrics difference* between the sets of interaction among loops V_1 and V_2,

$$d_I(V_1,V_2) = |(V_1 \setminus V_2) \cup (V_2 \setminus V_1)|$$

where V_1 and V_2 are the intersection sets of structure S_1 and S_2 , respectively. Note that $A \setminus B$ is the set of all elements that are in A, but not in B. Hence, we count the crossings present in either of the structures, but not in both. This *interaction loop distance* is a *metric*. This metric is very strict: all differences have the same weight. It does not take into account the backbone of the two structures. For illustating an

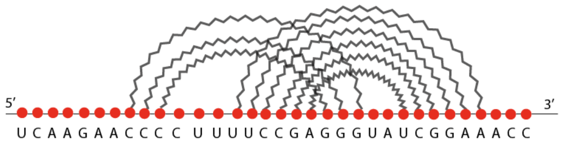
5′
3′
U C A A G A A C C C C U U U U C C G A G G G U A U C G G A A A C C

Figure 8: The diagram of PKB66 molecule extracted from (Taufer et al., 2008).

application of the proposed measure, we consider as examples two structures, S_1 and S_2. Let S_1 be PKB66 molecule illustrated in Figure 6 and let S_2 be PKB10 molecule illustrated in Figure 8. It is tRNA-like structure 3'end pseudoknot of ononis yellow mosaic virus.

Its intersection loops set is

$$\begin{aligned} V_2 = \; & \{(L_1,L_4),(L_1,L_5),(L_1,L_6),(L_1,L_7),(L_1,L_8),\\ & (L_1,L_9),(L_2,L_4),(L_2,L_5),(L_2,L_6),(L_2,L_7),\\ & (L_2,L_8),(L_2,L_9),(L_3,L_4),(L_3,L_5),(L_3,L_6),\\ & (L_3,L_7),(L_3,L_8),(L_3,L_9)\} \end{aligned}$$

The distance in terms of interaction among loops of the two considered structure is $d_I(V,V_2) = 3$. In fact, the cardinality of the difference between V and V_2 is 0 since each element of V is also in V_2, vice versa the cardinality of $V_2 \setminus V$ is 3 because of three pairs, $(L_1,L_9),(L_2,L_9),(L_3,L_9)$, are elements of V_2, but they are not in V.

4 RESULTS AND DISCUSSIONS

In this paper, we introduced a measure able to compare RNA secondary structures in terms of interactions among loops. In order to test the measure, we evaluated the evolutionary conservation of the pseudoknot domain of Vertebrate Telomerase RNA. Telomerase is a ribonucleoprotein enzyme that maintains telomere length by adding telomeric sequence repeats onto chromosome ends. The essential RNA component of telomerase provides the template for secondary structure of telomeric repeat synthesis. The most featured Vertebrate Telomerase RNA is the evolutionary conservation of four structural domains: the pseudoknot domain, the CR4-CR5 domain, the Box H/ACA domain and the CR7 domain (Chen et al., 2000). Applying the proposed methodology to the two pseudoknots, the distance is

$$d_I(V_H,V_S) = 0$$

where V_H and V_S are the intersection sets of the two pseudoknots domains of human and sharpnose shark telomerase RNAs, respectively. The result, $d_I(V_H,V_S) = 0$, shows that each interaction between loops belongs to both molecules. As consequences, the structure is conserved in accordance to the results present in the literature (Chen et al., 2000). Moreover, this measure, able to capture the interactions between loops, can be also applied to classify the molecules. Although two molecules of each pair are characterized by a functional similarity, the interaction among loops can differ. For example, we take into account a pair of molecules, extracted from (Taufer et al., 2008), that involves structural elements for translation initiation and ribosome recruitment found in the viral internal ribosome entry site (PKB223) and the V4 domain of 18S rRNA (PKB205) (Pasquali et al., 2005). Applying the proposed methodology to the two pseudoknots, the distance $d_I(V_{PKB223},V_{PKB205}) = 24$, where V_{PKB223} and V_{PKB205} are the intersection sets of the PKB223 and PKB205 molecules, respectively. This information can be considered as a *structural constrain* to guide the secondary structure folding. In fact, the biological presumption is that the RNA structure folds hierarchically. During the folding process, pseudoknot free structures are initially formed, whereas pseudoknots motifs are generated later to minimize the energy. Thus, a classification of the structures is useful to understand or classify how the structure evolves. Moreover, the measure can be also used to detect a mutation. At a structural level, the measure is able to capture the interactions between the loops. Each interaction between two loops is determined by a crossing of two base pairs. Taking the crossing among base pairs in consideration permits to define a more precise energy function than the standard one (Vernizzi et al., 2016).

The introduced measure is obtained taking advantage of the set theory. In particular, the symmetric difference of sets has been used. Other similarity functions can be applied to reach a more accurate measure based on interactions among loops. A list of useble similitarity functions is reported in Table 1.

Table 1: Similarity functions over two set X and Y.

Similarity Functions	Definition
Intersection	$S_B(X,Y) = \lvert X \cap Y \rvert$
Cosine	$S_C(X,Y) = \frac{\lvert X\cap Y \rvert}{\sqrt{\lvert X \rvert \lvert Y \rvert}}$
Dice	$S_D(X,Y) = \frac{2\lvert X\cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Hamming	$S_H(X,Y) = \lvert (X\cap Y)\cup(X\cup Y)^C \rvert$
Jaccard	$S_J(X,Y) = \frac{\lvert X\cap Y \rvert}{\lvert (X\cup Y)}$

From an algorithmic point of view, for each RNA molecule, in order to define the measure, we obtained a word that uniquely represents the secondary structure. Over this word, it is possible to define a set of rewriting rules that permits us to obtain the shape of each molecule. The shape is a topological concept widely used by Bon (Bon et al., 2008) and Reydis *et al.* (Reidys et al., 2011). Moreover, it is also possible to define an algorithm to compute some topological invariants, such as *genus* and *crossing number* (Vernizzi et al., 2016). Another possible procedure over the word can be easily defined to detect whether or not a pseudoknot belongs to a given class. Understanding if two structures are characterized by the same pseudoknots is useful for the choice of the particular algorithm for comparing the two structures taking into account the biological relevant operations such as addition, deletion, and substitution of nucleotides or base pairs.

5 CONCLUSIONS

The biological function of an RNA molecule depends on its structure. As a consequence, the molecule cannot sustain substantial changes to its secondary and tertiary structures to preserve the particular function. Thus, the knowledge of the structure is very important and the ability to compare the RNA structure motifs supports the study of function and evolution of RNA.

In this paper, we proposed a measure to compare RNA secondary structures with pseudoknots in terms of interactions among loops. From a biological point of view, it is useful to identify the conserved structures during the evolution since its primary structure is often unpreserved. In fact, this measure is able to detect the global properties of the molecules taking advantage of the set theory. Consequently, a benefit is that it can be computed quickly. Its properties make the measure easy to be handled theoretically. A statistical study over a large set of molecules can be performed in order to determine a new clusterization. This clusterization can be compared with others taken from differnt approaches present in the literature.

We plan to improve the developed software that implements the measure and the whole methodology presented in this paper in order to investigate and analyze in statistical terms the correlations between the proposed measure and the functions of RNAs. Moreover, we plan to evaluate the five similarity functions in order to classify the performance of the different similarity functions as measured. For reaching the goals, we have decided to compare molecules extracted from the Rfam (Nawrocki et al., 2015) database. This database classifies non-coding RNAs in families whose member posses a similar secondary structure, suggesting evolutionary relationships and similar functions. Moreover, this database provides a consensus secondary structure for each family.

ACKNOWLEDGEMENTS

We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM (www.topdrim.eu), number FP7-ICT- 318121.

REFERENCES

Blin, G. and Touzet, H. (2006). How to compare arc-annotated sequences: The alignment hierarchy. In *International Symposium on String Processing and Information Retrieval*, pages 291–303. Springer.

Bon, M., Vernizzi, G., Orland, H., and Zee, A. (2008). Topological classification of RNA structures. *Journal of molecular biology*, 379(4):900–911.

Burke, D. H., Scates, L., Andrews, K., and Gold, L. (1996). Bent pseudoknots and novel rna inhibitors of type 1 human immunodeficiency virus (hiv-1) reverse transcriptase. *Journal of molecular biology*, 264(4):650–666.

Chen, J.-L., Blasco, M. A., and Greider, C. W. (2000). Secondary structure of vertebrate telomerase rna. *Cell*, 100(5):503 – 514.

Corpet, F. and Michot, B. (1994). Rnalign program: alignment of rna sequences using both primary and secondary structures. *Computer applications in the biosciences: CABIOS*, 10(4):389–399.

Dill, K. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31):7133–55.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2005). Rna secondary structure prediction by centroids in a boltzmann weighted ensemble. *Rna*, 11(8):1157–1166.

Dulucq, S. and Tichit, L. (2003). Rna secondary structure comparison: exact analysis of the zhang–shasha tree edit algorithm. *Theoretical Computer Science*, 306(1-3):471–484.

Evans, P. (1999). Algorithms and Complexity for Annotated Sequences Analysis. PhD thesis, University of Victoria.

Evans, P. A. (2011). Finding common rna pseudoknot structures in polynomial time. *Journal of Discrete Algorithms*, 9(4):335 – 343.

Ferré-D'Amaré, A. and Doudna, J. (1999). Rna folds: insights from recent crystal structures. *Annual review of biophysics and biomolecular structure*, 28(1):57–73.

Harrison, M. A. (1978). *Introduction to formal language theory*. Addison-Wesley Longman Publishing Co., Inc.

Herrbach, C., Denise, A., and Dulucq, S. (2010). Average complexity of the jiang–wang–zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm. *Theoretical Computer Science*, 411(26):2423–2432.

Laskowski, R. and Thornton, J. (2008). Understanding the molecular machinery of genetics through 3D structures. *Nature Reviews Genetics*, 9(2):41–151.

Le, S.-Y., Owens, J., Nussinov, R., Chen, J.-H., Shapiro, B., and Maizel, J. V. (1989). Rna secondary structures: comparison and determination of frequently recurring substructures by consensus. *Bioinformatics*, 5(3):205–210.

Möhl, M., Will, S., and Backofen, R. (2010). Lifting prediction to alignment of rna pseudoknots. *Journal of Computational Biology*, 17(3):429–442.

Moulton, V., Zuker, M., Steel, M., Pointon, R., and Penny, D. (2000). Metrics on rna secondary structures. *Journal of Computational Biology*, 7(1-2):277–292.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: updates to the rna families database. *Nucleic Acids Research*, 43(D1):D130–D137.

Pasquali, S., Gan, H. H., and Schlick, T. (2005). Modular rna architecture revealed by computational analysis of existing pseudoknots and ribosomal rnas. *Nucleic Acids Research*, 33(4):1384–1398.

Quadrini, M., Culmone, R., and Merelli, E. (2017). Topological classification of rna structures via intersection graph. Accepted to 6th International Conference on the Theory and Practice of Natural Computing (TPNC).

Rastegari, B. and Condon, A. (2007). Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *Journal of computational biology*, 14(1):16–32.

Rastogi, T., Beattie, T. L., Olive, J. E., and Collins, R. A. (1996). A long-range pseudoknot is required for activity of the neurospora vs ribozyme. *The EMBO journal*, 15(11):2820.

Reidys, C. M., Huang, F., Andersen, J. E., Penner, R. C., Stadler, P. F., and Nebel, M. E. (2011). Topology and prediction of rna pseudoknots. *Bioinformatics*, 27(8):1076–1085.

Seki, H., Matsumura, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229.

Shapiro, B. A. (1988). An algorithm for comparing multiple rna secondary structures. *Computer applications in the biosciences: CABIOS*, 4(3):387–393.

Shapiro, B. A. and Zhang, K. (1990). Comparing multiple rna secondary structures using tree comparisons. *Bioinformatics*, 6(4):309–318.

Shen, L. X. and Jr, I. T. (1995). The structure of an rna pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *Journal of molecular biology*, 247(5):963–978.

Song, Y., Hua, L., Shapiro, B. A., and Wang, J. T. (2015). Effective alignment of rna pseudoknot structures using partition function posterior log-odds scores. *BMC Bioinformatics*, 16(1).

Staple, D. W. and Butcher, S. E. (2005). Pseudoknots: Rna structures with diverse functions. *PLOS Biology*, 3(6).

Taufer, M., Licon, A., Araiza, R., Mireles, D., Van Batenburg, F., Gultyaev, A. P., and Leung, M.-Y. (2008). Pseudobase++: an extension of pseudobase for easy searching, formatting and visualization of pseudoknots. *Nucleic acids research*, 37(suppl_1):D127–D135.

Vernizzi, G., Orland, H., and Zee, A. (2016). Classification and predictions of rna pseudoknots based on topological invariants. *Physical Review E*, 94(4):042410.