Topological Classification of RNA Structures via Intersection Graph

Michela Quadrini (orcid.org/0000-0003-0539-0290), Rosario Culmone (orcid.org/0000-0003-3333-0893), and Emanuela Merelli (orcid.org/0000-0002-1321-4134)

University of Camerino, via Madonna delle Carceri, Camerino, Italy
{michela.quadrini, rosario.culmone, emanuela.merelli}@unicam.it

Abstract. We introduce a new algebraic representation of RNA secondary structures as a composition of hairpins, considered as basic loops. Starting from it, we define an abstract algebraic representation and we propose a novel methodology to classify RNA structures based on two topological invariants, the genus and the crossing number. It takes advantage of the abstract representation to easily obtain two intersection graphs: one of the RNA molecule and another one of the relative shape. The edges cardinality of the former corresponds to the number of interactions among hairpins, whereas the edges cardinality of the latter is the crossing number of the shape associated to the molecule. The aforementioned crossing number together with the genus permits to define a more precise energy function than the standard one which is based on the genus only. Our methodology is validated over a subset of RNA structures extracted from Pseudobase++ database, and we classify them according to the two topological invariants.

Keywords: RNA Classification, Topological Invariants, RNA Algebraic Representation, Intersection Graph.

1 INTRODUCTION

Ribonucleic acid (RNA) is a single stranded molecule made of four different types of nucleotides, known as Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). Such single strand, referred to as *primary structure*, folds back on itself achieving *secondary* and *tertiary structures*. During such a process, called *folding process*, each nucleotide can interact at most with another one establishing a hydrogen bond performing Watson-Crick (G-C and A-U) and wobble (G-U) base pairs. The folding process can generate 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*, *internal loops*, *bulges*, *helixes* (or *stacks*) and *multi-loops*. Each one of them, generated when at least one base pair is formed, is a *loop*. Therefore, secondary structures are composed of loops. If no interaction among loops is present, the secondary structure is said to be *pseudoknots free*, as illustrated in Fig. 1 (A), otherwise it is called *pseudoknotted*,

as depicted in Fig. 1 (B). In this work, the phosphodiester bond, a chemical bound that links two consecutive nucleotides, is referred to as a **strong interaction** and is depicted by a black line, while the base pairs created during the folding process are called **weak interactions** and are illustrated by zig zag lines.

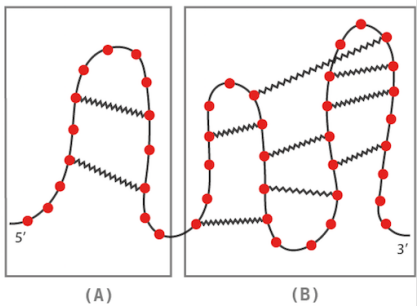

Fig. 1. RNA secondary structures

RNA molecules regulate a wide range of functions in biological systems. It has been recognized that in addition of being a carrier of genetic information, some RNA may also have enzymatic roles and may play a central part in the regulation of biological networks [5]. The pseudoknots, although it is known experimentally that they are fairly rare, usually impose some constraints on the sugar-phosphate backbone of the molecule. Their roles include forming the catalytic core of various ribozymes [13], self-splicing introns [1], and telomerase [17]. Additionally, they play critical roles in altering gene expression. For these reasons, starting from the primary structure of an RNA molecule, the prediction of the folding process is the main open problem of molecular biology [5]. Several deterministic and stochastic methods have been proposed for such prediction [2, 11]. Despite great progress, their overall success is limited, especially for long RNA molecules. Part of the difficulty lies in the prediction of RNA pseudoknots, which has been identified as an NP-complete problem [8]. Bon *et al.* [4] introduced a topological classification of RNA secondary structures with pseudoknots based on a topological invariant, the *genus*. Reidys *et al.* provided relevant contributions in the research area of combinatorial topology and developed several algorithms for predicting pseudoknots [14, 7]. Vernizzi *et al.* [18] added a new topological invariant, *the number of crossings*, to the aforementioned topological classification. Many different ways to represent RNA secondary structures are introduced in literature, such as the conventional diagram depicted in Fig. 2 (A), arc diagram illustrated in part (B) of Fig. 2, bracket representation and many others. The arc diagram representation can be regarded as a special case of the conventional diagram, where the vertices on a straight line (backbone) represent the nucleotides and base pairs are indicated using arcs.

In this work, we introduce a multiple context-free grammar that permits to associate a unique algebraic representation for each RNA molecule, both

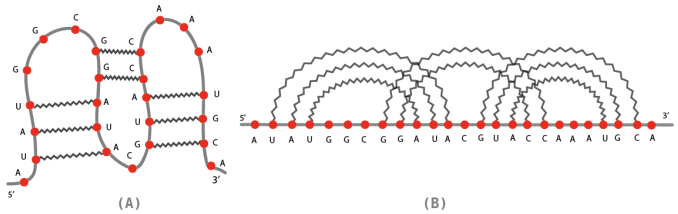

Fig. 2. Two different representations of RNA secondary structures

pseudoknot free and pseudoknotted. The main novelty of our approach, respect to the others present in the literature, is that we represent each RNA secondary structure as an algebraic composition of hairpins, considered as basic loops. Moreover, it permits to classify each RNA molecule in terms of genus and crossing number. Such crossing number and the genus, a non negative integer which depends only on the connectivity of the base pairs, are two topological invariants. They permit to improve the function for the energy calculation. Finally, a procedure, Pseudoknots Detection Procedure, is defined to identify the kind of pseudoknots of genus 1. In order to validate our methodology, we applied it to a subset of real RNA structures extracted from Pseudobase++ database, and we classified them according to their genus and crossing number.

The paper is organized as follows. In Section 2, we present a review of mathematical concepts necessary to understand the new proposed methodology, which is introduced in Section 3. The results are then commented in Section 4, whereas conclusions and future works are reported in Section 5.

2 MATHEMATICAL BACKGROUND

In this Section, some basic mathematical concepts will be introduced. The interested readers can refer to [12] for a complete treatment of topological invariants and to [10] for intersection graphs.

2.1 Topological Invariants

The global properties of RNA molecule are included in topological constraints encoded at the level of secondary structure. The topological invariants provide information regarding such constraints and, roughly speaking, they do not change under continuous stretching and bending of the topological space. The *genus* of an RNA molecule measures its complexity. Its geometrical interpretation is quite simple. In fact, the genus g of an arc diagram is the minimum number of handles that a sphere must have in order that each arc of the diagram can be illustrated without any crossing. An arc diagram that does not present any crossing can be drawn on a sphere. A graphical example is given at the top of Fig. 3. The sphere has no handles, so the genus associated to the structure is equal to 0. The arc diagram illustrated at the bottom right of Fig. 3 can be drawn without crossing

on a torus. Roughly speaking, the torus corresponds to a sphere with one handle and therefore the genus of the structure is 1.

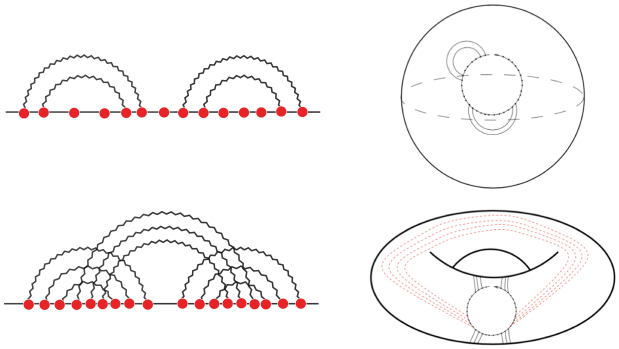

Fig. 3. Examples of the idea of genus

The genus permits to classify RNA secondary structures in equivalence classes; each class is determined by a value of genus g. In order to simplify the classification, we can observe that collapsing parallel arcs into one single arc and removing arcs which do not perform any cross, does not in fact change the genus value. This process determines the *shape* of the diagram. See Fig. 4 for an illustration of the process.

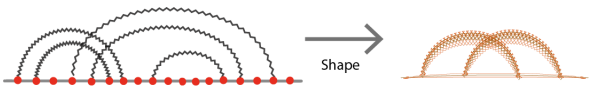

Fig. 4. The shape of a diagram

All the four types of RNA molecule with genus 1 are shown in Fig. 5.

Fig. 5. The four types of primitive pseudoknots with genus 1

A practical way for calculating the diagram genus consists in *fattening* the diagram, obtaining a double-line diagram, as illustrated in Fig. 6. Let P be the number of double lines (i.e., the number of base pairs) and let L be the number of closed loops, the genus of the diagram is the non negative integer defined by

$$g = \frac{P - L}{2} \ .$$

For instance, in Fig. 6 the diagram has 3 double lines and 1 closed loops.

Fig. 6. Steps to compute the genus of a structure

The genus has the property of being additive. Thus, for a structure comprised of two consecutive pseudoknots with genus g_1 and g_2 respectively, the genus of the whole structure is given by $g = g_1 + g_2$. For example, if the shape is composed of an H pseudoknot followed by a K pseudoknot, as illustrated in Fig. 7, each one has genus 1 and the genus of the whole structure is 2. In order to characterize the intrinsic complexity of a pseudoknot, the concepts of *irreducibility* and *nested* have been introduced. A shape is said to be *irreducible* if it cannot be disconnected by cutting the backbone. It is said to be *nested* if it can be removed by cutting the backbone twice, while the rest of the shape stays connected in a single component. The shape on the left of Fig. 7 is an example of a reducible one, whereas the motif on the right is irreducible.

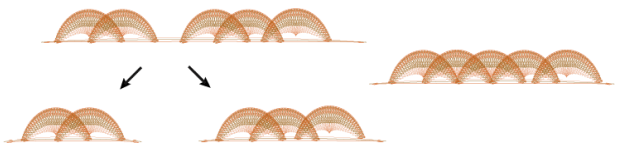

Fig. 7. A reducible shape (left) and an irreducible one (right)

Each arc diagram with genus greater than 0 is characterized by crossing arcs. Thus, the crossing arcs indicate the presence of at least one pseudoknot. If we take into account the four shapes of genus 1, introduced in Fig. 5, we can observe that they differ by the crossing number. Moreover, such crossing number and the genus do not uniquely identify the RNA shape. A simple example of this observation is given by the eight different pseudoknots with genus 2 and crossing number $\mathcal{N}_\mathcal{C}$ equals to 3 shown in Fig. 8.

The crossing number of a shape is a topological invariant and it has the property of being additive. In fact, if $\mathcal{D}$ is a reducible shape characterized by a sequence of two or more shapes, $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$, the crossing number, $\mathcal{N}_\mathcal{D}$, is given by the sum of the crossing number $\mathcal{N}_{\mathcal{D}_i}$ of each shape. Analogously, if $\mathcal{D}$ can be decomposed into nested parts $\mathcal{D}_i$, the crossing number, $\mathcal{N}_\mathcal{D}$, is given by the sum of the crossing number of each nested part. Thus, it is defined as follows:

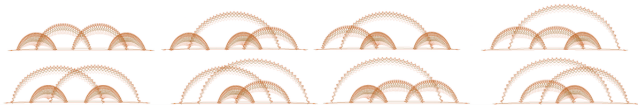

Fig. 8. The eight shapes with genus 2 and crossing number 3

1. Given an arc diagram $\mathcal{D}$, let $\mathcal{D} = \mathcal{D}_1 + \mathcal{D}_2 + \ldots \mathcal{D}_N$ be its decomposition in irreducible or nested parts $\mathcal{D}_i$;
2. For each diagram $\mathcal{D}_i$, we consider its shape $\mathcal{D}'_i$;
3. The crossing number $\mathcal{N}_C$ of $\mathcal{D}'$ is defined as the sum of the crossing number of each $\mathcal{D}'_i$.

2.2 Intersection Graph

Intersection graphs are relevant in both theoretical and applicative perspectives. In fact, they are able to provide several types of topological information about an arc diagram. For each arc diagram, its *intersection graph* is defined as follows:

1. each vertex corresponds to a loop of the diagram;
2. each edge corresponds to an interaction between two loops of the diagram.

An example of the intersection graph of the RNA structure illustrated in Fig. 2 is shown in Fig. 9.

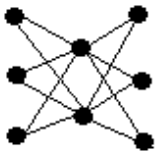

Fig. 9. The intersection graph of the RNA molecule shown in Fig. 2

3 MATERIAL AND METHODS

The topological classification of RNA secondary structures with pseudoknots, that we propose, is based on two topological invariants, *genus* and *crossing number*, and takes advantage of a new algebraic representation and of intersection graphs. To define the new representation it is necessary to introduce an operator able to model interactions among loops; it has been introduced in Section 3.1. Such operator is translated into a multiple context-free grammar, in Section 3.2. The procedures to obtain the intersection graph of an RNA molecule and the intersection graph of the relative shape are defined in Section 3.3, as well as the algorithm that permits to recognize the kind of pseudoknots of genus 1.

3.1 Operator to Model Interactions among Loops

In order to model RNA secondary structures, we define an operator *crossing*, $\bowtie_k$, able to model interactions among loops. The operator takes two arc diagrams and maps them into another one. It depends on a non integer parameter, k, which indicates that the resulting structure is obtained attaching the second arc diagram on the $k-$th nucleotides of the first one. According to the nature of RNA molecules, such operator is well-defined if each nucleotide of the resulting structure performs at most one weak interaction. 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. The new structure, obtained when k is equal to 0, is a concatenation between the two structures. In order to formally define the operator $\bowtie_k$, it is necessary to introduce new symbols, $\langle$, $\rangle$ and $\sharp$. Algebraically, each RNA secondary structure is identified by $(a_1^s, a_N^s)\langle\alpha\rangle$, where 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 (a_1^s, a_N^s). Each nucleotide that performs a weak interaction with another one, is marked by symbol $\sharp$, while the unpaired nucleotides are marked by ϵ. Formally, let S_1 and S_2 be two structures, where $S_1 = (a_1^s, a_N^s)\langle a_2^s \dots a_{N-1}^s\rangle$ and $S_2 = (b_1^s, b_M^s)\langle b_2^s \dots b_{M-1}^s\rangle$, the resulting structure, $S_1 \bowtie_k S_2$, is well defined if

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

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

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

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

3.2 Translating operator into MCFG

A context-free grammar is an inadequate formalism to describe arc diagrams with pseudoknots. It can be proved applying Ogdens Lemma [6]. As a consequence, a more expressive grammar is required. An appropriate choice is the so-called *Multiple Context-Free Grammar* (MCFG), introduced in [15]. 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 of nucleotides. The grammar 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 set of productions R is defined as follows:

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

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

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

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, $P \to L$ generates a hairpin. Note that a pseudoloop P is an RNA secondary structure without the head and the tail. Each loop L is a hairpin, $L \to x[\alpha^{+}]$, i.e., a Watson-Crick or a wobble base pair encloses a sequence of unpaired nucleotides, α^{+} .

Theorem 1. *Multiple context-free grammar G_{RNA}, introduced above, generates uniquely all RNA secondary structures.*

Proof. It is equivalent to prove that grammar G_{RNA} is not ambiguous. This property follows by the nature of the molecule, i.e., each nucleotide can perform at most one weak interaction and the primary structure is an ordered sequence of nucleotides. It is trivial to observe that the grammar is recursive to the right. This means that each production adds a hairpin starting from the end of the structure. Due to the biological constraint, the unambiguous property is guaranteed.

Theorem 2. *Each secondary structure can be uniquely decomposed in terms of a particular loop, i.e., hairpin.*

Proof. Each vertex which performs a weak interaction belongs to a unique hairpin. Since an unpaired nucleotide is either external or internal to a unique base pair the decomposition is unique.

3.3 From Arc Diagram to Intersection Graph

The multiple context-free grammar permits to associate a unique algebraic expression for each RNA secondary structure in terms of hairpins. Such algebraic expression contains each structural and biological information of the molecule. Obviously, two molecules having different backbones can be characterized by the same genus and same crossing number. We can observe that the two topological invariants cannot be influenced by the head and the tail of the structure, the unpaired nucleotides that characterize the loop or the number of nucleotides that two loops share. By removing the nucleotides, each weak interaction divides the

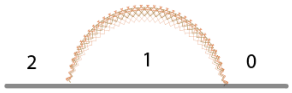

Fig. 10. Backbone components generated by an arc

backbone into three components, as illustrated in Fig. 10, which are enumerated from right to left starting from 0.
For each algebraic expression

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

the following abstract algebraic expression

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

is associated. Note that t is a non negative integer that represents the component of the backbone which the successive loop is attached to. Thus, operator $\bowtie_t$ is a bit different from the initial crossing operator: the initial one depends on nucleotides, whereas the second one depends on the backbone component. We decided to maintain the same symbol in order to not overload the notation. For each abstract structure, S', the intersection graph is associated by means of The Intersection Graph Procedure. It takes the input an abstract algebraic expression, that models the RNA molecule, and then returns an intersection graph as output. The core of the algorithm is based on the identification of the backbone component where the successive loop is attached on the identification of the numbers of crossing that the loop performs with the previous ones. It permits to detect the set of loops that cross each others. Another procedure, The Shape Intersection Graph Procedure, is defined over the intersection graph obtained by the previous algorithm. In fact, it takes the intersection graph of the molecule as input and returns the intersection graph of the relative shape as output, identifying and removing the edges which correspond to parallel arcs in the arc diagram. Finally, the kind of pseudoknots with genus 1 is defined by means of an additional procedure, The Pseudoknots Detection Procedure, based on edges of the shape intersection graph. The encoding of such procedures is omitted from this paper in order to not overload the readers with technicalities.

3.4 Example of Application

This methodology is applied to PKB10 molecule, extracted from Pseudobase++ database [16]. PKB10 is a tRNA-like structure 3′end pseudoknot of ononis yellow mosaic virus [9], which diagram, obtained from the database, is shown in Fig. 11. The algebraic expression of the structure is

$$S = \beta_1 x_1[\alpha_1] \bowtie_2 x_2[\alpha_2] \bowtie_1 x_3[\alpha_3] \bowtie_{12} x_4[\alpha_4] \bowtie_{11} x_5[\alpha_5] \bowtie_{10}$$
$$x_6[\alpha_6] \bowtie_9 x_7[\alpha_7] \bowtie_8 x_8[\alpha_8] \bowtie_7 x_9[\alpha_9]\beta_2$$

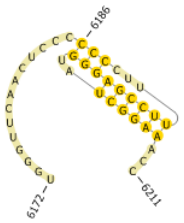

Fig. 11. The diagram of PKB10 obtained from Pseudobase++ database [16]

where $x_1 = x_2 = x_3 = (C, G)$, $x_4 = (A, U)$, $x_5 = (G, C)$, $x_6 = x_7 = (C, G)$, $x_8 = x_9 = (U, A)$, $\beta_1 = UGGGUUCAACUCCC$, $\alpha_1 = CUUUUCCGA$, $\alpha_2 = CCUUUUCCGAG$, $\alpha_3 = CCCUUUUCCGAGG$, $\alpha_4 = GGGUA$, $\alpha_5 = AGGGUAU$, $\alpha_6 = GAGGGUAUC$, $\alpha_7 = CGAGGGUAUCG$, $\alpha_8 = CCGAGGGUAUCGG$, $\alpha_9 = UCCGAGGGUAUCGGA$, $\beta_2 = ACC$.
The abstract algebraic expression is

$$S' = L \bowtie_2 L \bowtie_2 L \bowtie_2 \bowtie_3 L \bowtie_5 L \bowtie_7 L \bowtie_9 L \bowtie_{11} L$$

and the associated intersection graph, obtained applying The Intersection Graph Procedure, is $\mathcal{G} = (V, E)$ where $V = \{L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8\}$ and $E = \{(L_1, L_4), (L_1, L_5), (L_1, L_6), (L_1, L_7), (L_1, L_8), (L_2, L_4), (L_2, L_5), (L_2, L_6), (L_2, L_7), (L_2, L_8), (L_3, L_4), (L_3, L_5), (L_3, L_6), (L_3, L_7), (L_3, L_8)\}$. Applying The Shape Intersection Graph Procedure, we obtain $G' = (V', E')$, where $V' = \{L_1, L_2\}$ and $E' = \{(L_1, L_4)\}$. Finally, using The Pseudoknots Detection Proceduce we detect that the genus of the structure is 1 and we are in the presence of an H pseudoknot, thus $\mathcal{N}_{\mathcal{C}} = 1$.

4 RESULTS AND DISCUSSION

Our methodology permits to classify each RNA molecule in terms of *genus* and *crossing number* associated to its shape. Thus, for each equivalent class determined by the genus, it is possible to define a new classification. Such classification permits to define a more accurate energy function respect to the standard one. To test the methodology, we have analyzed a subset of real molecules extracted from Pseudobase++ database [16]. The molecules of the database are classified into groups in accord to different types of structure. We choose two groups, i.e., *HLIn* and *LL*, and the results of the analysis is shown in Table 1.
The results of the analyzed molecules correspond to the expected values. In fact, the genus of each molecule is 1 as well as the crossing number of the shape. These values are in accordance with the selected molecules, since each molecule is characterized by a H pseudoknot. Note that the same result can be obtained defining a procedure that associates for each molecule its shape, and applying an algorithm similar to The Intersection Graph Procedure. We propose the first approach because we believe that starting from the intersection graph of a

Table 1. Results of analysis

Molecule	Genus	Num of loops interactions	Crossing Num of Shape	Type of Pseudoknot
$PKB205$	1	16	1	$HLIn$
$PKB210$	1	63	1	$HLIn$
$PKB234$	1	63	1	$HLIn$
$PKB238$	1	80	1	$HLIn$
$PKB139$	1	24	1	LL
$PKB140$	1	68	1	LL
$PKB141$	1	27	1	LL
$PKB142$	1	35	1	LL
$PKB143$	1	35	1	LL
$PKB144$	1	32	1	LL
$PKB145$	1	30	1	LL
$PKB146$	1	25	1	LL
$PKB174$	1	112	1	LL
$PKB248$	1	18	1	LL
$PKB57$	1	28	1	LL

molecule, a new measure can be defined. Such measure will allow us to compute the distance between two RNA secondary structures in terms of interactions among loops.

5 CONCLUSIONS

In this work, a new algebraic representation of RNA secondary structures and an abstract representation have been defined. The former contains each structural and biological information of each molecule, the latter is obtained from the first one removing its primary structure. This simplification does not influence the two topological invariants, genus and crossing number, but easily allowed us to define three procedures. The Intersection Graph Procedure associates the intersection graph for each RNA molecule, while The Shape Intersection Graph Procedure, starting from the last structure, determines the intersection graph of The shape. Over the latter structure, the kind of pseudoknot of genus 1 is determined through The Pseudoknots Detection Procedure. Such methodology permits to classify each RNA molecule in terms of genus and crossing number associated to its shape. Thus, for each equivalent class determined by the genus, it is possible to define a new classification.

We have planned to improve the developed software that implements the whole methodology presented in this paper in order to analyze efficiently two database, Worldwide Protein Data Bank[3] and Pseudobase++ [16], with the scope of carrying out a more accurate topological classification than the one obtained by Bon et al. in [4]. A statistical study will be performed to detect the relations between genus and crossing number. Finally, the challenge will be

to define an algorithm for the RNA folding problem taking advantage of the classification obtained applying this proposed methodology.

References

1. Adams, P.L., Stahley, M.R., Gill, M.L., Kosen, A.B., Wang, J., Strobel, S.A.: Crystal structure of a group I intron splicing intermediate. Rna 10(12), 1867–1887 (2004)
2. Bellaousov, S., Mathews, D.H.: Probknot: fast prediction of rna secondary structure including pseudoknots. Rna 16(10), 1870–1880 (2010)
3. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide protein data bank. Nature Structural & Molecular Biology 10(12), 980–980 (2003)
4. Bon, M., Vernizzi, G., Orland, H., Zee, A.: Topological classification of RNA structures. Journal of molecular biology 379(4), 900–911 (2008)
5. Elliott, D., Ladomery, M.: Molecular biology of RNA. Oxford University Press (2017)
6. Harrison, M.A.: Introduction to formal language theory. Addison-Wesley Longman Publishing Co., Inc. (1978)
7. Huang, F.W., Reidys, C.M.: Topological language for RNA. Mathematical biosciences 282, 109–120 (2016)
8. Lyngsø, R.B., Pedersen, C.N.: RNA pseudoknot prediction in energy-based models. Journal of computational biology 7(3-4), 409–427 (2000)
9. Mans, R.M., Pleij, C.W., Bosch, L.: tRNA-like structures. Structure, function and evolutionary significance. Eur J Biochem 201(1) (1991)
10. McKee, T.A., McMorris, F.R.: Topics in intersection graph theory. SIAM (1999)
11. Metzler, D., Nebel, M.E.: Predicting RNA secondary structures with pseudoknots by MCMC sampling. Journal of mathematical biology 56(1), 161–181 (2008)
12. Munkres, J.R.: Analysis on manifolds. Westview Press (1997)
13. Rastogi, T., Beattie, T.L., Olive, J.E., Collins, R.A.: A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. The EMBO journal 15(11), 2820 (1996)
14. Reidys, C.M., Huang, F.W., Andersen, J.E., Penner, R.C., Stadler, P.F., Nebel, M.E.: Topology and prediction of RNA pseudoknots. Bioinformatics 27(8), 1076–1085 (2011)
15. Seki, H., Matsumura, T., Fujii, M., Kasami, T.: On multiple context-free grammars. Theoretical Computer Science 88(2), 191–229 (1991)
16. Taufer, M., Licon, A., Araiza, R., Mireles, D., Van Batenburg, F., Gultyaev, A.P., Leung, M.Y.: Pseudobase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic acids research 37(suppl_1), D127–D135 (2008)
17. Theimer, C.A., Blois, C.A., Feigon, J.: Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. Molecular cell 17(5), 671–682 (2005)
18. Vernizzi, G., Orland, H., Zee, A.: Classification and predictions of RNA pseudoknots based on topological invariants. Physical Review E 94(4), 042410 (2016)