Label Core for Understanding RNA Structures

Michela Quadrini[1,2][0000−0003−6203−6736], Emanuela Merelli[2][0000−0002−1321−4134], and Riccardo Piergallini[3][0000−0001−6203−6736]

[1] University of Padova, Department of Information Engineering, via Gradenigo 6/A, Padova, 35131
michela.quadrini@unipd.it

[2] University of Camerino, School of Science and Technology, Computer Science Department, Via Madonna delle Carceri, 9, Camerino (MC)
emanuela.merelli@unicam.it

[3] University of Camerino, School of Science and Technology, Mathematics Department, Via Madonna delle Carceri, 9, Camerino (MC)
riccardo.piergallini@unicam.it

Abstract. The RNA structure, the main predictor of biological function, is the result of the folding process. While the nucleotides in the RNA sequence rapidly coupled forming weak bonds, the spatial arrangement is a slow process. Although many computational approaches have been proposed to study the folding process of RNA, most of them do not consider the hierarchical aspect existing among the bonds.
In this work, we propose to collapse nucleotides and bonds underpinning the primary and secondary structure of RNA in a unique *label core* congruent with the spatial configuration. A label core is represented as a term of generalized context-free grammar properly defined to support RNA structural reduction and analysis.

Keywords: Generalized-context free grammar · label core grammar · RNA structure reduction.

1 Introduction

Ribonucleic acid (RNA) is a single stranded polymer, with a preferred $5'$-$3'$ direction, made of four different types of nucleotides, known as Adenine (`A`), Guanine (`G`), Cytosine (`C`) and Uracil (`U`). Each nucleotide is linked to the previous one by a phosphodiester bond, referred to as **strong bond**. Moreover, it can interact with at most another non-contiguous one, establishing a hydrogen bond, called **weak bond**. Such a process, known as **folding process**, induces complex *three-dimensional structure* (or shape). Such a shape is tied to its biological function. Discovering the relationships among nucleotides sequence, shape, and biological function has been considered one of the challenges in biology. RNAs play a variety of roles in cellular processes and are directly involved in the diseases for their ability to turn genes on and off. Disregarding the spatial configuration of the molecules and reducing nucleotides to dots, the molecule is abstracted in terms of *secondary structure*. Such an abstraction represents an intermediate

level between the sequence and the shape of the molecule. It is both tractable from a computational point of view and relevant from a biological perspective. As an example, under the action of antibiotics, many 16s ribosomal RNAs preserve the nucleotides sequence, but change the shape. Such changes are detected by secondary structure. This structure can be formalized as an *arc diagram*, where the nucleotides are represented by vertices on a straight line and the base pairs are drawn as arcs in the upper half-plane, see Fig. 1 for an example. An RNA secondary structure is said to be *pseudoknot-free* if the arc diagram does not present crossing among arcs, as illustrated in Fig. 1-**a**, otherwise it is called *pseudoknotted*, as depicted in Fig. 1-**b**.

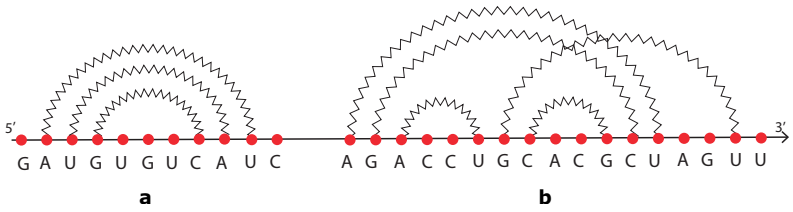

Fig. 1. An example of secondary structure represented as arc diagram. In **a**, the zigzagged arcs do not cross while, in **b**, pseudoknots are clearly visible as crossings of arcs

Here we propose an approach based on formal grammar to study the relationships between RNA structures and functions. In the literature different computational approaches have been exploited to face such a problem. Maestri and Merelli studied the relationships between RNA structure and functions by resorting to process calculi [9], while an algebraic language has been defined for representing and comparing secondary structures with arbitrary pseudoknots in polynomial time [10, 11]. Andersen *et al.* have exploited a combinatorial approach [1], and Yousef *et al.* have proposed an approach able to differentiate among species when the evolutionary distance increases [14]. Recently, Quadrini *at el.* have defined a context-free grammar to identify common substructures considering both the primary and secondary structures in [12].

In this work, we introduce the concept of **label core** of each molecule. For each secondary structure represented as an arc diagram, the label core is obtained by collapsing groups of consecutive unpaired nucleotides into into a single one, and consecutively parallel arcs are collapsed into a single one. Two arcs, identified by the pairs (i_1, j_1) and (i_2, j_2) are consecutively parallel if $i_1 = i_2 - 1 < j_2 = j_1 + 1$. In the literature, two similar concepts have been introduced, the core and shape [13]. The shape is determined by removing nucleotides and any arcs that do not cross, and collapsing the parallel arcs of the structure into single arcs, while the definition of the core preserve any arcs that do not cross.

The label core permits both the classification of the RNAs in terms of equivalent class of secondary structures as the core or shape and the consideration of the nucleotides sequence. Moreover, it is a particular arc diagram, where each vertex represents a finite ordered set of nucleotides rather than a single one, and it is unique for each arc diagram. As an example, the label core of the structures in Figure 1 is shown in Figure 2. To gain the ability of syntactically

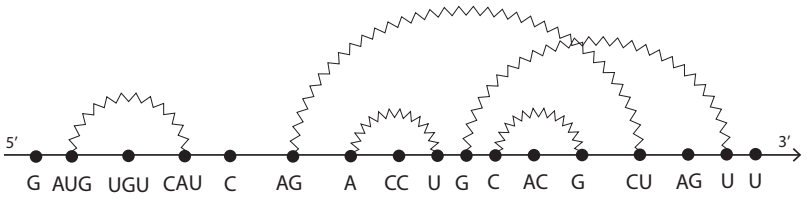

Fig. 2. The core of the structure shown in Figure 1

representing each RNA secondary structure, we define a generalized context-free grammars, **Label Core Grammar**, that uniquely represents the label core of each molecule as a string. Such a string consists of sequences of characters, which represent the nucleotides, enclosed in special symbols, $\langle$, $\rangle$, and equipped with natural numbers. Each sequence enclosed by the special characters represents a block of consecutive nucleotides. Moreover, if the special symbol is followed by a number h, it means that the consider block of nucleotides performs weak bonds with another block enclosing other h blocks. This representation of the structure provides several advantages, among which it allows us to analyze substructures taking into account the nucleotides synthesis and the local folding.

This paper is organized as follows. In Section 2, we give the necessary background on formal grammars. In Section 3, we define a grammar, called *Label Core Grammar*, that describe a language corresponding the label core of RNA secondary structure. In Section 4, we use the words generated by the language to analyze the RNA structures in order to understand the relations between the sequence and the its secondary structure. In particular, we use two different measures, CG-skew and AU-skew, to analyze the structures. The paper ends with some conclusions and future perspective, Section 5.

2 Generalized Context-free Grammars

In this section, we give the background on formal grammar needed for this work. Context-free grammars are not expressive enough to model RNA pseudoknotted structures, as formally proved by Ogden's Lemma, a generalization of the pumping lemma for context-free languages [5]. A natural extension of context-free grammars is Generalized context-free grammars (GCFG), in which

nonterminal symbols may generate n-tuples of strings that are obtained by merging m-tuples of other strings by using general rewriting partial functions. Such formalism has same the generative capacity of Type 0 grammars. For a more detailed description, the interested reader can refer to [7].

Definition 1 (Generalized context-free grammar (GCFG)). *A GCFG $\mathcal{G}$ is a $5-$tuple (N, T, F, P, S) where:*

- *N is a finite set of nonterminal symbols;*
- *T is a finite alphabet, also called set of terminal symbols;*
- *F is a finite set of merging partial functions;*
- *P is a finite set of rewriting rules; and*
- *$S \in N$ is the start symbol.*

Let $\mathbb{T}$ be the set of all n-tuples of strings in T^*, with $n \geq 1$. Each rewriting rule in P has the form

$$A_0 \to f[A_1, A_2, \ldots, A_q]$$

where:

- *$A_i \in N$, for all $i = 0, \ldots, q$;*
- *$f\colon \underbrace{\mathbb{T} \times \cdots \times \mathbb{T}}_{q \text{ times}} \mapsto \mathbb{T}$ is a partial function in the set F, it produces an n-tuple of strings on the alphabet T and its arguments are q n_i-tuples for $i = 1 \ldots q$ of strings on the same alphabet.*

If $q = 0$, i.e., $f \in \mathbb{T}$, then the rewriting rule is called terminating *rule.*

Definition 2. *Let $\mathcal{G} = (N, T, F, P, S)$ be a GCFG. For each nonterminal $A \in N$, the set $\mathcal{L}_\mathcal{G}(A)$ is defined as the smallest set that satisfies the following two conditions:*

- *if a rule $A \to \theta$ of P is a terminating rule, then θ belongs to the language accepted by A, denoted $\theta \in L_\mathcal{G}(A)$;*
- *if $\theta_i \in L_\mathcal{G}(A_i)$ for $i = 1, \ldots, q$, $A \to f[A_1, A_2, \ldots, A_q] \in P$ and $f[\theta_1, \theta_2, \ldots, \theta_q]$ is defined, then $f[\theta_1, \theta_2, \ldots, \theta_q] \in \mathcal{L}_\mathcal{G}(A)$.*

The language generated by $\mathcal{G}$, denoted by $\mathcal{L}_\mathcal{G}$, is $\mathcal{L}_\mathcal{G}(S)$.

Definition 3 (Derivation tree). *A* derivation tree *of a GCFG $\mathcal{G} = (N, \mathcal{T}, \mathcal{F}, P, S)$ is defined as follows:*

- *for a terminating rule $A \to \theta$, the tree whose root is labelled with A and has only one child labelled with θ is a derivation tree of θ;*
- *if T_i, $i = 1, \ldots, q$, is a derivation tree of $\theta_i \in \mathbb{T}$ whose root is labelled with A_i, $A \to f[A_1, A_2, \ldots, A_q]$ is in P and $f[\theta_1, \theta_2, \ldots, \theta_q]$ is defined, then a tree such that*
  - *the root is labelled with A;*
  - *the root has q children;*
  - *for all $i = 1, \ldots, q$ the subtree rooted at the i^{th} child is isomorphic to T_i is a derivation tree of $f[\theta_1, \theta_2, \ldots, \theta_q]$;*
- *there is no other derivation tree.*

3 Label Core Grammar

RNA molecules can be formally described in a hierarchical way. An RNA is a finite sequence, called *primary structure*, composed of four different nucleotides, adenine (`A`), guanine (`G`), cytosine (`C`) and uracil (`U`), that is represented as a word over an alphabet. By folding back onto itself, establishing weak bonds, an RNA molecule forms a complex shape, also called *tertiary structures*. The tertiary structure of a molecule represents the spatial configuration of the molecule, based on the relative position of atoms, obtained by NMR spectroscopy or X-Ray diffraction [8, 6]. Moreover, it can be abstracted in terms of *secondary structure*, which consists of a nucleotides sequence and a set of disjoint base pairs that correspond to the weak bonds. As mentioned in Section 1, it can be represented as an arc diagram by putting the dots representing the nucleotides and the strong bonds (the backbone of the molecule) on the x-axis and realizing the weak bonds as semi-circular zigzag arcs in the upper half-plane. Formally, the arc diagram is a labeled graph over the vertex set $[\ell] = \{1, \ldots, \ell\}$, in which each vertex has degree ≤ 3, and the edges are all the segments. We introduce the concept of **label core**. For each RNA secondary structure represented as an arc diagram, it label core is obtained by collapsing groups of consecutive unpaired nucleotides into a single one, while consecutively parallel arcs are collapsed into a single one. Two arcs, identified by the pairs (i_1, j_1) and (i_2, j_2) are consecutively parallel if $i_1 = i_2 - 1 < j_2 = j_1 + 1$. Such a core is unique for each arc diagram. Moreover, it can be considered as a particular arc diagram, where each vertex represents a finite string over the alphabet $\mathcal{A}_{RNA}$ rather than the single character, as shown in Figure 2. To gain the ability to syntactically represent the label core of each RNA secondary structure, we define a generalized context-free grammars. Each term of the generated language uniquely represents a label core. It is a word formed by sequences of characters enclosed by between "$\langle$" and "$\rangle$", which can be followed by a natural number h. Each subsequence represents a set of consecutive nucleotides and each natural number h indicates that the considered sequence is linked with a previousone overtaking h of them. In order words, the two linked elements are separated by other h elements. For brevity, as used in [3] by Giegerich *et al.* for tree grammars, we add a *lexical level* to the generalized context-free grammar concept, allowing strings in place of single symbols over $\mathcal{A}$. Let $\mathcal{A}_{RNA} = \{\mathtt{A}, \mathtt{U}, \mathtt{G}, \mathtt{C}\}$ be the alphabet of RNA. The **Label Core Grammar** is a $\mathcal{C}_{RNA} = (N, T, P, S, F)$, where $N = \{S, S'\}$, $T = \mathcal{A}_{RNA} \cup \{\langle\ ,\ \rangle\} \cup \{(\ ,\ )\} \cup \{1, \ldots, N\}$, $F = \{\overline{f}\}$ and the set of rewriting rules P is defined as follows:

$$
\begin{array}{lll}
S & ::= \epsilon & \textit{empty structure} \\
 & \mid S' & \textit{non-empty structure} \\
S' & ::= S\langle b\rangle & \textit{primary structure} \\
 & \mid f(S', (h, b)) & \textit{secondary structure}
\end{array}
$$

with $b \in \mathcal{A}^{+}_{RNA}$ and $h \in \{1, \ldots, N\}$. The partial function f depends on $M = \nu(S)$ and $\psi(S')$ defined as follows, respectively

$$\nu(S) = \begin{cases} \emptyset & \text{if } S = \epsilon \\ \{\ell\} \cup \{\ell - h - 1\} \cup \nu(S') & \text{if } S = S'\langle b_h \rangle \\ \nu(S') & \text{if } S = S'\langle b \rangle \end{cases}$$

and

$$\psi(S') = \begin{cases} 1 & \text{if } S' = b,\ S' = b_h \\ 1 + \psi(S) & \text{if } S' = S\langle b \rangle,\ S' = S\langle b_h \rangle \end{cases}$$

The partial function f is defined as follows

$$\bar{f}(S', (h, b)) = \begin{cases} S'\langle b_h \rangle & \text{if } h < \psi(S') = \ell \text{ and } \ell - \bar{h} - 1 \notin M \text{ and} \\ & \quad |b_h| = |S\langle \ell - h - 1 \rangle| \\ \perp & \text{otherwise} \end{cases}$$

The partial function f, when it is defined, depend on an integer h that does not correspond to the nucleotides included into the weak bond, but it represents the number of components composed of collapsed nucleotides or weak bonds that are overtaken by the considered weak bond. Moreover, such weak bond links an ordered set of nucleotides with another one. For this reason, the partial function f depends on a further condition, i.e., the cardinality of the two linked stacks must be the same. We describe, step by step, the unique way to represent the core of the pseudoknotted component represented in Fig. 1 using the productions of the grammar. In other words, we illustrate a procedure analogous to a deterministic recursive descent parser for $\mathcal{C}_{RNA}$ grammar. The result is that we can determine uniquely a derivation tree. The first step consists in recognizing the vertex of the structure applying the production $S' := S\langle b \rangle$ of the grammar. In this case, $b = \texttt{U}$. The second step is to apply the production $S' := f(S', (h, \langle b \rangle))$ to represent the bond determined by the second to last stack of nucleotides, made by only one nucleotide, U. In this case, the value of h is 5, because there are 5 stacks between the two considered stacks of nucleotides. The third step consists of the formalization of the stack AG, using production $S' := S\langle b \rangle$. Proceeding this way, we obtain the string

$$\langle \texttt{G} \rangle \langle \texttt{AUG} \rangle \langle \texttt{UGU} \rangle \langle \texttt{CAU} \rangle_1 \langle \texttt{C} \rangle \langle \texttt{AG} \rangle \langle \texttt{A} \rangle \langle \texttt{CC} \rangle \langle \texttt{U} \rangle_1 \langle \texttt{G} \rangle \langle \texttt{C} \rangle \langle \texttt{AC} \rangle \langle \texttt{G} \rangle_1 \langle \texttt{CU} \rangle_7 \langle \texttt{AG} \rangle \langle \texttt{U} \rangle_5 \langle \texttt{U} \rangle$$

Such scheme works in general to give a unique algebraic expression of each motif of RNA secondary structure. This observations yields the following:

Theorem 1. *The Label Core Grammar, $\mathcal{C}_{RNA}$, generates uniquely all RNA label core.*

4 RNA Structural Analysis

The grammar described in the previous section generates a language whose words uniquely represents the label core of RNA secondary structures. Here we use two

different measures, *CG-skew* (cythosine-guanine ratio) and *UG-skew* (adenine-uracil ratio) for analyzing the words of the language. The two measures was introduced to identify particular sequence in a genome at which replication starts and ends [4].

Definition 4 (CG-skew). *The CG-skew of a strand is a measure between* -1 *and* 1 *for dominance in occurrence of cytosine compared to guanine:*

$$CG\text{-}skew = \frac{C - G}{C + G}$$

where G *and* C *represent the frequency of occurrence in the considered sequence.*

Definition 5 (AU-skew). *The AU-skew of a strand is a measure between* -1 *and* 1 *for dominance in occurrence of cytosine compared to guanine:*

$$AU\text{-}skew = \frac{A - U}{A + U}$$

where A *and* U *represent the frequency of occurrence in the considered sequence.*

To understand the nucleotides sequence contribution in the weak bonds formation, we analyze the word generated by the grammar using the two measures. Since such strings contain numbers, which indicate the formation of weak bonds between the associated block and another previous one, we are able to analyze substructures taking into account the nucleotides synthesis. It is equivalent to analyze subwords of the form $\omega[1 : i]$, i.e., subwords from the first character to the i-th one that corresponds a substructure formed by the first i nucleotides. The value of i depends on the presence of weak bonds. As an example, we consider the word generated by the grammar $\mathcal{C}_{RNA}$ that identifies the label core of the structure illustrated in Fig. 1

$$\langle \mathtt{G} \rangle \langle \mathtt{AUG} \rangle \langle \mathtt{UGU} \rangle \langle \mathtt{CAU} \rangle_1 \langle \mathtt{C} \rangle \langle \mathtt{AG} \rangle \langle \mathtt{A} \rangle \langle \mathtt{CC} \rangle \langle \mathtt{U} \rangle_1 \langle \mathtt{G} \rangle \langle \mathtt{C} \rangle \langle \mathtt{AC} \rangle \langle \mathtt{G} \rangle_1 \langle \mathtt{CU} \rangle_7 \langle \mathtt{AG} \rangle \langle \mathtt{U} \rangle_5 \langle \mathtt{U} \rangle$$

as determined in Section 3. For this term, we analyze the six subwords, i.e., GAUGUGUCAU, GAUGUGUCAUCAGACCU, GAUGUGUCAUCAGACCUGCACG, GAUGUGUCAUCAGACCUGCACGCU, GAUGUGUCAUCAGACCUGCACGCUAGU, and GAUGUGUCAUCAGACCUGCACGCUU. As a consequence, we are able to analyze substructures taking into account the nucleotides synthesis and the local folding. Following this schemas, we have selected two different sets of molecules from RNA STRAND database [2].

A group (Group A) is composed of 13 molecules of 5s ribosomal RNA that we have download from RNAStrand Databases selecting all molecules of this class validated by MNR or X-ray, the other one (Group B) is composed of 31 molecules of 16S. The relative lists are reported in Appendix A. Moreover, each of them has been validated by MNR or X-ray. From the analysis, it follows that every time the first block of weak bonds is created, there is an imbalance of the presence of nucleotides of cytosine with respect to guanines which is approximately double of the whole sequence, while no evidence for the imbalance of the presence of nucleotides of uracil with respect to adenines. Although this result has no statistical value, it represents a starting point for further analysis.

5 Conclusion

In this work, we have introduced the concept of *label core* of RNA molecules. Considering an RNA secondary structure represented as an arc diagram, the label core is determined by collapsing consecutive nucleotides and arcs parallel arcs. We have defined a generalized context-free grammar, called *Label Core Grammar*, to represent syntactically the label core of a molecule. The words generated by the grammar consists of a sequence of characters, which represent nucleotides, equipped with the information related to weak bonds. Such representation has favored the analysis of RNA substructures taking into account the nucleotides synthesis and local folding. In particular, we have applied two different measures, CG-skew and AU-skew, over two different sets of molecules, and we consider the results are as a starting point for further analysis. Our long-term goal is to use the information on nucleotides sequence contribution, combined with environmental conditions, to predict the formation of intra and inter molecule weak bonds. In particular, we intend to define stochastic grammar to predict RNA secondary structures. Moreover, the fact that we take into account the sequence of characters is divided into block allows us to identify the unpaired nucleotides that can interact with other RNAs. We also intend to use this theoretical result to represent the RNA primary and secondary structures as a vector space to apply machine learning techniques with the aim to identify targets of RNA-RNA interactions. Understand the targets and how the mechanism of RNA-RNA interactions is a fundamental task to link the structure and biological unction, both healthy and sick cells.

References

1. Andersen, J.E., Huang, F.W., Penner, R., Reidys, C.: Topology of RNA-RNA interaction structures. Journal of Computational Biology **7**(19), 928–943 (2012)
2. Andronescu, M., Bereg, V., Hoos, H.H., Condon, A.: RNA STRAND: the RNA secondary structure and statistical analysis database. BMC bioinformatics **9**(1), 340 (2008)
3. Giegerich, R., Steffen, P.: Implementing Algebraic Dynamic Programming in the Functional and the Imperative Programming Paradigm. In: Boiten, E., Möller, B. (eds.) Mathematics of Program Construction. MPC 2002. Lecture Notes in Computer Science, vol. 2386, pp. 1–20. Springer, Berlin, Heidelberg (2002)
4. Grigoriev, A.: Analyzing genomes with cumulative skew diagrams. Nucleic acids research **26**(10), 2286–2290 (1998)
5. Harrison, M.A.: Introduction to formal language theory. Addison-Wesley Longman Publishing Co., Inc. (1978)
6. Holbrook, S.R., Kim, S.H.: RNA crystallography. Biopolymers: Original Research on Biomolecules **44**(1), 3–21 (1997)
7. Kasami, T., Seki, H., Fujii, M.: Generalized context-free grammars and multiple context-free grammars. Systems and computers in Japan **20**(7), 43–52 (1989)
8. Kjems, J., Egebjerg, J.: Modern methods for probing RNA structure. Current opinion in biotechnology **9**(1), 59–65 (1998)

9. Maestri, S., Merelli, E.: Process calculi may reveal the equivalence lying at the heart of RNA and proteins. Scientific report **9**(559) (2019)
10. Quadrini, M., Tesei, L., Merelli, E.: An algebraic language for RNA pseudoknots comparison. BMC Bioinformatics **20**(4), 161 (2019)
11. Quadrini, M., Merelli, E.: Loop-loop interaction metrics on rna secondary structures with pseudoknots. In: BIOINFORMATICS. pp. 29–37 (2018)
12. Quadrini, M., Merelli, E., Piergallini, R.: Loop Grammars to Identify RNA Structural Patterns. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS. pp. 302–309. SciTePress (2019)
13. Reidys, C.: Combinatorial Computational Biology of RNA. Springer (2011)
14. Yousef, M., Khalifa, W., lhan Erkin Acar, Allmer, J.: MicroRNA categorization using sequence motifs and k-mers. BMC Bioinformatics **18**(1), e170 (2017)

Appendix

In this appendix we list the molecules of the two group that we have download from RNA Strand database.

List of molecules of Group A.

PDB_00004	PDB_00030	PDB_00048	PDB_00158	PDB_00249
PDB_00250	PDB_00252	PDB_00253	PDB_00288	PDB_00339
PDB_00570	PDB_00752	PDB_01100		

List of molecules of Group B.

PDB_00069	PDB_00227	PDB_00228	PDB_00409	PDB_00456
PDB_00457	PDB_00458	PDB_00459	PDB_00478	PDB_00589
PDB_00645	PDB_00703	PDB_00769	PDB_00771	PDB_00773
PDB_00775	PDB_00777	PDB_00791	PDB_00793	PDB_00933
PDB_00935	PDB_01107	PDB_01220	PDB_01222	PDB_01224
PDB_01226	PDB_01228	PDB_01230	PDB_01232	PDB_01234
PDB_01241	PDB_01243	PDB_01245	PDB_01247	PDB_01282
PDB_01284				