Towards a Shape Language for Interpreting RNA Folding*

Adane Letta Mamuye
School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
adaneletta.mamuye@unicam.it

Emanuela Merelli
School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
emanuela.merelli@unicam.it

Luca Tesei
School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
luca.tesei@unicam.it

ABSTRACT

In this paper we introduce a shape language for representing RNA secondary structures in a non-standard, non-linear way. The main motivation is to propose a new interpretation of RNA folding as a self-adaptability process, within the $S[B]$ paradigm, towards a minimum free energy configuration. An RNA secondary structure is decomposed first by distinguishing between pseudoknot free and pseudonotted sub-structures. For pseudoknot free sub-structures a proper formal language is defined. To address the representation of pseudoknotted sub-structures the crucial aspects of RNA irreducible shapes and their associated automatic groups are introduced.

Categories and Subject Descriptors

I.1.3 [**Symbolic and Algebraic Manipulation**]: Languages and Systems—*Special-purpose algebraic systems*; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Formal languages, Algorithms

Keywords

RNA secondary structures, RNA shapes, $S[B]$ paradigm, formal languages, group theory, topological data analysis.

1. INTRODUCTION

Ribonucleic acid (RNA) is a linear molecule composed by four different nucleoacids: adenine (A), guanine (G), cytosine (C) and uracil (U). Such molecules perform a variety of biological functions inside the cell [11]. Moreover, they play a central role in protein synthesis, enzymatic catalysis and in the regulation of genome organization and gene expression. RNA structures are formed by folding a sequence of nucleotides, the so called primary structure. In 2-dimension, the folding process generates an RNA secondary structure that corresponds to a specific shape, also called its structural motif. Each structure is responsible of the function itself.

The secondary structure can be divided into sub-structures that are either *pseudoknot free* or *pseudoknotted*. In molecular biology, the prediction of RNA pseudoknotted structures is an important task for supporting disease diagnosis. In the last decade, several comparative sequence analysis and dynamic programming methods have been used to predict the right configuration of RNA secondary structures [4, 14]. In particular, Reidys et al. provided relevant contributions in the research area of combinatorial topology [13, 5]. They introduced the concept of RNA *irreducible-shapes* as the building blocks to capture the recursive structures within the RNA multiple context-free grammar representation [1]. Despite the excellent results, accurate prediction is still an ongoing challenge in computational biology, being an NP-complete problem [7].

Since an RNA molecule exhibits an auto-regulative mechanism similar to the adaptability process of complex systems [8], a new way of modeling the prediction of an RNA secondary structure can be investigated. The idea is to represent an RNA configuration by an algebraic structure that codes simultaneously the RNA functional behavior and its structural motif. Each configuration is a particular instance of a *fold space* that evolves until the configuration with minimum free energy is reached. This process can be represented by successive transformations of the algebraic structure. The result is a new concept of ′shape′ (the algebraic structure) that delivers both local and global information, i.e. a representation of the current secondary structure and the corresponding free energy.

Even though many bio-inspired methods have been presented in the literature for describing complex systems, none of them provides a language characterization that allows one to represent a system as an expression that simultaneously

*We acknowledge the financial support of the Future and Emerging Technologies (FET) program within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FP7 FET-Proactive Call 8 Dynamics Multi-level Complex Systems (DyMCS), Grant Agreement TOPDRIM, Number FP7-ICT-318121. We thank Christian Reidys for useful discussions about this topic.

[1] To avoid ambiguity in using the term ′shape′ we will use, throughout the paper, ′irreducible-shape′ to refer to Reidys' shapes, which are actually graphs.

.

shows the local and global information in a unique contextual semantics. We introduce a *shape language* as an RNA domain specific language whose syntax and semantics allow us to represent a shape and its behavior as a self-adaptive system. A ′shape′, at any time, expresses both the *stability* (the type and the number of secondary sub-structures) and the *quality* (the amount of free energy of the RNA configuration) of the RNA structure. The reduction mechanism will allow us to simulate the evolution of a shape as a reachability problem throughout the process of RNA adaptation. We will embed the RNA shape language into the $S[B]$ paradigm, a framework for modeling complex systems. $S[B]$ was introduced by the authors as a two-level entangled model, namely the S *global* or *structural* components and the B *local* or *behavioral* level [9].

We formalize the loops of pseudoknot free structures as members of specific formal languages and we use topologization (a procedure to transform a data set into a topological object) for deriving topological shapes of *genus one* from RNA pseudoknotted strucures. Then, we associate to each of them an *automatic group* with the relative finite-state automata [2], which allow us to determine if a given expression is in canonical form. Basing on these mathematical tools, the part of the shape language for pseudoknotted secondary structures can be defined.

2. FORMALIZATION OF RNA LOOPS

Along the RNA linear molecule, each nucleotide of the primary structure can form a base pair by interacting with one other nucleotide, forming Watson-Crick bases pairs (C-G and A-U) or Wobble base pairs (G-U). The RNA secondary structure is due to the creation of other pairings that generate loops (structural elements) namely *hairpins*, *bulges*, *internal loops*, *multi-branched loops* and *helixes* (or *stacks*). Moreover, an RNA secondary structure can be of two types: *pseudoknot free* or *pseudoknotted*. As illustrated in Figure 1 (left picture), a pseudoknot free structure is composed of a set of *non-crossing-serial* interactions, while a pseudoknotted one is formed by *crossing-serial* interactions (right picture). The five structural elements (loops) can be used repeatedly, in various combinations, to form different RNA pseudoknot free structures. The two types of structures can also be represented by diagrams as shown in Figure 2. The set of vertices are the nucleotides and the set of arcs are the basing pairings. A pseudoknot free structure is a diagram without crossing arcs whereas diagram with crossings represents a pseudoknotted structure.

2.1 Loops of RNA Pseudoknot Free Structures

Each loop of a pseudoknot free structure can be represented as a word of a language. Given a finite set of symbols Σ, called alphabet, a language is any subset of strings formed by the elements of Σ. Let $\Sigma = \{A, G, C, U\} \cup \{(,)\}$ be the alphabet of RNA, where A, G, C, U represent the four nucleotides adenine, guanine, cytosine and uracil, respectively, and the brackets are used to enclose loops. In the following, to associate a language to each loop, we use $a, b, \ldots$ to denote any of the nucleotides in Σ. The corresponding complements, w.r.t. Watson-Crick or Wobble base pairs, are denoted by $\overline{a}, \overline{b}, \ldots$. For example, an *helix* loop made by two base pairs is represented as $\{ab\overline{b}\overline{a} \mid a, b \in \Sigma\}$. For loops with unpaired nucleotides (which might be in a hairpin loop, internal loop, bulge loop or multi-branched loop), we use

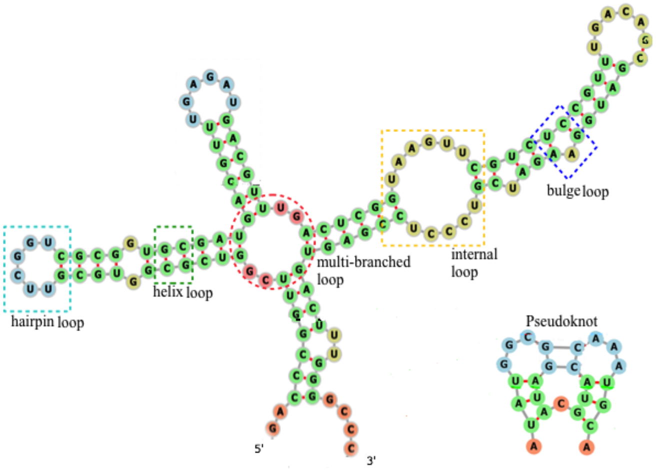
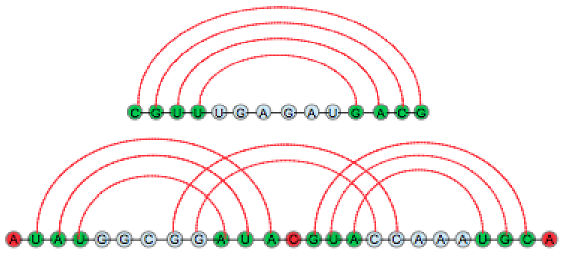


Figure 1: On the left, a pseudoknot free structure. On the right, a pseudoknotted structure. These structures have been drawn with the *Forna* online RNA drawing web server [6].

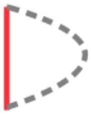


C G U U U G A G A U G A C G
A U A U G G C G G A U A C G U A C C A A A U G C A

Figure 2: On the left, the diagram of the pseudoknot free structure of Figure 1. On the right, the diagram of the pseudoknotted structure of Figure 1.

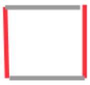
strings from the alphabet $\Sigma_\bullet = \{A_\bullet, G_\bullet, C_\bullet, U_\bullet\}$. Accordingly, the five languages corresponding to the RNA loops are defined in the following, where the * iteration operator applied on an alphabet denotes the set of all finite strings that can be formed with the symbols of the alphabet, together with the empty string.

Hairpin loop



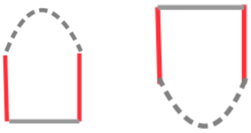
$$L_{\text{hairpin}} = \{(a\,\alpha\,\overline{a}) \mid a \in \Sigma, \alpha \in \Sigma_\bullet^*\}$$

Helix loop



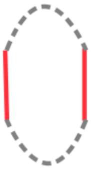
$$L_{\text{helix}} = \{(a_1 a_2 \cdots a_n \overline{a_n}\,\overline{a_{n-1}} \cdots \overline{a_1}) \mid n > 1, a_i \in \Sigma, 1 \leq i \leq n\}$$

Bulge loop



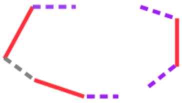
$$L_{\text{bulge}} = \{(ab\overline{b}\,\alpha\,\overline{a}) \mid a, b \in \Sigma, \alpha \in \Sigma_\bullet^*\} \cup \{(a\,\alpha\,b\overline{b}\overline{a}) \mid a, b \in \Sigma, \alpha \in \Sigma_\bullet^*\}$$

Internal loop



$$L_{\text{inner}} = \{(a\,\alpha\,b\overline{b}\,\beta\,\overline{a}) \mid a, b \in \Sigma, \alpha, \beta \in \Sigma_\bullet^*\}$$

Multi-branched loop



$$L_{\text{multi}} = \{(a_1\,\alpha_1\,a_2\overline{a_2}\,\alpha_2\,\cdots\,\alpha_{n-1}\,a_n\overline{a_n}\,\alpha_n\,\overline{a_1}) \mid n > 1, a_i \in \Sigma, \alpha_i \in \Sigma_\bullet^*, 1 \leq i \leq n\}$$

A pseudoknot free secondary structure can be generated by concatenating words $\alpha\beta_1\beta_2\cdots\beta_n\delta$ such that for all i, $1 \leq i \leq n$, β_i is in one of the languages defined above and $\alpha, \delta \in \Sigma_\bullet^*$ are possible initial and final non-loops. The order in which the loops are presented in the word is arbitrary. For instance, a "left-to-right" order (following the sequence of nucleotides, i.e. from the 5'-end to the 3'-end) can be used. Using this order, the pseudoknot free structure in the right part of Figure 1 can be represented by the following string: $G_\bullet A_\bullet(CCGG)_{\text{helix}}(CGUU_\bullet U_\bullet G)_{\text{bulge}}(GUAC)_{\text{helix}}$ $(UC_\bullet G_\bullet GUGUU_\bullet G_\bullet AUG)_{\text{multi}}(UCGCGCGA)_{\text{helix}}$ $(GG_\bullet UGG_\bullet U)_{\text{inner}}\cdots$ $(UU_\bullet G_\bullet A_\bullet C_\bullet A_\bullet G_\bullet C_\bullet G)_{\text{hairpin}}G_\bullet C_\bullet C_\bullet C_\bullet$[2].

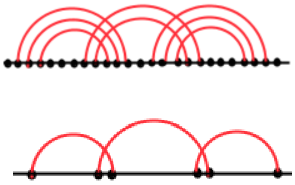


Figure 3: On the left, the RNA diagram of the pseudoknotted structure in Figure 2 (right). On the right, the corresponding RNA shape in which parallel arcs have been substituted with only one arc and unpaired vertices of the backbone have been discarded.

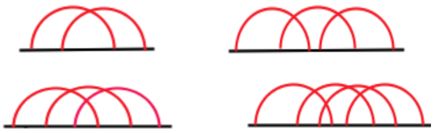


Figure 4: The four irreducible-shapes (H, K, L and M) of genus $g = 1$ are presented from left to right, respectively.

2.2 Irreducible Shapes of RNA Pseudoknotted Structures

The strategy that we use for formulating a language of RNA pseudoknots is to extract RNA shapes from RNA diagrams [5]. This process works by replacing, in RNA diagrams, parallel arcs with one single arc and by discarding all unpaired vertices of the backbone and all arcs of length one, as shown in Figure 3. This shape is tailored to preserve the topological information of the associated RNA diagram.

Topology is a branch of mathematics that studies shapes and properties of shapes that are invariant under homeomorphisms [3]. A topological invariant is a property of a topological space which is invariant under any deformation that does not produce rips. The *genus* is one of the topological invariants, which counts the number of holes in an orientable surface. For instance, a torus has a single hole, thus its genus g is equal to 1.

For shapes of fixed topological genus there exist only finitely many *irreducible-shapes*. A shape is called irreducible if it cannot be broken into two disconnected pieces by cutting a single horizontal edge [1]. If the genus is $g = 1$, it can be shown that there are exactly four irreducible-shapes, called H, K, L and M [5], which are depicted in Figure 4. Any pseudoknotted structure can be associated to a topological space. Given a dimension γ, the structure is called γ-*structure* if it can be constructed by concatenating and nesting irreducible-shapes of genus $g \leq \gamma$ [13, 5]. Thus any 1-*structure*, of arbitrary topological genus, can be obtained by concatenating and nesting the four irreducible-shapes H, K, L and M. For instance, the concatenation of H, K and L, depicted in Figure 5, is a 1-*structure* with genus $g = 3$.

More than 95% of all known pseudoknot structures are composed by the four irreducible-shapes [5]. Thus, we use

[2] Note that here the subscripts with the loop names are given only to help reading the sequence, but they are not part of the language.

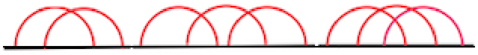


Figure 5: 1-*structure* of genus $g = 3$ composed of H, K and L shapes.

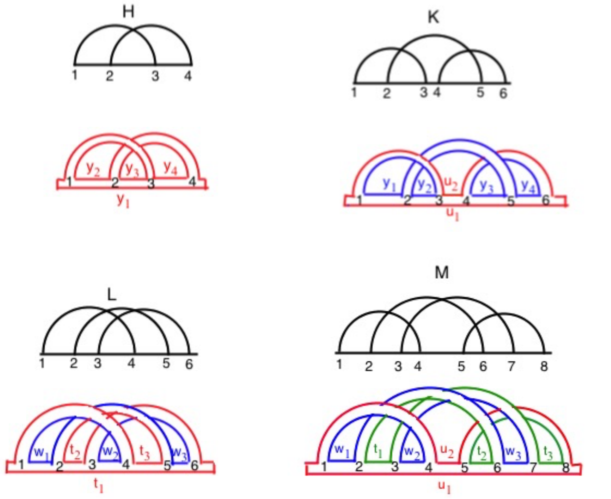


Figure 6: The four irreducible shapes and the corresponding fatgraphs in which the distinct cycles of the boundaries are depicted in different colors.

irreducible-shapes and the *group* algebraic structure, which has been used for characterizing the words of a formal language [2], to define the language of RNA shapes of genus $g = 1$. A group G can be presented, written $G = \langle S \mid R \rangle$, by a set $S = \{s_1, ..., s_m\}$ of generators and a set $R = \{r_1, ..., r_n\}$ of relations among the generators [10]. In the following we start with each of the four irreducible-shapes and we obtain a corresponding group presentation.

The first step is to move from a shape to a fatgraph by fatting the arcs of the shape into ribbons. Figure 6 shows the fatgraphs obtained from the irreducible-shapes H, K, L and M. It is easy to see from the figure that the boundaries of the fat arcs (colored in red, blue and green) are possibly interlaced disjoint cycles, depending on the particular shape. For instance, the boundaries of the fatgraph of shape H form a unique red cycle. In case of K and L there are two disjoint cycles, colored in red and blue, while for M there are three disjoint cycles, colored red, blue and green. The original arcs connecting the vertices of the shape along the horizontal line (representing the backbone of the original RNA diagram) can be viewed as the elements of these cycles. To close the cycle, an arc connecting vertex 1 to the last vertex on the right of each shape is added. As an example, consider the arc labelled y_1 of the fatgraph of H connecting vertex 1 with vertex 4. Then, we can say that the unique cycle of shape H has elements y_1, y_2, y_3 and y_4. Similarly, the red cycle of K has elements u_1 and u_2, while the other blue cycle has elements y_1, y_2, y_3 and y_4. Finally, the red cycle of M has elements u_1 and u_2, the blue cycle has elements w_1, w_2 and w_3 while the green cycle has elements t_1, t_2 and t_3.

For each cycle, it is possible to associate a cyclic group with a finite presentation. All the groups of the irreducible-shapes are finitely generated groups and all groups with finite presentations are *automatic groups* [12]. An automatic group is a finitely generated group equipped with several finite-state automata. These automata can tell if a given word representation of a group element is in a "canonical form". This machinery can be used to define the part of our language for representing pseudoknotted secondary structures.

3. CONCLUSIONS AND FUTURE WORK

We introduced a formal language to represent pseudoknot free RNA secondary structures and we introduced the necessary mathematical tools to describe also pseudoknotted structures. In a future work we will use the resulting shape language to represent the RNA folding as an $S[B]$ system evolving towards a minimal energy configuration.

4. REFERENCES

[1] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of rna structures. *Journal of Molecular Biology*, 379(4):900–911, 2008.

[2] D. Epstein, J. Cannon, D. Holt, S. Levy, M. Paterson, and W. Thurston. *Word processing in Groups.* Jones & Bartlett, London, UK, 1992.

[3] A. Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, UK, 2001.

[4] I. Hofacker, M. Fekete, and P. Stadler. Secondary Structure Prediction for Aligned RNA Sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[5] F. W. Huang and C. M. Reidys. Shapes of Topological RNA Structures. Available at `http://arXiv:1403.2908`, 2014.

[6] P. Kerpedjiev, S. Hammer, and I. Hofacker. Forna (Force-Directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams. *Manuscript submmitted for publication.*, 2014.

[7] R. B. Lyngsøand N. S. Pedersen. RNA Pseudoknot Predictions in Energy-Based Models. *Journal of Computational Biology*, 7:409–427, 2000.

[8] E. Merelli, N. Paoletti, and L. Tesei. Adaptability Checking in Multi-Level Complex Systems. *Science of Computer Programming*, `http://dx.doi.org/10.1016/j.scico.2015.03.004`.

[9] E. Merelli, M. Pettini, and M. Rasetti. Topology Driven Modeling: the IS Metaphor. *Natural Computing*, 14(3):421–430, 2015.

[10] J. S. Milne. Group Theory, 2010. Available at `http://www.jmilne.org/math/`.

[11] K. V. Morris and J. S. Mattick. The Rise of Regulatory RNA. *Nat. Rev. Genet*, 15:423–437, 2014.

[12] S. Rees. Hairdressing in Groups: a Survey of Combings and Formal Languages. In *The Epstein Birthday Schrift*, volume 1 of *Geometry and Topology Monographs*, pages 493–509. MSP, 1998.

[13] C. M. Reidys, F. W. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, and M. E. Nebel. Topology and Prediction of RNA Pseudoknots. *Bioinformatics*, 27(8):1076–1085, 2011.

[14] M. Zuker. Mfold: Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.*, 31:3406–3415, 2003.