Partitioning RNAs into pseudonotted and pseudoknot-free regions modeled as Dual Graphs

Louis Petingi[*] and Tamar Schlick[†]

January 19, 2016

Abstract

Dual graphs have been applied to model RNA secondary structures. The purpose of the paper is two-fold: we present new graph-theoretic properties of dual graphs to validate the further analysis and classification of RNAs using these topological representations; we also present a linear-time algorithm to partition dual graphs into topological components called *blocks* and determine if each block contains a *pseudoknot* or not. We show that a block contains a pseudoknot if and only if the block has a vertex of degree 3 or more; this characterization allows us to efficiently isolate smaller RNA fragments and classify them as pseudoknotted or pseudoknot-free regions, while keeping these sub-structures intact. Even though non-topological techniques to detect and classify pseudoknots have been efficiently applied, structural properties of dual graphs provide a unique perspective for the further analysis of RNAs. Applications to RNA design can be envisioned since modular building blocks with intact pseudoknots can be combined to form new constructs.

Keywords: Graph Theory, RNA Secondary Structures, Partitioning, Biconnectivity, Pseudoknots.

1 Introduction

Graph theory is a well-established field of mathematics with applications to areas where the objects can be modeled as discrete structures called *graphs* or

[*]College of Staten Island, City University of New York, Computer Science Department, Staten Island, New York, email: louis.petingi@csi.cuny.edu. Corresponding Author

[†]New York University, Department of Chemistry and Courant Institute of Mathematical Sciences, New York, New York, email: sclick@nyu.edu

1

networks. The study of combinatorial properties of these networks, such as communication, chemical, and biological networks, can be guided by graph-theoretical principles and algorithms. Specific examples include the study of chemical structures (e.g., hydrocarbons, drug compounds) [16, 21], genetic and cellular relationships [10, 22], and transportation networks [24].

In mathematical terms, an undirected graph $G = (V, E)$ is a discrete object described by a finite set of *vertices* V and a set E of unordered pair of vertices called *edges*, where each edge represents a connection between two vertices.

The graphs described in this paper were introduced in 2003 by Gan et. al [8], called dual graphs, were applied to model RNA secondary structures (2D). The 2D elements of RNA molecules consist of double-stranded (stem) regions by base pairing such as Adenine-Uracil, Guanine-Cytosine, Guanine-Uracil, and single stranded loops; stems and loops are mapped to the vertices and edges of the corresponding dual graph, respectively. Dual graphs are needed to represent pseudoknots, structures involving an inter-wining of two-base-paired regions of the RNA. These are common elements in many biologically important RNAs.

Given a graph $G = (V, E)$, let the degree of a vertex $u \in V$ be the number of edges incident at u in G. In this paper we introduce a partitioning algorithm for dual graph representations of RNA 2D structures to recognize pseudoknots. Our algorithm partitions a dual graph into graph-theoretic components called *blocks* and then determines whether each block contains a pseudoknot; we show that a block contains a pseudoknot if and only if the block has a vertex of degree 3 or more. Thus our methodology provides a systematic approach to partition an RNA 2D structure, modeled as a dual graph, into smaller RNA regions containing pseudoknots, while providing a new topological perspective for the analysis of RNAs.

Pseudoknots can be classified into two main groups: *standard* and *recursive* pseudoknots [6, 28]. The latter is distinguished from the former by having nested pseudoknots within a pseudoknot. While our partitioning algorithm can detect general pseudoknots, it is not within the scope of this work to classify them. With extensions of our graph-theoretical techniques, however, it may be possible to analyze and treat standard and the more complex recursive pseudoknots structures further, as needed for specific biological applications.

In the next section, we present background material relevant to this paper, as well as notation and mathematical definitions of RNA primary, secondary, and of pseudoknot structures. In Section 3 we describe our partitioning approach of a dual graph G into components $G' \subseteq G$ called blocks,

2

as well as new combinatorial properties of dual graphs. In Section 4, we characterize these blocks, and show that a block contains a pseudoknot in the RNA 2D structure if and only if the block has a vertex of degree 3 or more. This characterization permits us to isolate pseudoknots, without *breaking* them so their structural properties can be further studied; moreover we also present an alternative way to visualize the presence of pseudoknots in blocks by considering their planar *geometric-duals* ([13], pg.113). In Section 5 we illustrate algorithmic tests performed on dual graph representations of existing RNA motifs, and classify each block of a dual graph as having a pseudoknot or not. We summarize the findings and outline new directions in Section 6. An Appendix section includes definitions, mathematical proofs, and supporting material.

2 RNA background and definitions

Modeling of RNAs as graphs began in 1978 when Waterman introduced topological representations of RNA to analyze the secondary structure of tRNA [27]. In 1990, Shapiro and coworkers used a tree representation of 2D structures to measure structural similarities [25].

In 2003, Gan et. al introduced *tree* and *dual* graph-theoretic representations of RNA 2D motifs in a framework called RAG (RNA-As-Graphs) [7, 8, 9, 15]. While tree graphs are intuitive and easily applied to many areas of RNA research such as partitioning (e.g., Kim et al. [19]), complex RNA secondary structures with pseudoknots (*PKs*) can only be represented by dual graphs (see [18] for a survey on these topological representations); a pseudoknot is an intertwining of two-based-paired regions (stems) of an RNA (see Figure 1).

The structural configuration of pseudoknots does not lend itself well to computational detection due to its overlapping nature. The base pairing in pseudoknots is not well-nested making the presence of pseudoknots in RNA sequences more difficult to predict by the dynamic programming [4] and context-free grammars standard methods [2]. Our methodology, based on topological properties of dual graphs, provides a new perspective for the problem of detection and classification of pseudoknots and of general RNAs.

Following (Kravchenko, 2009 [17]), we define our biological variables as follows.

Definition 1 *General terms:*

    i. RNA primary structure: a sequence of linearly ordered bases $x_1, x_2, \ldots, x_r$, where $x_i \in \{A, U, C, G\}$.

ii. canonical base pair: a base pair $(x_i, x_j) \in \{(A, U), (U, A), (C, G), (G, C),$
   $(G, U), (U, G)\}$.

iii. RNA secondary structure without pseudoknot - or regular structure,
   encapsulated in the region $(i_0, \ldots, k_0)$: an RNA 2D structure in which
   no two base pairs $(x_i, x_j), (x_l, x_m)$, satisfy $i_0 \leq i < l < j < m \leq m_0$
   (i.e., no two base pairs intertwined).

iv. a base pair stem: a tuple $(x_i, x_{i+1}, \ldots, x_{i+r}, x_{i+(r+1)}, \ldots, x_{j-1}, x_j)$ in
   which (x_i, x_j), $(x_{i+1}, x_{j-1}), \ldots, (x_{i+r}, x_{i+(r+1)})$ form base pairs.

v. loop region: a tuple $(x_1, x_2, \ldots, x_r)$ in which $\forall_{i \leq j \leq r}(x_i, x_j)$ does not
   form a base pair.

vi. a pseudoknot encapsulated in the region $(i_0, \ldots, k_0)$: if $\exists l, m, (i_0 < l <$
   $m < k_0)$ such that (x_{i_0}, x_m) and (x_l, x_{k_0}) are base pairs.

A graphical representation is an intuitive and natural way to depict an RNA 2D structure (see Figure 1-(a),(b)), in which the x-axis is labeled according to the primary linearly ordered sequence of bases (Definition 1-i), and a stem (Definition 1-iv) is represented by arcs connecting base pairs. A region on the x-axis between the end-points of the arcs representing stems is called a *segment*.

A dual graph can be equivalently defined from the graphical representation of an RNA 2D structure as follows (Figure 1).

Definition 2 *The dual graph is derived by mapping stems and the segments between stems (x-axis), of the graphical representation of an RNA 2D structure, to the vertices and edges of the dual graph, respectively.*

In the next section we propose our partitioning approach of a dual graph G, into subgraphs $G' \subseteq G$, called blocks. In addition we show new combinatorial properties of dual graphs that will be later invoked to prove that a block contains a pseudoknot if and only if the block has a vertex of degree 3 or more.

3 Partitioning of a graph into blocks and combinatorial properties of dual graphs

Let $G = (V, E)$ be a connected graph; the following definitions will be used below. Unless otherwise stated, we follow the notation of Harary [13],
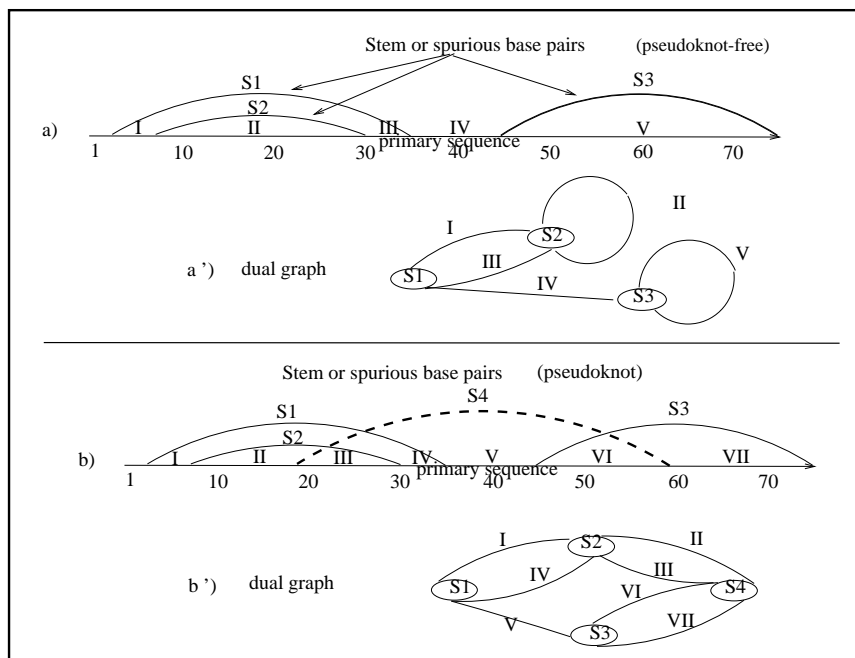
Figure 1: *Graphical and dual graph representations of an RNA 2D structure.* (a) graphical representation of a pseudoknot-free RNA primary sequence and embedded stems or base pairs; (a′) corresponding dual graph representation. (b) graphical representation of a pseudoknotted RNA 2D structure; (b′) corresponding dual graph.

Definition 3 *Connectivity*

- i. *A vertex-set $X \subseteq V$ is a vertex-disconnecting set if deletion of X from G, denoted by $G - X$, results in a disconnected graph.*

- ii. *A vertex v is an articulation point or cut-vertex if $G - v$ results in a disconnected graph (i.e., at least two components remain).*

- iii. *The vertex-connectivity, $\kappa(G)$, is the minimum number of vertices whose removal from G results in a disconnected graph or in a isolated vertex. If G is a single edge, then $\kappa(G) = 1$.*

- iv. *A connected component is non-separable if it does not have an articulation point (or cut-vertex). Please note that single edges or isolated points are non-separable.*

5

v. A block is a maximal (edge-wise) non-separable graph.

The concepts of articulation points and maximal non-separable components (blocks) are related. Indeed, articulation points partition any graph into blocks (see Fig. 2). Except when a graph G is composed of just two vertices adjacent by one or several (parallel) edges ($\kappa(G) = 1$), any maximal non-separable graph is bi-connected (i.e., $\kappa(G) \geq 2$). The fact that blocks are maximally non-separable subgraphs allows us to isolate pseudoknots, without breaking their structural properties.
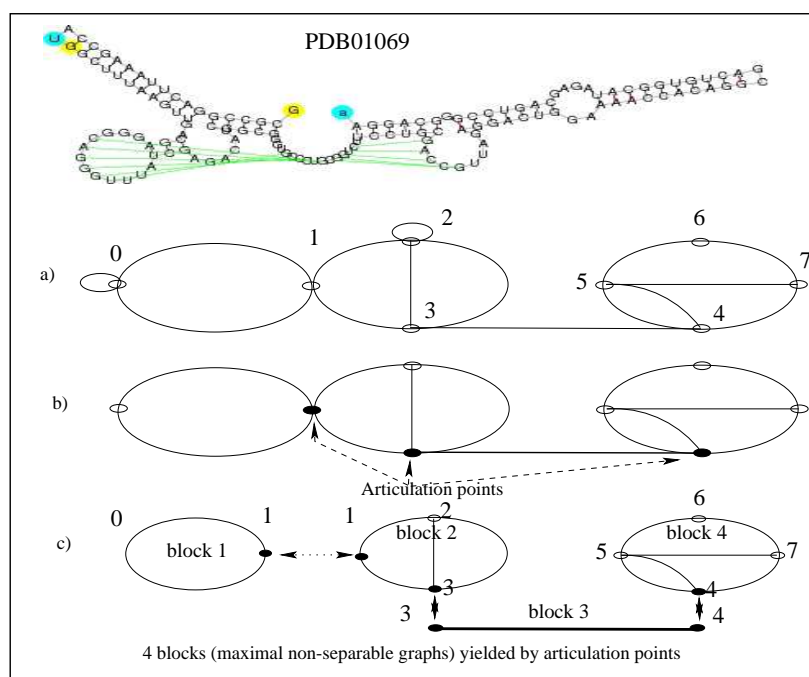

Figure 2: *Identification of articulation points and partitioning of the dual graph corresponding to PDB01069 RNA 2D (Catalytic Ribozyme RNA) into blocks.*

Our partitioning algorithm is based on the classical result for identifying block components in a connected undirected graph introduced to John Hopcroft and Robert Tarjan (1973, [14]), that runs in linear computational time.

We next state new combinatorial properties associated with dual graphs that will be invoked to show that a block contains a pseudoknot if and only

if it has a vertex of degree at least 3.

A graph is *Eulerian* if there exist a trail (see Definition 5-iii, Appendix 7.1) from a vertex v_0 of G, ending at vertex v_k, covering all the edges of the topology; if $v_0 = v_k$ then the graph is an *Eulerian cycle* (see [13], pg. 64). Dual graph representations of general RNA 2D structures, and specifically of PKs, can be easily shown to be Eulerian graphs from Definition 2. By starting from the origin on the x-axis of the graphical representation and traversing to the right, a unique trail in its dual graph can be described, where all edges are covered.

Lemma 1 *The dual graph representations of RNA 2D structures and of PKs are Eulerian.*

As depicted in Figure 1-(b), the alternating sequence of stems and segments $\{S_1, I, S_2, II, S_4, III, S_2, IV, S_1, V, S_3, VI, S_4, VII, S_3\}$ of the graphical representation (b) forms an Eulerian trail in its dual graph (b′).

As a consequence of Lemma 1, it follows that at most two vertices are of odd degree in a dual graph representation of an RNA 2D structure or a PK (see [13], pg. 64). From this point on, we delete self-loops in dual graphs as they correspond to stems, in a graphical representation, not containing or not being crossed (intertwined) by another stem. In addition under the assumption that two different stems cannot share the same bases, each endpoint of a stem in the graphical representation can be adjacent to at most two other stems, and thus the maximum degree of a vertex of a dual graph is 4.

These facts together with Lemma 1, yield the following corollary that will be later referenced to prove our main results.

Corollary 2 *Dual graph representation of an RNA 2D structure (PKs) has the following properties:*

1. *The graph is Eulerian.*

2. *The maximum degree of any vertex is four.*

3. *The graph has at most two vertices of odd degree.*

7

4 Mapping PKs to blocks with certain degrees and combinatorial properties of dual graphs

4.1 Partitioning of dual graphs into pseudonotted and pseudoknot-free blocks

Now we can prove that once a dual graph representing an RNA 2D structure has been partitioned into blocks, a block is contains a pseudoknot if and only if it contains a vertex of degree 3 or more. The mathematical proofs of the lemmas stated in this section, are shown in Appendix 7.2.

In preparation to the main results of this chapter, we first define the following.

Definition 4 *For any graph G, blocks can be partitioned into three classes,*

1. *Single edges.*

2. *Cycles.*

3. *Blocks containing a vertex v of degree at least 3.*

From Definition 1-iii, an RNA 2D structure is regular (pseudoknot-free) and encapsulated in a region $(i_0, \ldots, k_0)$, if no two base pairs $(x_i, x_j), (x_l, x_m)$, satisfy $i < l < j < m$, $i_0 \leq i, j, l, m \leq m_0$. Under the previous assumption that self-loops are deleted, this definition yields the following lemma,

Lemma 3 *Each block in the dual graph representation of a regular RNA 2D structure is either a bridge or a cycle of length $l, l \geq 2$ (see Definition 4-1,2).*

Conversely we show the following.

Lemma 4 *If an RNA 2D structure contains a pseudoknot, then its corresponding dual graph contains a block having a vertex of degree 3 or more (see Definition 4-3).*

Lemma 3 and Lemma 4 yield our main result as follows.

Corollary 5 *Given a dual graph representation of RNA 2D structure, a block represents a pseudoknot if and only if the block has a vertex of degree 3 or more.*

To summarize our partitioning algorithm, we performed the following steps.

8

1. *Partition the dual graph into blocks by application of Hopcroft and Tarjan's algorithm, as described in Section 3.*

2. *Analyze each block to determine if it has a vertex of degree at least 3. If that is the case then the block contains a pseudoknot, according to Corollary 5. If not then the block represents a pseudoknot-free structure.*

4.2   Other relevant combinatorial properties of dual graphs

The analysis required to identify and study different types of pseudoknots, based on combinatorial properties of dual graphs, seems to be justified. For example every dual graph known so far has a planar embedding, i.e., it can be drawn in a plane so no two edges cross, even though to show planarity for general dual graphs remains an open problem. The planarity property of dual graphs is not only supported by simple observations but also by the fact that the degree of any vertex of a dual graph is at most four. Every planar graph G can be divided into regions, where each edge in E divides exactly two regions of G. The *geometric-dual*[1] (see [13], pg. 113) representation $G_d = (V_d, E_d)$ of a planar graph G is obtained as follows: for each region of the graph G we have a corresponding vertex in G_d, and if two regions of the graph G have a common edge bordering them, then the corresponding vertices in G_d will be adjacent.

It can be easily shown, from Corollary 5, that the geometric-dual of a graph G representing a pseudoknot-free region (under the assumption that G is planar) is composed of a vertex v_{outer}, corresponding to the outer-region of G, and v_{outer} is adjacent to a vertex v by r parallel edges, if v maps to a cycle (phase) on r vertices in G (bridges in G correspond to self-loops in G_d emanating from v_{outer}). Consequently deleting the vertex v_{outer} from G_d results in just isolated vertices (see Fig. 3). Conversely, deleting v_{outer} from G_d when G represents a pseudoknotted region, results in a graph containing connected components with at least two vertices corresponding to pseudoknotted blocks (see Fig. 4). These facts strongly suggest that the use of dual graphs and their geometric-duals, can shed light into the connectivity and degree analyses of standard pseudoknots (see Definitions 7 and 8, Appendix 7.3), as well as more complex pseudoknotted structures (i.e., recursive pseudoknots); it also provides a fertile theoretical ground for

---

[1]geometric-duals of planar graphs are also known as "dual-graphs" in the literature, which unfortunately coincide with the denomination of the graph representations of RNA 2D discussed in this paper.
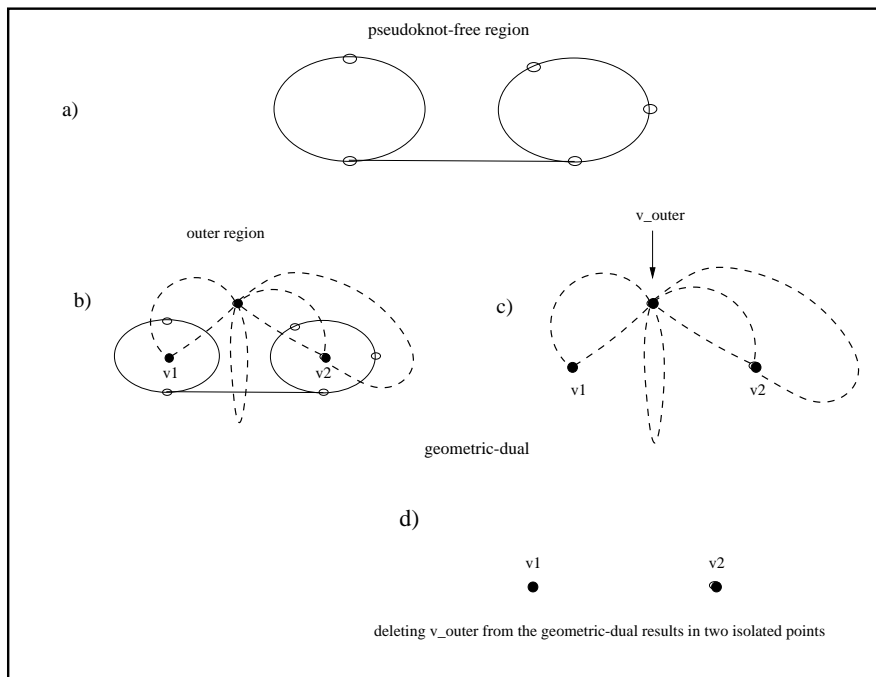
the study of RNA structure and function.

Figure 3: *A pseudoknot-free region.* a) dual graph representation of a pseudoknot-free RNA; b) and c) corresponding geometric-dual; d) deleting v_{outer} results in just isolated points.

5 Experimental results

We illustrate our partitioning algorithm, described at the end of Section 4.1, on the dual graph representations of two RNA 2D structures, based on the New York University's RAG database [15]. Our partitioning algorithm was implemented in C++ and run on a Hewlett-Packard Pavilion Dv6 (2.4 GHz) notebook. Each partitioning takes less than one second because of the linear computational complexity of Hopcroft and Tarjan's algorithm.

Consider the PDB01069 RNA 2D structure, *Post-Cleavage State of the Thermoanaerobacter Tengcongenis GlmS Ribozyme*, known to be the only catalytic RNA to require a small-molecule activator for catalysis (see Klein et al. [20]). Its dual graph is decomposed into 4 blocks as illustrated in
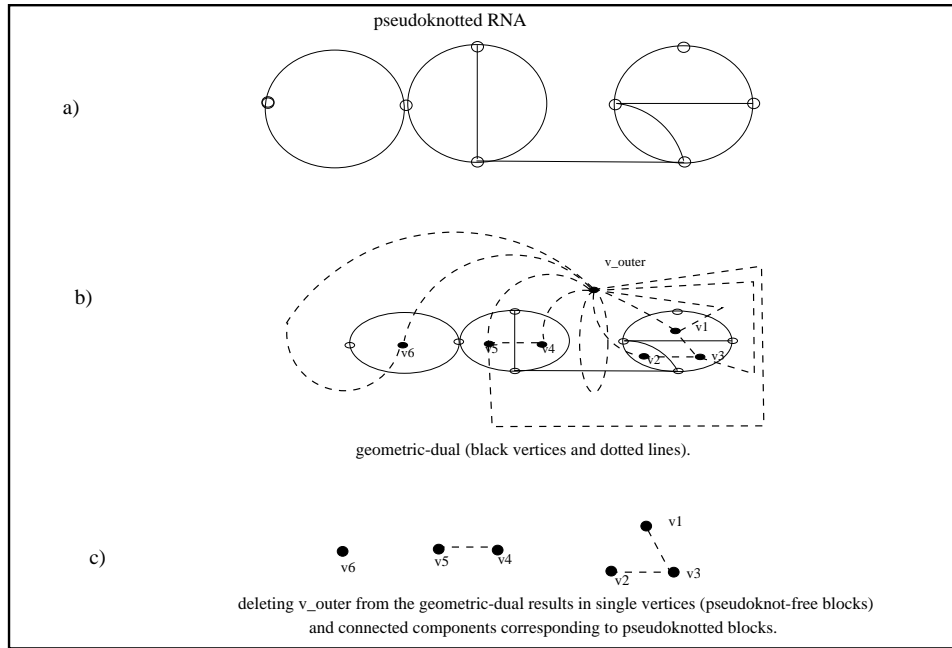
Figure 4: *A pseudoknotted region.* a) dual graph representation of a pseudo-knotted RNA; b) corresponding geometric-dual G_d; c) deleting v_{outer} from G_d results in either isolated vertices corresponding to pseudoknot-free blocks or connected components (pseudoknotted blocks).

Figure 2. According to Corollary 5, block 1 and block 3, a cycle and an edge, respectively, correspond to regular regions, while blocks 2 and 4, correspond to pseudoknots.
We next consider the dual graph representation of PKB236 (see Fig. 5), *Regulatory Pseudoknot of the Interferon-gamma Gene 5′-UTR*, thought to be involved in regulatory translation (see Ben-Asouli et al. [1]); in this case the only block is the dual graph itself. As this block contains a vertex of degree 3 or more, then this block is a pseudoknot.

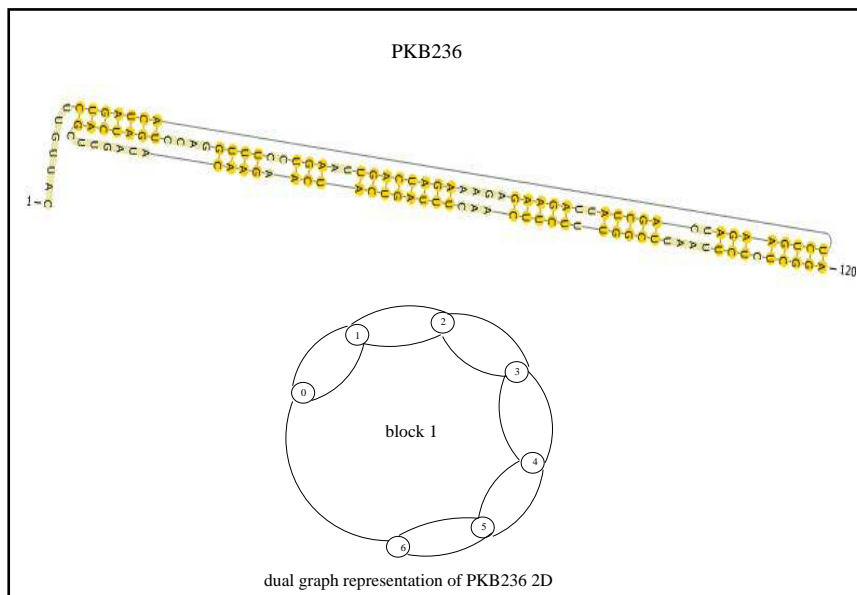Appendix 7.4 depicts the output generated when our algorithm was run on the aforementioned RNA structures.

Figure 5: *Partition of the dual graph corresponding to motif PKB236 (Regulatory Pseudoknot of the Interferon-gamma gene 5′-UTR).*

6   Concluding remarks and future work

We have presented a partitioning approach of the dual graph representation of RNA 2D structures into maximal non-separable components called blocks. Partitioning of a graph into blocks can be efficiently accomplished by application of Hopcroft and Tarjan's algorithm to identify articulation points. From mathematical definitions of RNA 2D structures and of pseudoknots, we proved that an RNA 2D structure contains a pseudoknot if and only if the dual graph representation has a block in which one of the vertices is of degree 3 or more. Our partitioning algorithm suggests that dual graphs along with the graph-theoretic properties described in the paper can help analyze connectivity and features of both standard (see Appendix 7.3) and recursive pseudoknots. Ultimately partitioning and classification of dual graphs could guide the discovery of modular regions of RNA and thus be exploited for design of novel RNAs constructed from these building blocks.

12

References

[1] Y. Ben-Asouli et al., Human Interferon-γ mRNA Autoregulates Its Translation through a Pseudoknot that Activates the Interferon-Inducible Protein Kinase PKR, *Cell* 108, 2002, pp. 221-232.

[2] A. Condon et. al, Classifying RNA pseudoknotted structures, *Theoretical Computer Science* 320(1), 2004, pp. 3550.

[3] R.M. Dirks, and N.A. Pierce, A partition function algorithm for nucleic acid secondary structure including pseudoknots; *J. Comput. Chem* 24 (13) 2003, pp. 16641677.

[4] R. M. Dirks, and N.A. Pierce, An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J Computation Chemistry* 25, 2004, pp. 1295-1304.

[5] J. C. Burnett and J. J. Rossi, RNA-based therapeutics: current progress and future prospects, *Chem Biol.* 19(1), 2012, pp. 60–71.

[6] B. Dost et al., Structural Alignment of Pseudoknotted RNA,*Journal of Computational Biology* 15(5), 2008, pp. 489–504.

[7] D. Fera et al., RAG: RNA-As-Graphs web resource, *BMC Bioinformatics* 5, 2004, pp. 88.

[8] H. H. Gan et al., Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design, *Nucleic Acids Res.* 31(11), 2003, pp. 2926 – 2943.

[9] H. H. Gan et al., RAG: RNA-As-Graphs database–concepts, analysis, and features, *Bioinformatics* 20(8), 2004, pp. 1285 –1291.

[10] K. C. Gunsalus et al., Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis, *Nature* 436(7052), 2005, pp. 861–865.

[11] P. Guo, The emerging field of RNA nanotechnology, *Nat Nanotechnol.* 2010, pp. 833–842.

[12] A. J. Hamilton and D. C. Baulcombe, A species of small antisense RNA in post transcriptional gene silencing in plants, *Science* 286(5441), 2009, pp. 950–952.

[13] F. Harary, *Graph Theory*, Addison-Wesley, Mass. 1969.

[14] J. Hopcroft, and R. Tarjan, Efficient algorithms for graph manipulation. Communications of the ACM 16 (6), 1973, pp. 372–378.

[15] J. A. Izzo et al., RAG: an update to the RNA-As-Graphs resource, *BMC Bioinformatics* 12, 2011, pp. 219.

[16] M. Johnson, Structure-activity maps for visualizing the graph variables arising in drug design. *J. Biopharm. Stat.* 3(2), 1993, pp. 203–236.

[17] A. Kravchenko, Predicting RNA Secondary Structures Including Pseudoknots, *University of Oxford* internal report, 2009.

[18] N. Kim et al., Network Theory Tools for RNA Modeling, *WSEAS Transactions on Math.* 12(9), 2013, pp. 941-955.

[19] N. Kim et al., RNA Graph Partitioning for the Discovery of RNA Modularity: A Novel Application of Graph Partition Algorithm to Biology, *PLOS ONE* 9(9), 2014.

[20] D.J. Klein et al., Structural Basis of glmS Ribozyme Activation by Glucosamine-6-phosphate, *Science* 313, 2006, pp. 1752-1756.

[21] M. Mandado et al., Chemical graph theory and n-center electron delocalization indices: a study on polycyclic aromatic hydrocarbons, *J. Comput. Chem.* 28(10), 2007, pp. 1625–1633.

[22] R. Milo, Superfamilies of evolved and designed networks, *Science* 303(5663), 2004, pp. 1538–1542.

[23] L. Petingi, Identifying and Analyzing Pseudoknots based on Graph-Theoretical Properties of Pseudoknots: A Partitioning Approach, *CUNY Graduate Center Academic Works*, Internal Report, 2015.

[24] P. Santi et al., Quantifying the Benefits of Vehicle Pooling with Shareability Networks, *PNAS* 111(37), 2014, pp. 13290-13294.

[25] B. Shapiro and K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, *Comput. Appl. Biosci.* 6(5), 1990, pp. 309 – 318.

[26] P. A. Sharp, The centrality of RNA, *Cell* 136(4), 2009, pp. 577–580.

[27] M. S. Waterman, Secondary structure of single-stranded nucleic acids, *Adv. Math. Suppl. Stud.* 1, 1978, pp. 167–212.

[28] T. K. Wong et al., Structural alignment of RNA with complex pseudo-knot structure, *J Comput Biol.*18(1), 2011, pp. 97–108.

7 Appendix

7.1 Graph Theory definitions

Let $G = (V, E)$ be a graph with vertex-set V and edge-set E. We next present general graph-theoretic definitions following Harary [13].

Definition 5 *General graph-theoretic terms:*

    i. *Let $H_1.x.H_2$ represent the graph composed of two graphs, H_1, and H_2, sharing the same vertex x.*

    ii. *A walk between two vertices u and v in graph $G = (V, E)$, is an alternating sequence of vertices and edges $< v_o = u, e_1, v_1, \ldots, e_k, v_k = v >$ such that $e_i = (v_{i-1}, v_i)$ is an edge of G.*

    iii. *A trail between two vertices u and v in graph $G = (V, E)$, is a walk between u and v with no repetition of edges.*

    iv. *A path between two vertices u and v in graph $G = (V, E)$, is a walk (or trail) between u and v with no repetition of vertices.*

    v. *A graph is Eulerian if there exist a trail from a vertex v_0 of G, ending at vertex v_k, covering all the edges of the topology, and if $v_0 = v_k$ then the graph is an Eulerian cycle.*

7.2 Proofs of lemmas stated in Section 4.1

From Definition 1-iii, an RNA 2D structure is regular (pseudoknot-free) and encapsulated in a region $(i_0, \ldots, k_0)$, if no two base pairs $(x_i, x_j), (x_l, x_m)$, satisfy $i < l < j < m$, $i_0 \leq i, j, l, m \leq m_0$. Under the previous assumption that self-loops are deleted, this definition yields the following lemma,
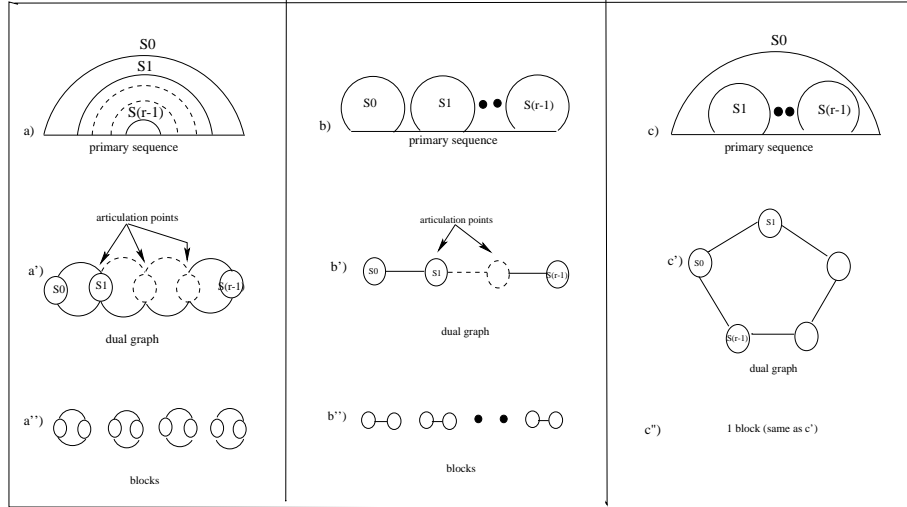
Figure 6: *Classification of PK-free regions and graphical/dual/block representations.* (a) graphical, dual, and block representations of r nested-stems - (b) graphical, dual, and block representations of r adjacent stems - (c) graphical, dual, and block representations of a stem containing $r-1$ adjacent stems.

Lemma 3 *Each block in the dual graph representation of a regular RNA 2D structure is either a bridge or a cycle of length $l, l \geq 2$ (see Definition 4-1,2).*

Proof. Consider the graphical representation of a regular RNA 2D structure; we will proceed by construction. A regular (pseudoknot-free) region can be recursively defined as follows (see Fig. 6): (a) a region composed of r nested-stems; (b) r adjacent stems, (c) a stem containing a sequence of $r-1$ adjacent stems; (d) a single stem (represented as an isolated vertex in its dual graph, not illustrated in Fig. 6). In a transformation, a set of stems identified by properties a, b, and c in the graphical representation, are reduced (converted) into a single stem, while its corresponding dual graph is generated (see Definition 2). The blocks obtained from the dual graph representations of these properties, are either cycles of length 2, single edges, a cycle of length r, or an isolated vertex, respectively. Consider a sequence of transformations of dual graphs $G_1 \Rightarrow G_2 \Rightarrow \ldots \Rightarrow G_n$, where the dual graph G_{i+1} is obtained from dual graph G_i by following the precedence rules in which, first, internal stems of the ones identified by properties (a) through (c) of the graphical representation are reduced into a single stem, while the

16

corresponding dual graph is generated; in the dual graph we distinguish the vertex corresponding to the outer-stem. Because only distinguished vertices could be later made adjacent to other vertices in a transformation, the blocks generated by the sequence of transformations from G_1 through G_{n-1} will remain blocks in G_n, with the possible addition of blocks composed of single edges. ∎

To illustrate Lemma 3, consider Figure 7 depicting the graphical representation of a pseudoknot-free region. The stems S_0, S_1, and S_2, identified by property (a), with corresponding dual graph with distinguished vertex S_0, are then reduced into a single stem in the graphical representation. Similarly, by property (a), we reduce the pairs of nested-stems S_3, S_4, and S_5, S_6, to two single stems with distinguished vertices S_3 and S_5, in the dual graph, respectively. As the stem S_9 contains a sequence of 3 (reduced) stems, by application of property (c), it can be reduced to a single stem with dual graph composed of a cycle on 4 vertices, and distinguished vertex S_9. Finally, by property (b), we connect the sequence of 3 (reduced) stems (i.e., S_0, S_9, and S_8) by single edges in the dual graph.
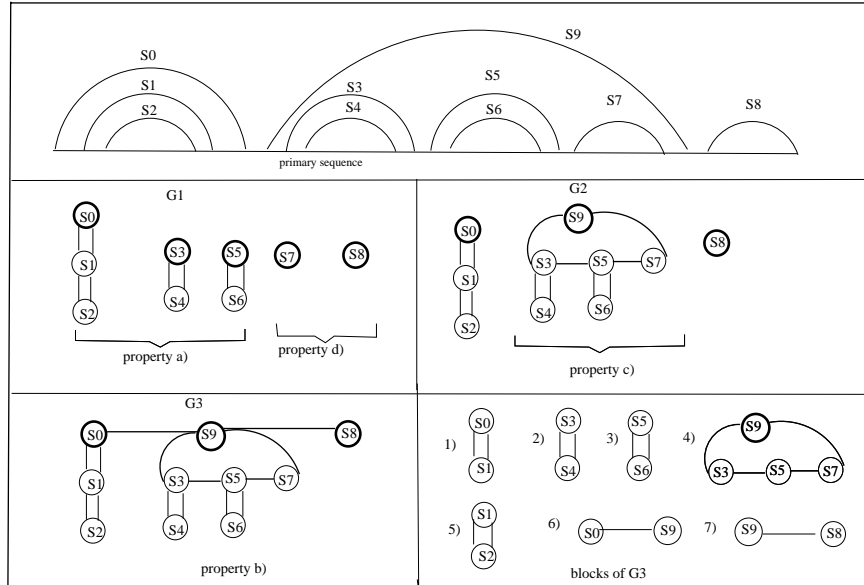

Figure 7: *An example illustrating Lemma 3.*

Conversely we show the following.

Lemma 4 *If an RNA 2D structure contains a pseudoknot, then its corresponding dual graph contains a block having a vertex of degree 3 or more (see Definition 4-3).*

Proof. By Definition 1-vi, if an RNA 2D structure contains a pseudoknot, there exist a stem *crossing* (interweaving) another stem. Let us denominate these interweaving stems, in the graphical representation, S_1 and S_2, respectively. There exist then three independent paths, X_1, X_2, and X_3, from S_1 to S_2 (see Figure 8-(a)), following the primary sequence of the graphical representation; these three paths correspond to trails in the dual graph representation (see Definition 5-iii of Appendix 7.1). We first note that $X_2 \cup X_3$ forms an Eulerian cycle G_1 in the dual graph representation (see Definition 5-v, and Lemma 1), beginning and ending at S_2, having S_1 as one of its vertices. Because an Eulerian cycle is the union of simple cycles ([13], pg. 64) (Figure 8-(b)), then the articulation points of G_1 have maximum possible degree 4 (see Corollary 2-2); when we add then the trail X_1, from S_1 to S_2 to G_1, X_1 cannot touch (include) any of the articulation points of G_1. Let $G_2 = B_1.a.B_2.b.B_3.c.B_4 \ldots B_r$ (see Definition 5-i, Appendix 7.1) be a subgraph of G_1 describing a sequence of blocks $B_1, B_2, \ldots, B_r$, S_1 is a vertex of B_1, and S_2 is a vertex of B_r, in which the set $\tilde{A} = \{a, b, c.....\}$ is the set of articulation points connecting the blocks of G_2. Let G^* be the graph obtained by adding the trail X_1 to G_2. Clearly $\kappa(G^*)$ (see Definition 3-iii) is at least 2 as deleting a single articulation point in $\tilde{A}$ won't disconnect G^* as X_1 does not have a vertex in $\tilde{A}$, thus G^* is a non-separable graph (Definition 3-iv). As both S_1 and S_2 have degree at least 3 in G^*, then there is a block containing G^* (possibly itself) having a vertex of degree 3 or more. ∎

7.3 Degrees in Pseudoknots

In this section we introduced the concept of *degree* in pseudoknots as defined by Dost et al. [6], to suggest the study of structural properties of RNA secondary structures and of pseudoknots, represented as dual graphs.

Intuitively speaking the degree of a pseudoknot refers to the minimum number of *turns* (plus one), the primary sequence must be object to avoid the crossing of base pairs. As an example consider Figure 9-a, in which two turns (i.e., the pseudoknot is of degree 3) of the primary sequence suffice so no two base pairs intertwine.

Let $x_1, x_2, \ldots, x_n$ be a sequence of linearly ordered bases. An alternative notation for a secondary structure is: $M = \{(i, j) | 1 \leq i < j \leq n, (x_i, x_j)$ is a base pair$\}$. Let also define $M_{i_0, k_0} \subseteq M$, as $M_{i_0, k_0} = \{(i, j) \in M | i_0 \leq i <$
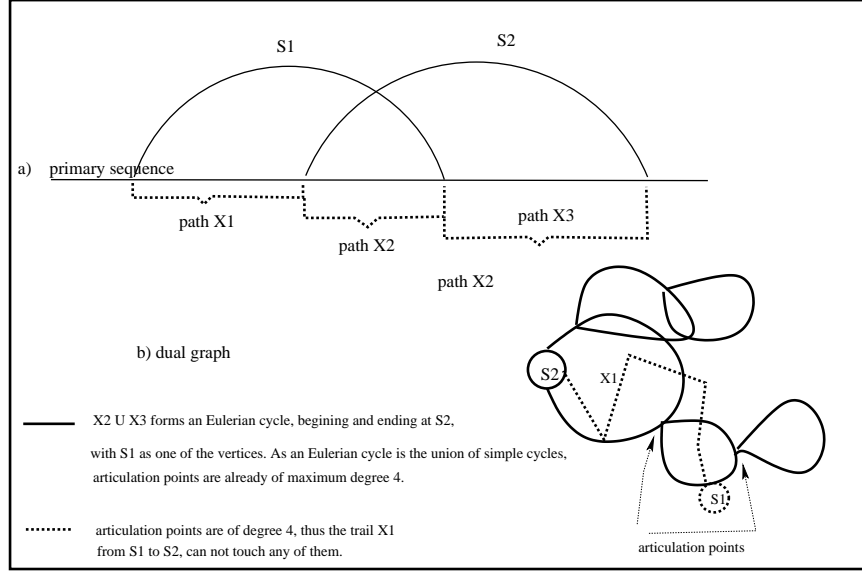
18

Figure 8: *Supporting illustration for the proof of Lemma 4*

$j \leq k_0\}$. The secondary structure, in the absence of crossing or intertwining base pairs is called *regular*, and has the following definition.

Definition 6 *An RNA 2D structure M_{i_0,k_0} is regular iff $M_{i_0,k_0} = \emptyset$ or $\exists (i,j) \in M_{i_0,k_0}$ such that*

    i. *$M_{i_0,k_0} = M_{i_0,i-1} \cup M_{i+1,j-1} \cup M_{j+1,k_0} \cup (i,j)$ (no base pairs cross the partitions).*

    ii. *Each of $M_{i_0,i-1}, M_{i+1,j-1}, M_{j+1,k_0}$ is regular.*

The following two definitions describe the concept of *degrees* in pseudo-knots (see Figure 9).

Definition 7 *M_{i_0,k_0} is a simple pseudoknot iff M_{i_0,k_0} is regular or $\exists j_1, j_2 \in \mathfrak{S}^+, (i_0 \leq j_1 < j_2 \leq k_0)$ such that the resulting partition, $D_1 = [i_0, j_1 - 1], D_2 = [j_1, j_2 - 1], D_3 = [j_2, k_0]$, satisfies the following:*

    i. *$M_{i_0,k_0} = (S_L \cup S_R)$, where $S_L = \{(i,j) \in M_{i_0,k_0} | i \in D_1, j \in D_2\}$ and $S_R = \{(i,j) \in M_{i_0,k_0} | i \in D_2, j \in D_3\}$ (i.e., base-pairs crossing regular regions).*

ii. S_L and S_R are regular.

Definition 8 M_{i_0,k_0} *is a standard pseudoknot with degree d $(d \geq 3)$ iff M_{i_0,k_0} is regular or $\exists\ j_1, \ldots, j_d \in \Im^+, (i_0 \leq j_1 < \ldots < j_d \leq k_0)$ which divide $[i_0, k_0]$ into d parts, $D_1 = [i_0, j_1 - 1], D_2 = [j_1, j_2 - 1], \ldots, D_d = [j_d, k_0]$, and satisfy the following:*

i. $M_{i_0,k_0} = \bigcup_{l=1}^{d-1} S_l$, where $S_l = \{(i, j) \in M_{i_0,k_0} | i \in D_l, j \in D_{l+1}\}, 1 \leq l < d$.
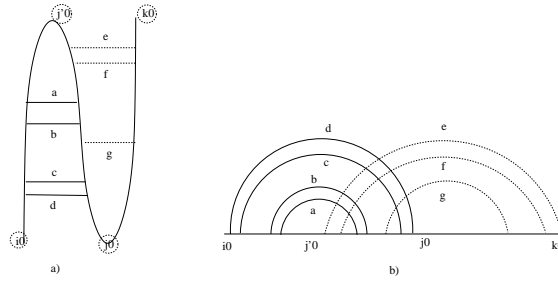
ii. S_l is regular for all $1 \leq l < d$.

Figure 9: Pseudoknot - (a) degree representation of a standard pseudoknot - (b) graphical representation of the pseudoknot where the x-axis is labeled by the primary sequence.

Note that a simple pseudoknot is a standard pseudoknot of degree 3.

The concept of degree in pseudoknots can be also extended to the more complex recursive PKs, composed themselves of smaller pseudoknotted regions [6, 28].

7.4 C++ algorithmic tests performed on RNA motifs mentioned in Section 5

———————— Motif :PDB01069 ————————
====================New Block ================
(7,5) - (7,4) - (6,7) - (5,6) - (4,5) - (4,5) -
degree of 7 is 3
degree of 4 is 3
degree of 5 is 4
—- this block represents a pseudoknot —-

==================New Block ================

(3,4) -

—- this block represents a regular-region —-

==================New Block ================

(3,1) - (2,3) - (2,3) - (1,2) -

degree of 3 is 3

degree of 2 is 3

—- this block represents a pseudoknot —-

==================New Block ================

(0,1) - (0,1) -

—- this block represents a regular-region —-

————— Summary information for Motif :PDB01069 ——————————
———

———— Total number of blocks: 4

———— number of PK blocks: 2

———— number of regular blocks : 2

————————————————————————————————————-

——————————— Motif :PKB236 ————————————

==================New Block ================

(6,0) - (5,6) - (5,6) - (4,5) - (4,5) - (3,4) - (3,4) - (2,3) - (2,3) - (1,2) - (1,2)
- (0,1) - (0,1) -

degree of 6 is 3

degree of 5 is 4

degree of 4 is 4

degree of 3 is 4

degree of 2 is 4

degree of 0 is 3

degree of 1 is 4

—- this block represents a pseudoknot —-

————— Summary information for Motif :PKB236 ——————————
—

———— Total number of blocks: 1

———— number of PK blocks: 1

———— number of regular blocks : 0

————————————————————————————————————-