A Graph-Theoretical Approach for Partitioning RNA Secondary Structures into Pseudonotted and Pseudoknot-free Regions

Louis Petingi

***Abstract*—Dual graphs have been applied to model RNA secondary structures with pseudoknots, or intertwined base pairs. In this paper we present a linear-time algorithm to partition dual graphs into topological components called blocks and determine whether 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. Applications to RNA design can be envisioned since modular building blocks with intact pseudoknots can be combined to form new constructs.**

***Index Terms*—RNAs secondary structures, pseudoknots, graph theory.**

I. Introduction

Graph theory is a field of mathematics with applications to many research areas where the objects can be represented as discrete structures called *graphs*.

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.

Waterman [18] was one of the first researchers to represent RNAs as graphs. The graph representations discussed in this paper, called dual graphs, were introduced in 2003 by Gan et. al [7], and were applied to model RNA secondary structures (2D). The 2D elements of RNA molecules consist of double-stranded (stem) regions defined 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 (later we present an alternative definition of dual graphs). Dual graphs are needed to represent pseudoknots (*PKs*), structures involving an intertwining of two-base-paired regions of the RNA. These are common elements in many biologically important RNAs.

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 PKs. Our algorithm partitions a dual graph into graph-theoretic components called *blocks* and then determines whether each block contains a pseudoknot; a block contains a pseudoknot if and only if the block has a vertex of degree 3 or more. Thus our procedure 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 [5], [19]. The latter is distinguished from the former by having nested pseudoknots within a pseudoknot. While our partitioning algorithm can detect general pseudoknots, it cannot classify them. Extensions, however, 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 III we describe our partitioning approach of a dual graph G into components $G' \subseteq G$ called blocks. In Section IV, we characterize these blocks, as either pseudoknotted or pseudoknot-free. In Section V and Section VI we illustrate algorithmic tests performed on different motifs. We summarize the findings and outline new directions in Section VI. An Appendix section includes definitions, mathematical proofs, and supporting material.

II. Background and Definitions

In 2003, Gan et. al introduced *tree* and *dual* graph-theoretic representations of RNA 2D motifs in a framework called RAG (RNA-As-Graphs) [6], [7], [8], [11]. Dual graphs can represent complex RNA secondary structures with pseudonots; 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 PKs is not well-nested, making the presence of PKs in RNA sequences more difficult to predict by the dynamic programming [3], [4] and context-free grammars standard methods [2]. Our methodology, based on topological properties of dual graphs, suggest a new way to look at the problem of detection and classification of PKs and of general RNAs.

Following (Kravchenko, 2009 [12]), we define our biological variables as follows.

Definition 1: General terms:

a. *RNA primary structure*: a sequence of linearly ordered bases $x_1, x_2, \ldots, x_r$, where $x_i \in \{A, U, C, G\}$.
b. *canonical base pair*: a base pair $(x_i, x_j) \in \{(A,U), (U,A), (C,G), (G,C), (G,U), (U,G)\}$.
c. *RNA secondary structure without pseudoknot - or regular structure, encapsulated in the region* $(i_0, \ldots, k_0)$:

Manuscript received July 19, 2016; revised July 29, 2016. This work was supported in part by PSC-CUNY Grant # 68318-00 46 from the City University of New York Research Foundation.

L. Petingi is with the Department of Computer Science, College of Staten Island, City University of New York, Staten Island, NY, 10314, USA e-mail: louis.petingi@csi.cuny.edu.

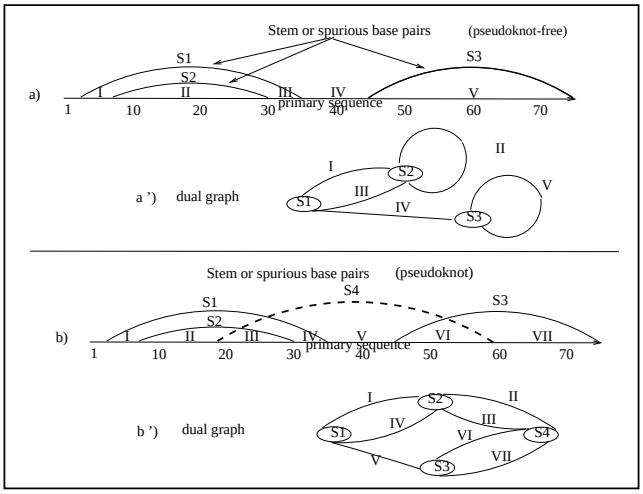

Fig. 1. *Graphical and dual graph representations of an RNA 2D structure. (a) graphical representation of a pseudoknot-free RNA primary sequence and embedded stems or base pairs; (a′) corresponding dual graph representation. (b) graphical representation of a pseudoknotted RNA 2D structure; (b′) corresponding dual graph.*

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).

d. *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.

e. *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.

f. *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-a), and a stem (Definition 1-d) 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 defined 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.

III. Graph Partitioning Algorithm

Our graph-theoretic partitioning methodology is based on identifying *articulation points* in the dual graph representation of an RNA secondary structure. An articulation point partitions a graphs into connected components, that is, its deletion disconnects a graph.

We need to define the following.

Definition 3: Connectivity

a. 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.

b. 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).

c. 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 $\kappa(G) = 1$.

d. 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.

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

Indeed, identification of articulation points allow us to identify blocks (see Fig. 2). Since a block is a maximally non-separable component, a pseudoknot cannot be then contained in two different blocks. Thus identification of blocks allows us to isolate pseudoknots (as well as pseudoknot-free blocks), without breaking their structural properties.

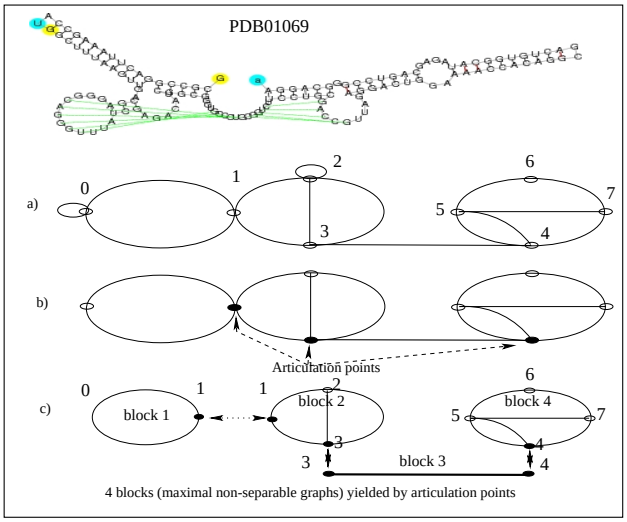

Fig. 2. *Identification of articulation points and partitioning of the dual graph corresponding to PDB01069 RNA 2D (Catalytic Ribozyme RNA) into blocks.*

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) [10]) that runs in linear computational time.

Classification of blocks as either pseudoknotted or pseudoknot-free are discussed in the next section.

IV. Classification of blocks and partitioning algorithm

The mathematical proofs of the lemmas stated in this section, are shown in Appendix B.

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,

a. Single edges.

b. Cycles.

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

From Definition 1-c, an RNA 2D structure is regular (pseudoknot-free) and encapsulated in a region $(i_0, \ldots, k_0)$,

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

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 assumption that self-loops are deleted, this definition yields the following lemma.

Lemma 1: Each block in the dual graph representation of a regular RNA 2D structure is either a bridge or a cycle of length $l, l \geq 2$.

Conversely we show the following.

Lemma 2: If an RNA 2D structure contains a pseudoknot, then its corresponding dual graph contains a block having a vertex of degree 3 or more.

Lemma 1 and Lemma 2 yield our main result as follows.

Corollary 3: 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 perform the following steps.

1. *Partition the dual graph into blocks by application of Hopcroft and Tarjan's algorithm.*
2. *Analyze each block to determine whether contains a vertex of degree at least 3. If that is the case then the block contains a pseudoknot, according to Corollary 3. If not then the block represents a pseudoknot-free structure.*

V. Illustrative examples of the partitioning algorithm

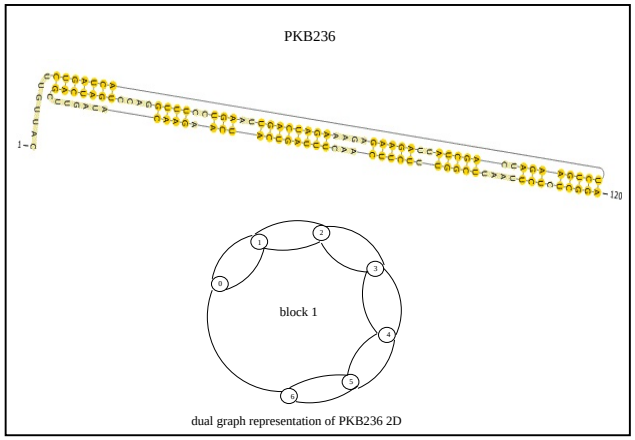

Fig. 3. *Partition of the dual graph corresponding to motif PKB236 (Regulatory Pseudoknot of the Interferon-gamma gene 5′-UTR).*

We illustrate our partitioning algorithm on the dual graph representations of two RNA 2D structures, based on the New York University's RAG database [11]. Our partitioning algorithm was implemented in C++ and runs on a Hewlett-Packard Pavilion Dv6 (2.4 GHz) notebook; the time taken for each partitioning is insignificant 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. [15]). Its dual graph is decomposed into 4 blocks as illustrated in Figure 2. According to Corollary 3, 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. 3), *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, this block is a pseudoknot.

In the next section we depict the output generated by the partitioning algorithm when tests were performed on the aforementioned RNA structures.

VI. C++ algorithmic tests performed on RNA motifs

In this section (a, b) represents an edge of a dual graph with end-vertices a and b.

The following is the output yielded by our partitioning program when tests were performed on motifs PDB01069 and PKB236.

```
———————— 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
————————————————————————-
```

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VII. Conclusions 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, providing a systematic way to classify different RNAs regions. 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.

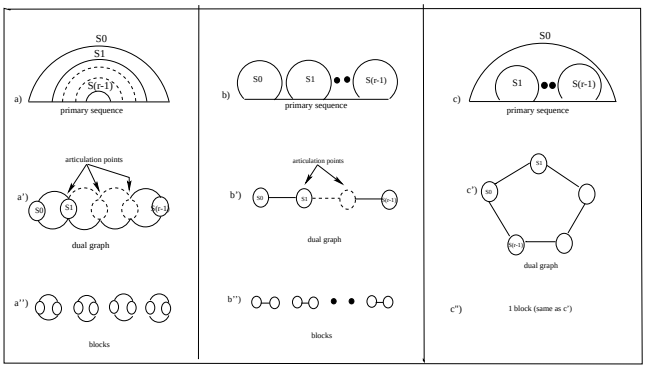

Fig. 4. *Classification of PK-free regions and graphical/dual/block representations.* (a) graphical, dual, and block representations of r nested-stems - (b) graphical, dual, and block representations of r adjacent stems - (c) graphical, dual, and block representations of a stem containing $r-1$ adjacent stems.

Appendix A
Graph Theory formulations and 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 [9].

Definition 5: General graph-theoretic terms:

a. Let $H_1.x.H_2$ represent the graph composed of two graphs, H_1, and H_2, sharing the same vertex x.
b. 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.
c. 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.
d. 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.
f. 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.

Dual graph representations of RNAs, and 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. Also it is easy to show the degree of any vertex in a dual graph is at most 4 as its corresponding stem in the graphical representation can be adjacent to at most 4 other stems.

Claim 4: The dual graph representations of RNA 2D structures and of PKs are Eulerian. In addition the degree of a vertex v is at most 4.

Appendix B
Proof of Lemmas stated in Section III

From Definition 1-c, 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 1: Each block in the dual graph representation of a regular RNA 2D structure is either a bridge or a cycle of length $l, l \geq 2$.

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. 4): (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. 4). 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 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 1, consider Figure 5 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.

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

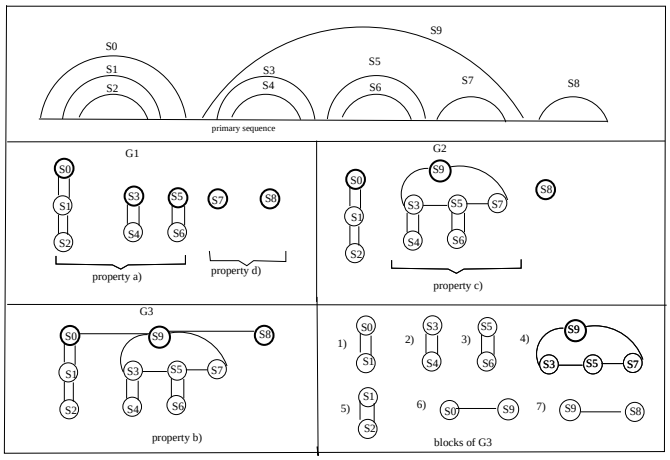

Fig. 5. *An example illustrating Lemma 1.*

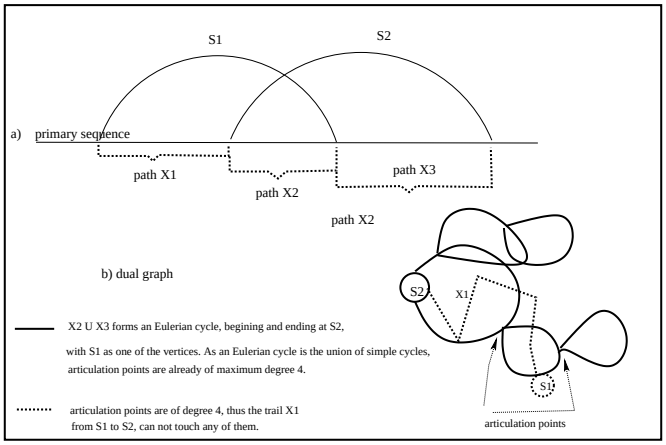

Fig. 6. *Supporting illustration for the proof of Lemma 2*

Conversely we show the following.

Lemma 2: If an RNA 2D structure contains a pseudoknot, then its corresponding dual graph contains a block having a vertex of degree 3 or more.

Proof. By Definition 1-f, 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 6-(a)), following the primary sequence of the graphical representation; these three paths correspond to trails in the dual graph representation (see Definition 5-c). We first note that $X_2 \cup X_3$ forms an Eulerian cycle G_1 in the dual graph representation (see Definition 5-e, and Claim 4), beginning and ending at S_2, having S_1 as one of its vertices. Because an Eulerian cycle is the union of simple cycles ([9], pg. 64) (Figure 6-(b)), then the articulation points of G_1 have maximum possible degree 4 (Claim 4); 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-a) 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-c) 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-d). 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. □

ACKNOWLEDGMENT

We would like to thank Tamar Schlick for providing editing suggestions to make the manuscript more suitable from a biological context. Two earlier versions of this work, [16], and [17], were archived as technical reports.

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

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

[7] 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.

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

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

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

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

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

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

[14] 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.

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

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

[17] L. Petingi, and T. Schlick, "Partitioning RNAs into pseudonotted and pseudoknot-free regions modeled as Dual Graphs", q-bio.QM,arXiv:1601.04259 (2016), http://arxiv.org/abs/1601.0425.

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

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