Graph-Theoretic Partitioning of RNAs and Classification of Frameshifting Elements in Viruses

Louis Petingi, *Member, IAENG*[*]

Abstract—**Dual graphs have been applied to model RNA secondary structures with pseudoknots, or intertwined base pairs. In previous works, a linear-time algorithm was introduced to partition dual graphs into maximally connected components called blocks and determine whether each block contains a pseudoknot or not. In this paper we present a methodology that uses our previous results and classify pseudoknots into the classical H,K,L, and M types, based upon a novel representation of RNA secondary structures as dual directed graphs (i.e., digraphs). This classification would help with the systematic analysis of RNA structures and specifically when classifying viral mRNA pseudoknots that stimulate frameshifting responsible for viral replication.**

Keywords: Graph Theory, RNA Secondary Structures, Partitioning, Bi-connectivity, Pseudoknots.

1 Introduction

In this work we are extending the results obtained in [21] in which algorithms were introduced to classify complex structures called pseudoknots, based on a novel graph-theoretic representation of RNAs; here, based on these algorithms, we specifically classify a class of viral ribosomal frameshifting signals obtained from the **PseudoBase** database from the Institute of Theoretical Biology, Leiden University, Netherlands [22].

Let $G = (V, E)$ be undirected graph composed of by a finite set of *vertices* V and a set E of unordered pairs $e = (v_1, v_2)$ of vertices. called *edges*, where each edge represents a relation between two vertices.

In 2003, Gan et al. [5] introduced dual graphs 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 can represent complex RNA structures called pseudoknots (*PKs*), which result when two base-paired regions intertwine. Pseudoknots have been associated with a diverse range of important RNA activities as for example in viral gene expression and genome replication (e.g., hepatitis C, and SARS-CoV viruses). Even though emphasis has been recently placed on viral translational initiation and elongation, the broader roles of pseudoknots are well-documented [4, 15].

In [16, 17] a linear-time partitioning algorithm was introduced based on the dual graph representation of RNA 2Ds. This algorithm partitions a dual graph into connected components called *blocks* and then determines whether each block contains a pseudoknot or is a regular region. Thus our procedure provides a systematic approach to partition an RNA 2D, into smaller classified regions, while providing a topological perspective for the analysis of RNAs.

In [18] pseudoknots were classified into two main groups: *recursive* and *non-recursive*. The former is distinguished from the latter because they contain an internal pseudoknotted or regular region that does not intertwine with external stems within the PK. In addition, if the PK is recursive, the partitioning algorithm uniquely identifies each recursive region.

In the classical literature, pseudoknots have been classified and predicted by folding algorithms into four types: H,K,L, and M [11]. Even though each type is defined in terms of how few stems intertwine, pseudoknots can be complex structures, recursive, and be comprised of several stems. In [21], we showed that based on representing RNAs as **dual directed graphs**, each PK type can be identified through a series of reductions to a unique representative. Representation of RNA secondary structures as dual directed graphs, is more precised than its dual undirected counterpart, as it is possible, as will be discussed in Section 3, that a undirected dual graph can model two or more distinct RNA 2Ds; this conflict is

[*]Manuscript received January 7, 2023 and revised in January 9, 2023.

This work was supported in part by PSC-CUNY Grant 61249-00-49 from the City University of New York Research Foundation.

Louis Petingi is a professor of Computer Science at the College of Staten Island (City University of New York), Staten Island, NY, 10314 USA (email: louis.petingi@csi.cuny.edu).

avoided by the novel representation.

The stimulatory nature of PKs in viral replication, and specifically in **frameshifting**, has been widely studied (see for example [2] and [4]). In a recent publication (2021), Bhatt et al. [3] presented a detailed study of programmed ribosomal frameshifting in translation of the SARS-CoV-2 virus. Interesting enough evidences show that the simplest H type pseudoknot (or related structures, see Section 3) is predominantly present in eukaryotic families of viral mRNAs. Related questions follow from this observation;

1. Are there families of viruses or retroviruses in which other types of PKs (i.e., K, L, M) stimulate frameshifting.
2. Are there structural differences in pseudoknots comprising the FSEs (i.e., frameshifting elements) in eukaryotic cells versus prokaryotic cells.
3. Are there viral mRNAs in which PKs are not present.

Our methodology will be able to shed some light on these outstanding questions.

In the next section, we present background material and definitions relevant to this paper, while reviewing the partitioning algorithm introduced in [16, 17], as well as its applications, as for example the development of a library of building blocks for RNA design by fragment assembly [9, 10, 24]; additionally it is shown how the partitioning algorithm can detect recursive PKs and their recursive regions. In Section 3, we discuss the algorithms presented in [21] and, in addition, a family of viral mRNA PKs obtained from a public database of RNA motifs, is classified into the different types. In Section 4 we summarize the findings and describe ongoing and future work.

2 Background

2.1 Biological and Topological Definitions

In 2003, Gan et al. [5] introduced *dual* graph-theoretic representations of RNA 2D motifs in a framework called RAG (RNA-As-Graphs) [12].

We define our biological variables as follows.

Definition 1. *General terms:*

a. *RNA primary sequence: 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* $(i_0, \ldots, k_0)$*: 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* $(x_i, x_{i+1}, \ldots, x_{i+r}, x_{j-r}, \ldots, x_{j-1}, x_j)$ *in which* (x_i, x_j), $(x_{i+1}, x_{j-1}), \ldots, (x_{i+r}, x_{j-r})$ *form base pairs.*

e. *segment region: is a tuple* $(x_i, x_{i+1}, \ldots, x_{i+r})$ *in which* (x_i, x_j) *is not a base pair whenever* $j - i \geq 1$.

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 (i.e., at least two base pairs intertwined).*

A dual graph can be easily derived from the graphical representation of an RNA 2D structure: each stem is modeled by a vertex of the dual graph, and following the primary sequence in linear order (i.e., from the 5' end to the 3' end), a segment between stems S_i and S_j is represented by an edge (S_i, S_j) in the dual graph (see Fig. 1).

In the next section we present our partitioning approach of a dual graph G, into subgraphs $G' \subseteq G$, called blocks.

2.2 Graph Partitioning Algorithm

The graph-theoretic partitioning algorithm is based on identifying *articulation points* of the dual graph representation of an RNA 2D. An articulation point is a vertex of a graph whose deletion disconnects a graph or an isolated vertex remains. 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 these block components allows us to isolate pseudoknots (as well as pseudoknot-free blocks), without breaking their structural properties.

An algorithm for identifying (bi-connected) block components in a graph was introduced by John Hopcroft and Robert Tarjan (1973, [7]), and runs in linear computational time.

A *hairpin* loop occurs when two regions of the same strand, usually complementary in nucleotide sequence when read in opposite directions, base-pair to form a double helix that ends in an unpaired loop. A self-loop in the dual graph, i.e., an edge having the same vertex as the end-points, represents a hairpin, and as it does not connect two different vertices (i.e., stems), it is formally deleted from the dual graph.

Corollary 1. *[16, 17] Given a dual graph representation of RNA 2D structure, a block represents a pseudoknot if*

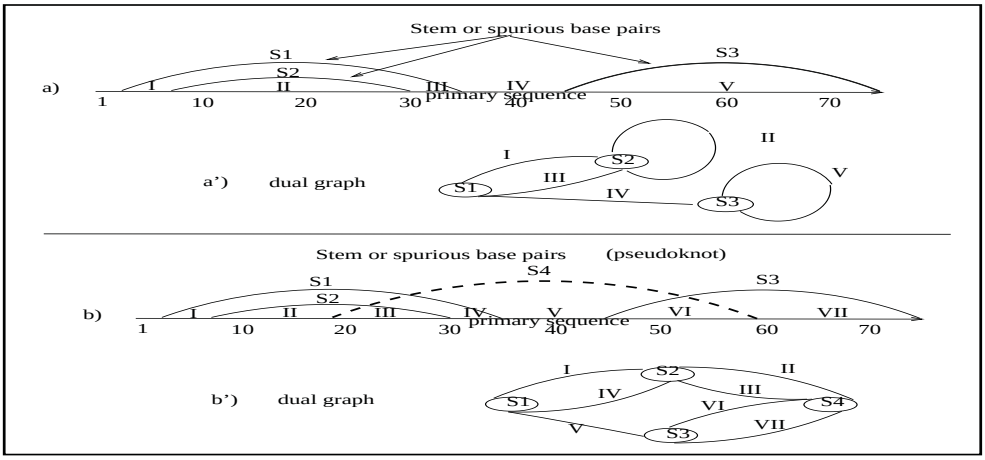

Figure 1: *Graphical and dual graph representations of an RNA 2D structure. (a) graphical representation of a pseudoknot-free RNA primary sequence and embedded stems or base pairs; (a′) corresponding dual graph representation. (b) graphical representation of a pseudoknotted RNA 2D structure; (b′) corresponding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

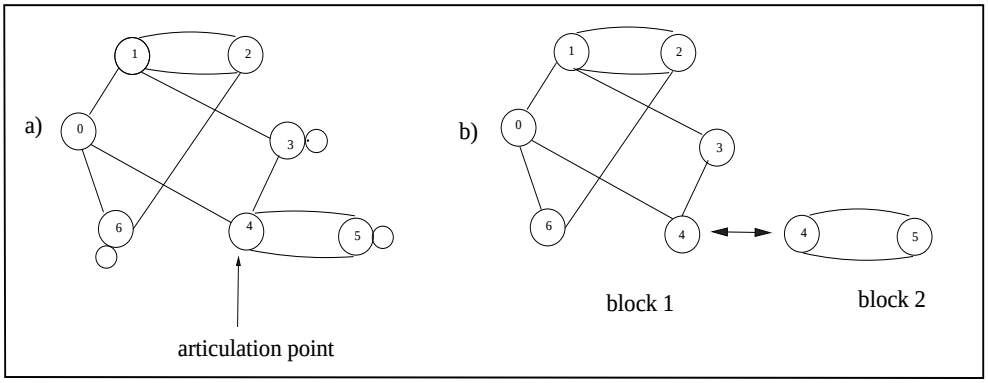

Figure 2: *Identification of a) articulation points and b) partitioning of a dual graph.*

and only if the block has a vertex of degree (Definition 1-f) at least 3 *where the degree of a vertex u is the number of edges incident at u.*

From Definition 1-c, an RNA 2D structure is a regular-region (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 $i < l < j < m$, $i_0 \leq i, j, l, m \leq m_0$, otherwise the region is a pseudoknot; this definition yields the following main result.

Corollary 1 yields the following algorithm,

Algorithm 1. Partitioning

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 1. If not then the block represents a pseudoknot-free structure.

Consider as an example the dual graph shown in Figure 2. This graph is decomposed into 2 blocks. According to Corollary 1, block 1 is a pseudoknot as it has a vertex of degree at least 3, while block 2, a cycle, corresponds to a regular region.

In the next section we extend our algorithm to classify PKs as either recursive or non-recursive; the algorithm can also identify each recursive region.

2.3 Classification of Pseudoknots as either Recursive or Non-recursive and Identification of each Recursive Region

The RNA 2D dual graph and graphical representations depicted in this section are based upon New York University's *RAG*-database [8], and *R-Chie* visualization software [13], respectively.

A recursive pseudoknot is a pseudoknot $M_{i,j}$ in a region $[i, j]$ that contains a pseudoknotted or regular region $M_{k,l}$, $i < k < l < j$, and there does not exist a base pair

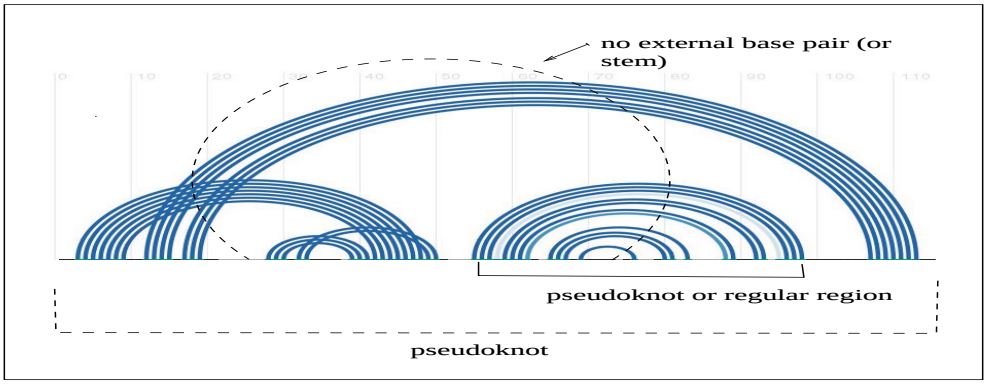

Figure 3: *Recursive pseudoknot.*

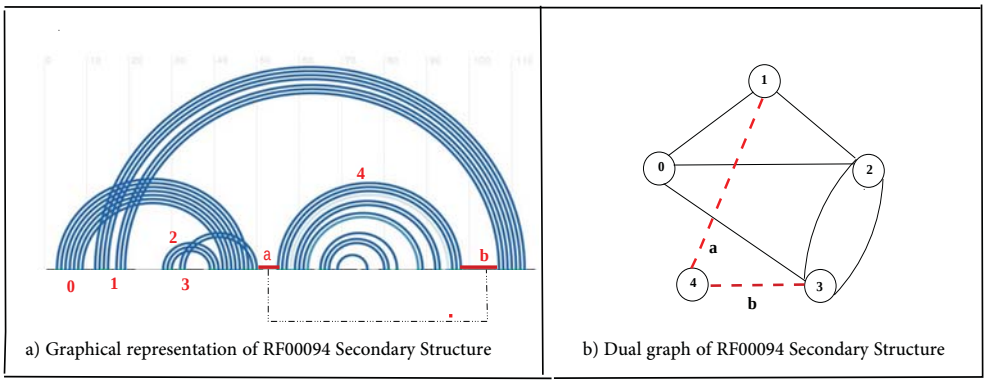

Figure 4: *Hepatitis Delta Virus Ribozyme secondary structure. a) Graphical representation. b) Dual graph representation.*

(x_c, x_d), such that x_d is a base of $M_{k,l}$, and x_c is a base of $M_{i,j}$ external to $M_{k,l}$ (see Fig. 3).

A pseudoknotted block can be classified as recursive by determining the edge-connectivity of the block. The *edge-connectivity* is defined as the minimum number of edges that if they are deleted then the resulting graph is disconnnected. As an example consider the *Hepatitis Delta Virus Ribozyme* (see Fig. 4), necessary for viral replication. The stem labeled 4 in the graphical representation (or vertex labeled 4 in the dual graph) is attached to the pseudoknot by the segments a and b in its graphical representation, or edges labeled a and b in the dual graph representation. As by deleting two edges in the dual graph, vertex labeled 4 becomes disconnected, then stem 4 is a recursive region of the PK, thus the dual graph edge-connectivity is 2. The proof of the following lemma was shown in [18].

Lemma 1. *The dual graph representation of a pseudoknotted block is recursive if and only if the block has edge-connectivity 2.*

It can be determined that a pair if edges is a disconnecting set by application of Depth-First-Search [6] in time $(|E|^3)$, allowing us to find every internal recursive region of a recursive pseudoknot, if such pair of edges exist.

As an example of a non-recursive pseudoknot consider the *Translational repression of the Escherichia coli alpha operon mRNA* ([23]), illustrated in Fig. 5. The dual graph representation of this motif 2D has edge-connectivity 3, thus it is not a recursive PK. The algorithm is written in C++ and is archived for public use [19].

3 Classification of H, K, L, M Pseudoknots Types

In the classical literature, pseudoknots have been classified and predicted by folding algorithms into four types: H,K,L, and M [11]. Even though each type is defined in terms of how few stems intertwine, pseudoknots can be complex structures, recursive, and be comprised of several stems. In a recent paper [21], we showed that based on dual directed graphs, each PK type can be identified through a series of reductions to a unique representative. This methodology will permit us to systematically analyze several motifs simultaneously and develop more precise RNA folding algorithms. The results and algo-

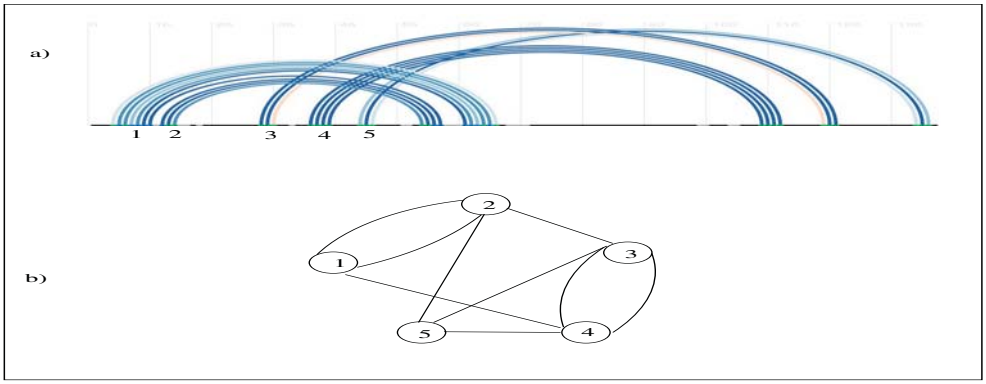

Figure 5: *Translational repression of the Escherichia coli alpha operon mRNA*. a) Graphical representation; b) Dual graph representation.

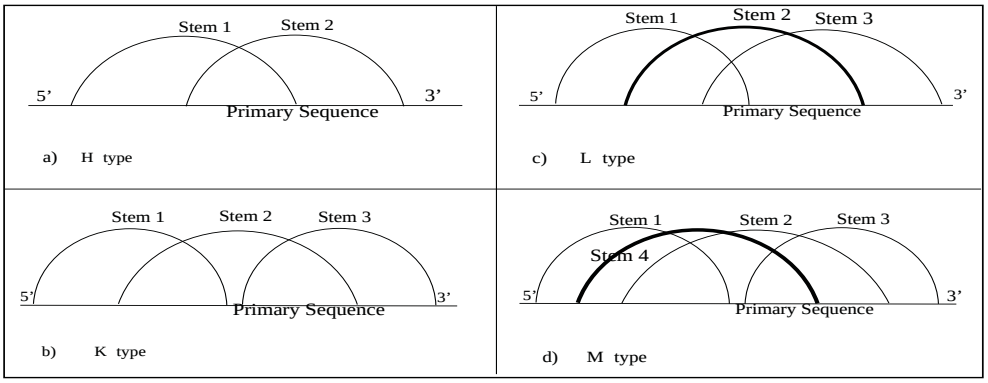

Figure 6: *Graphical representation of the H,K,L, and M types. a) H-type. b) K-type, c) L-type, and d) M-type.*

rithms discussed in Section 2.2 and Section 2.3, based upon undirected dual graphs, can be extended to directed dual graphs, as the direction of a directed edge can be ignored. Moreover, representation of RNA secondary structures based on dual directed graphs, is more precised than its dual undirected counterpart, as a dual undirected graph can model two or more distinct RNA 2Ds; this conflict is avoided by the novel representation. For example, the distinct directed graphs representing K and L types (see Fig. 7), have the same corresponding undirected dual graph, i.e., by replacing directed edges by undirected edges.

The definitions of pseudoknots follow that ones stated by Antczak et al. [1], and Kucharik et al. [11].

A H-type pseudoknot occurs when a nucleotide of a loop or bulge pairs with a nucleotide of a single-stranded region outside the loop; this type can be alternatively illustrated from the graphical representation (see Fig. 6-a) by the intertwining of two stems.

A K-type PK results when two nucleotides from different loops (or bulges) pair to form a double helical segment (see Fig. 6-b). Similarly we can also describe L and M types pseudoknots, derived from the H and K types, respectively, by addition of a stem (darker color) as shown in Figure 6-c and Figure 6-d.

Although two PKs of the same type of maybe structurally different, in [21] it is shown that based upon the representation of RNAs as dual digraphs, each motif type can be reduced to a unique digraph type-representative, by performing a sequence of topological reductions, to one of the types depicted in Fig 7; otherwise the PK is none of these types.

3.1 Application of the Algorithms to Classify a Class of Viral mRNAs

In this section we classify a class of viral ribosomal frameshifting signals obtained from the **PseudoBase** database from the Institute of Theoretical Biology, Leiden University, Netherlands [22]. Each viral FSE (frameshifting element) pseudoknot will be classified as one of the types discussed in Section 3 (i.e., H, K, L, and M types) and we also determine if the corresponding frameshifting stimulating pseudoknot is recursive or non-recursive (see Section 2.3). We analyze the frameshift-

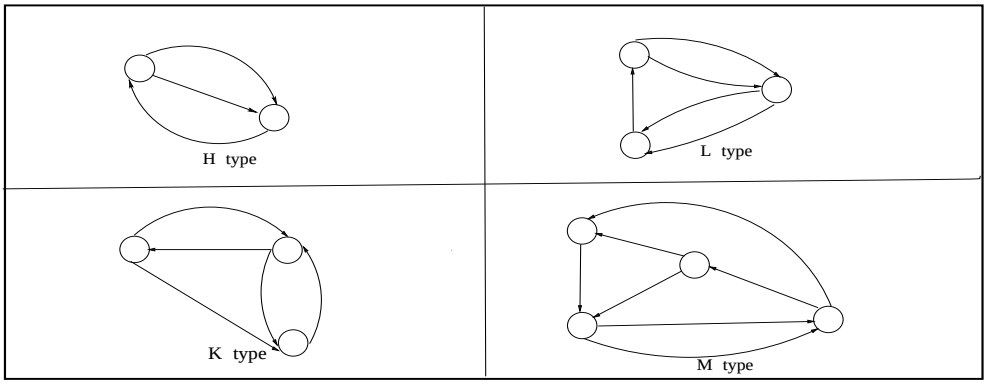

Figure 7: *By application of Reductions 1 and 2 [21], a directed graph representing a RNA motif, will be reduced to a unique representative of the 4 different types (H,K,L,M).*

Table 1: Frameshifting Stimulating Pseudoknots

Virus	**Type**	**Recursive**
Alfuy	H	No
Beet chlorosis	H	No
Bernie	H	No
Bovine leukemia	H	No
Beet western yellows	H	No
Barley yellow dwarf	H	No
Cucurbit aphit-borne	H	No
Equine arterities	H	No
Equine infectious anemia	H	No
Feline immunodeficiency	H	No
Human coronavirus 229E	K	No
Homo sapiens (Hs-Ma3)	H	No
Infectious bronchitis	H	No
Japanese encephalitis	H	Yes
West Nile (Kunijn)	H	Yes
Lactate d-e (strain C)	H	No
Middelburg	H	No
Mouse mammary tumor	H	No
Murray Valley encephalitis	H	No
Mus musculus (mouse)	H	No
Pea enation mosaic	H	No
Potatoe leafroll (S type)	H	No
Potatoe leafroll (W type)	H	No
Porcine respir. syn. (N. Ame.)	H	No
Porcine respir. syn. (Eur.)	H	No
Rous sarcoma	H	Yes
SARS-CoV	H	Yes
Seal louse	H	No
Simian retrovirus-1	H	No
Sugarcane yellow leaf	H	No
Usutu	H	No
Visna-Maedi	H	No
White Bream	H	No
West Nile	H	Yes

ing element of several viruses (see Table 1). As it was predicted, most frameshifting stimulating PKs are of the simple form (i.e., H type), however we found one K-type pseudoknot (i.e., Human Coronavirus 229E), interesting enough related to the SARS-CoV and SARS-CoV-2 strains. Regarding the latter, it is noted that although their corresponding pseudoknots are of the simplest form (i.e., H-type), they contain very distinctive recursive regions.

4 Conclusions and Ongoing Work

The Covid-19 pandemic accelerated the study of viral replication and the need to develop therapeutics to control infectivity. We suggest that pseudoknots not only play a significant role, but a predominant one in viral transmissibility for most viruses, and our proposed techniques aim to shed some light in this area, as well to better understand the roles of PKs in general RNA functionality. Evidences show that the simplest H-type pseudoknot (or related structures) are predominantly present in eukaryotic families of viral mRNAs, however as it is shown in Table 1, other types of pseudoknots stimulate frameshifting; this analysis opens the door for determining structural properties of PKs as for example what roles recursive fragments play in these structures. Our proposed techniques could make substantial progress on this area of research.

References

[1] 7 Antczak, M., Popenda, M., Zok, T., Zurkowski, M., Adamiak, R.-W, and Szachniuk, M.: New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. Bioinformatics **34** (8): 1304-1312 (2018).

[2] Barends, S., Rudinger-Thirion, J., Florentz, C., Giegé, R., Pleij, C., Kraal, B.: tRNA-Like Structure

Regulates Translation of Brome Mosaic Virus RNA. J Virol. **78** (8): 4003–4010 (2004).

[3] Bhatt et al.: Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome, Science **372**, 1306–1313 (2021).

[4] Brierley, I., Pennell, S., Gilbert, R.-J.: Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nat Rev. Microbiol. **5** (8): 598-610 (2007).

[5] Gan, H.-H., Pasquali, S., Schlick, T.: Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucleic Acids Res. **31**(11), 2926–2943 (2003).

[6] Harary, F.:Graph Theory, Addison-Wesley, Mass. 1969.

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

[8] Izzo, J.-A., Kim, N., Elmetwaly, S., Schlick, T.: RAG: an update to the RNA-As-Graphs resource. BMC Bioinformatics **12**, 219 (2011).

[9] Jain, S., Bayrak, C.-S., Petingi, L., Schlick, T.: Dual Graph Partitioning Highlights a Small Group of Pseudoknot-Containing RNA Submotifs. Genes **9**(8), 371 (2018).

[10] Jain, S., Saju, S., Petingi, L., Schlick, T.: An extended dual graph library and partitioning algorithm applicable to pseudoknotted RNA structures, Methods **162**, 74–84 (2019).

[11] Kucharik, M., Hofacker, I.-L., Stadler, P.-F., Qin, J.: Pseudoknots in RNA folding landscapes, Bioinformatics **32**(2), 187–194 (2016).

[12] Kim, N., Petingi, L., Schlick, T.: Network Theory Tools for RNA Modeling. WSEAS Transactions on Math. **12**(9), 941–955 (2013).

[13] Lai, D., Proctor, J.-R., Zhu, J.-Y., Meyer, I.-M.:R-chie : a web server and R package for visualizing RNA secondary structures, Nucleic Acids Research **40**(12), e95 (2012). e-RNA Website, url - https://www.e-rna.org/r-chie/rfam.cgi. Last accessed 4 January 2019.

[14] Livieratos, I.-C. et al.: Analysis of the RNA of Potato yellow vein virus: evidence for a tripartite genome and conserved 3'-terminal structures among members of the genus Crinvirus, J. Gen. Virol. **85**, 2065–2075 (2004). e-RNA Website, url - https://www.e-rna.org/r-chie/rfam.cgi. Last accessed 4 January 2019.

[15] Pesells, A., Serganov, A.: Structure and function of pseudknots involved in gene expression control. Wiley Interdiscip. Rev. RNA **5**(6), 803–822 (2014).

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

[17] Petingi, L., Schlick, T.: Partitioning and Classification of RNA Secondary Structures into Pseudonotted and Pseudoknot-free Regions Using a Graph-Theoretical Approach, IAENG International Journal of Computer Science, **44**(2), pp 241-246, 2017.

[18] Petingi, L., Schlick, T.: Graph-Theoretic Partitioning of RNAs and Classification of Pseudoknots. In: Holmes I., Martín-Vide C., Vega-Rodríguez M. (eds) Algorithms for Computational Biology. AlCoB 2019. Lecture Notes in Computer Science, **11488**, Springer, Cham.

[19] Petingi, L.: Dual Graph Partitioning Code. url - https://github.com/Louis-Petingi/Partition-Algorithm-3. Last accesed 3 January, 2019.

[20] Petingi L: Graph-Theoretic Partitioning of RNAs and Classification of Pseudoknots-II. CUNY Academic Works. Internal Report (July 2021). Also in arXiv:2109.03236 [q-bio.BM].

[21] Petingi L.: Graph-Theoretic Partitioning of RNAs and Classification of Pseudoknots-II, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2022, 6-8 July, 2022, London, U.K., pp 124-130.

[22] PseudoBase database: Institute of Theoretical Biology, Leiden University, Netherlands. Webpage: https://www.ekevanbatenburg.nl

[23] Schlax, P-J., Xavier, K.-A., Gluick, T.-C., Draper D-E.: Translational repression of the Escherichia coli alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. J Biol Chem. **276** (42), 38494–38501 (2001).

[24] Zhu, Q., Petingi, L., Schlick, T. RNA-As-Graphs Motif Atlas—Dual Graph Library of RNA Modules and Viral Frameshifting-Element Applications. Int. J. Mol. Sci. (2022), **23**(16), 9249. https://doi.org/10.3390/ijms23169249