Research Article **Open Access**

Adane L. Mamuye*, Matteo Rucco, Luca Tesei, and Emanuela Merelli

Persistent Homology Analysis of RNA

DOI 10.1515/mlbmb-2016-0002
Received September 14, 2016; accepted November 20, 2016

Abstract: Topological data analysis has been recently used to extract meaningful information from biomolecules. Here we introduce the application of persistent homology, a topological data analysis tool, for computing persistent features (loops) of the RNA folding space. The scaffold of the RNA folding space is a complex graph from which the global features are extracted by completing the graph to a simplicial complex via the notion of clique and Vietoris-Rips complexes. The resulting simplicial complexes are characterised in terms of topological invariants, such as the number of holes in any dimension, i.e. Betti numbers. Our approach discovers persistent structural features, which are the set of smallest components to which the RNA folding space can be reduced. Thanks to this discovery, which in terms of data mining can be considered as a space dimension reduction, it is possible to extract a new insight that is crucial for understanding the mechanism of the RNA folding towards the optimal secondary structure. This structure is composed by the components discovered during the reduction step of the RNA folding space and is characterized by minimum free energy.

Keywords: RNA folding space; persistent homology; persistent structural features

1 Background

Ribonucleic acid (RNA) is a biological molecule that plays a key role in various biological processes. Discoveries of the past decade witnessed that RNA is involved in catalytic activity, protein synthesis and gene regulation [22]. Understanding the recurrent interactions among RNA molecular components and their structures are vital to detect prominent information and insight on the RNA mechanism behind its roles. Base-pairing is the most specific interaction in RNA as it involves the hydrogen-bonding of nucleotides. Such interactions are fundamental for understanding RNA folding, function and evolution. RNA secondary structures are defined by list of base pairs in which each base pair appears at most once [10]. Understanding secondary structures is important for inferring structure-function relationships.

RNA secondary structure can be determined by experimental techniques, such as X-ray crystallography and NMR. However, they are time-consuming, expensive and in some cases infeasible [16]; thus, in the last three decades numerous computational methods, such as comparative sequence analysis and dynamic programming based on thermodynamic models, have been established for dealing with the prediction of secondary structures. Though comparative sequence analysis is the most reliable approach to predict RNA secondary structures, it requires many multiple homologous sequences and is labour intensive [20]. Thermodynamics based approaches can also be less accurate than comparative-based algorithms [16] and are computationally expensive. As a result, improving the accuracy of predicting RNA secondary structure remains an

***Corresponding Author: Adane L. Mamuye:** School of Science and Technology, Computer Science Division, University of Camerino, 62032, Camerino (MC), Italy, E-mail: adaneletta.mamuye@unicam.it
Luca Tesei, Emanuela Merelli: School of Science and Technology, Computer Science Division, University of Camerino, 62032, Camerino (MC), Italy, E-mail: emanuela.merelli@unicam.it, luca.tesei@unicam.it
Matteo Rucco: School of Science and Technology, Computer Science Division, University of Camerino, 62032, Camerino (MC), Italy and Italian National Council of Research, Institute of Applied Mathematics and Information Technologies, 16149, Genova (Ge), Italy, E-mail: matteo.rucco@unicam.it

ongoing challenge in computational biology. Moreover, it is still an open question to what extent the final structure assumed by RNA is determined by the minimal free energy with respect to the kinetic folding [12].

Biomolecules are typically characterised by a finite, usually large, number of non-identical interacting entities that often are complex systems. In terms of data volume, such molecules can be associated with big datasets. To extract useful information from these data, a suite of new techniques called topological data analysis (TDA) has recently shown a lot of potentiality in discovering important features that traditional techniques could not find [6, 7]. TDA, which comprises of a set of techniques inspired from algebraic topology, is a useful approach for analysing multidimensional complex data. Persistent homology appears as a fundamental tool in TDA and is used in a variety of disciplines [8, 9, 21, 26], including biomolecules [5, 13, 30–32], for the topological simplification of complex data at a different spatial resolution.

Xia et al. [30] presented a multi-resolution persistent homology analysis to reveal the intrinsic topological invariants of DNA and RNA molecules. Additionally, persistent homology has been proposed for studying evolution by considering a set of genomes and calculating the genetic distance between each pair of sequences [9]. Finally, persistent homology has been employed to analyse the RNA suboptimal structure space and the secondary structure space of 5s rRNA [19]. In addition to persistent homology, Mapper, another algorithm within TDA, has been used for identifying the dominant states, clusters, in the folding and unfolding pathways of RNA hairpin [4].

The number of possible secondary structures of a sequence, in the folding space, grows exponentially with its sequence length [29] and corresponds to a multidimensional space [14]. This paper proposes a new method for analysing the RNA folding multidimensional space by using persistent homology. Previously, topology was introduced for RNA structures by Penner and Waterman [23]. Then combinatorial topology was used to construct topological shapes from secondary structures; Bon et al. [3] contributed with a novel topological classification of RNA pseudoknots based on topological genus; Reidys et al. [25] presented a pseudoknot prediction algorithm based on topological shapes. In these two last papers, the way of building a topological space from RNA pseudoknot structure was obtained via the notion of fatgraphs. However, to the best of our knowledge, no attempt has been made to use persistent homology to analyse the RNA folding space.

The remainder of the paper is organised as follows. In Section 2 we discuss the basics of RNA secondary structure. Section 3 illustrates the methodology followed in our study. In Section 4 we explain how our approach is tested on RNA sequence data while conclusion and future work are given in Section 5.

2 RNA Secondary Structure

RNA is a single strand biological molecule that consists of a sequence of nucleotides, adenine (**A**), guanine (**G**), cytosine (**C**) and uracil (**U**), connected by a phosphodiester bond. The sequence of nucleotides, called the primary structure or the backbone, folds back on itself. This folding is due to the formation of hydrogen bonds between two non-neighbour nucleotides. In particular, each nucleotide of the primary structure can form a base pair by interacting with at most one other nucleotide; such pairing is due to Watson-Crick bases pairs (**G-C** and **A-U**) and wobble base pair (**G-U**). These base pairs, called canonical base pairs, determine the RNA secondary structure.

Each RNA secondary structure can be of two types: pseudoknot free or pseudoknotted. A pseudoknot free structure is composed of a set of non-crossing-serial interactions, while a pseudoknotted one is formed by crossing-serial interactions, see Figure 1. An RNA secondary structure can be composed of several loops (structural elements) namely *helix*: double helical region consistsing of consecutive base pairs without unpaired nucleotides in between, *hairpin*: sequence of unpaired nucleotides enclosed by a single base pair, *internal loop*: loops with two base pairs and two regions of unpaired nucleotides, *bulge*: internal loop with one or more unpaired nucleotides on one side of the duplex and *multi-branched loop*: join point for three or more helices that can also contain unpaired nucleotides in the loop, as illustrated in Figure 1. Each loop is

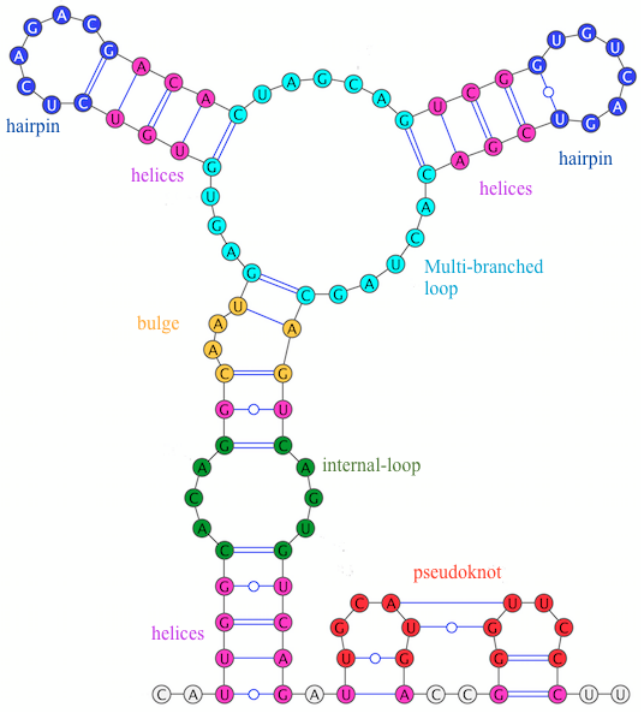

Figure 1: Structural elements: the five structural elements of RNA structure, namely helix (pink), hairpin(blue), bulge(orange), internal-loop (green) and multi-branched loop (cyan) together with the pseudoknot(red).

characterised by its length, i.e. the number of unpaired nucleotides in the loop, and its degree, given by the number of base pairs delimiting the loop (including the closing pair) [15].

3 Methodology

Techniques inspired by topology are important to understand large and complex datasets and to derive useful knowledge from them. In this section we report on how to extract significant structural features from the RNA folding space by using a data driven approach, i.e. topological data analysis.

3.1 From Graph to Weighted Graph

The RNA folding space can be represented as a complex graph. In a such graph the nodes are the nucleotides and the edges are the phosphate backbone and all the possible base pairs linked by hydrogen bonds. Under this representation, each structure belonging to the entire ensemble of the RNA folding space is represented by a sub-graph of the complex graph. Then, the whole graph can be considered as the scaffold of the RNA folding space. To compute the edge weights of the base pairs, the RNA base-pairing probabilities are taken into account. We use the *RNAstructure* web server to predict the base-pair probabilities with a partition function calculation [1]. It gives as output the base pair probability for each pair. We use these probability values weights of the base pair interactions. An example of weighted of the backbone interactions are considered as a fixed value that is adjusted at the time of analysis. The weighted graph constructed in this way is depicted in Figure 2.

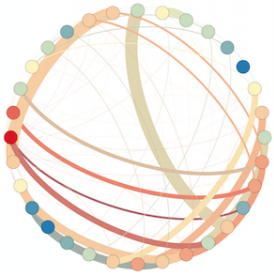

Figure 2: Weighted graph: the thickness of each edge represents the corresponding base pair probability between two nodes (nucleotides).

3.2 From Weighted Graph to Simplicial Complex

The weighted graph is the scaffold of the folding space from which a simplicial complex representation is derived. A simplicial complex is a simple combinatorial object in which simplices, namely: *vertex* for 0-simplex; *edge* for 1-simplex; *triangle* for 2-simplex; *tetrahedron* for 3-simplex (and so on), are nested together to represent topological spaces. In general, a simplicial complex K needs to fulfil two conditions: the intersection between two simplices must be the empty set or it must be equal to a facet (for example the edge or a vertex of a triangle) with dimension less than the dimensions of the intersecting simplices. The dimension of a simplicial complex is equal to the dimension of its biggest simplices.

Given a weighted graph, there are different ways to construct a filtered simplicial complex out of it. In this work, we use the clique weight rank persistent homology procedure as implemented in the jHoles software [1] [2], see Figure 3. The clique of dimension k (k-clique) is composed of all the maximum cliques (m-cliques) defined on m <k nodes. Therefore, a k-clique can be seen as a k-1 simplex [2].

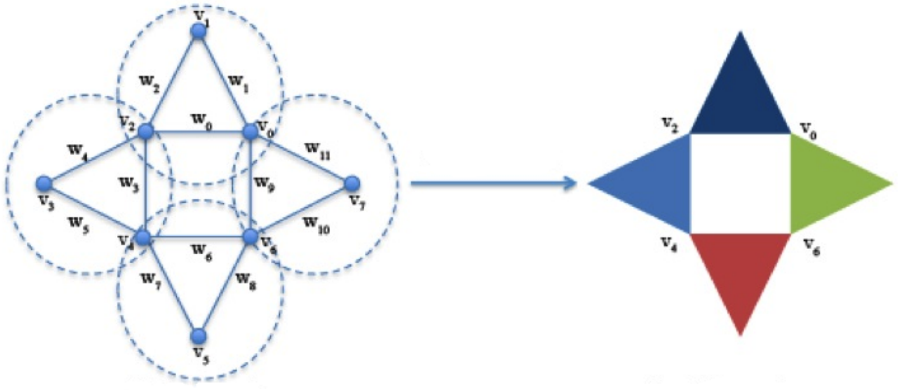

Figure 3: jHoles simplical complex construction: on the left, a weighted undirected graph with eight vertices and twelve edges. The four maximal cliques are highlighted by dashed circles. On the right, a simiplicial complex corresponding to the weighted graph [21].

We can also construct an abstract simplicial complex from the adjacency matrix of an undirected graph by using the Vieoris-Rips simplex construction under the assumption that the adjacency matrix represents a pair-wise distance matrix. In Vietoris-Rips two points are connected if the proximity distance between them is at most a fixed distance ϵ. Given a metric space X and a proximity parameter $\epsilon > 0$, the Vietoris-Rips

1 jHoles is a Java open source software that can be downloaded at http://www.jholes.eu

simplicial complex $(X; \epsilon)$ has finite subsets of X of diameter less than ϵ as its simplices. Intuitively, and by looking at Figure 4, we surround each point by a sphere, then we blow-up each sphere simoultanously up to ϵ and we highlight the intersections among spheres. We add a k-simplex any time we see a subset of k points with common intersection. Vietoris-Rips filtration is an increasing sequence of simplical complexes as shown in Figure 4. We use JavaPlex [27] to construct the Vietoris-Rips complex over the adjacency matrix.

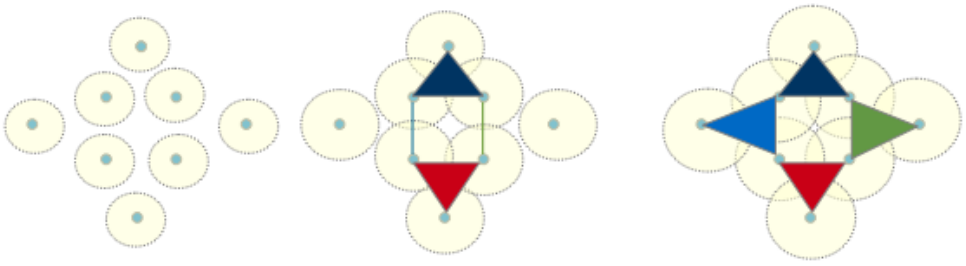

Figure 4: Vietoris-Rips simplical complex construction. From left to right: filtration by increasing the proximity distance (the radii of the spheres).

3.3 Persistent Homology on Filtered Simplicial Complex

TDA measures the features of a topological space based on topological invariants. Topological invariants, algebraic objects which are invariant under homeomorphisms, are more suitable to analyse the topological space [6, 11]. Persistent homology, one of the topological invariants, can be computed through an incremental algorithm. It takes a collection of filtered simplices as input. Then it iterates over the set of the filter values and, at each iteration, it adds the corresponding simplices. Moreover, it computes the homological groups of the newly formed topological space. The dimensions of the homological groups are called Betti numbers. Betti numbers are special counters for n-dimensional holes [6]. Persistent homology measures the lifespan of the topological space: the features that are present after the last iteration are called persistent, otherwise they are classified as irrelevant.

To apply persistent homology on filtered simplicial complexes, we use the Java software suites named jHoles and JavaPlex. jHoles implements the clique weight rank persistent homology algorithm to recover accurate long-range information from the weighted graph [2]. jHoles persistent homology engine is jPHEngine (an optimised fork of JavaPlex). jHoles categorises the set of weights of the input graph and then computes all the maximal cliques by using the Bron-Kerbosch algorithm. Then it ranks each clique with the index corresponding to the minimum (or maximum) value among the weights of the edges within a clique. Because each k-clique is equivalent to a k-1 simplex, jHoles manages each ranked clique as a filtered simplex. The collection of filtered simplices is then analysed by persistent homology. Persistent homology spans over the sequence of filter values and for each filter it adds the corresponding filtered simplices. At each step the algorithm computes the Betti numbers and lists the generators of the topological features. A generator is a simplex involved in the topological loop. For example a $2D$ hole, that is listed in the homological group H_1, is formed by a family of 1-simplices arranged in a circular motif. Moreover, each 1-simplex is formed by a pair of vertices that are 0-simplices. The 0-simplices are the generators of a 1-simplex, which are the generator of the 2D hole. Recall that Vietoris-Rips takes into account the proximity distance. In this work, proximity is given by the probability of pairing of the nucleotides, the higher is the probability the higher is the proximity. The identified generators are those holes (i.e. loops) that persist over the filtration process.

In our setting we focus on the generators of the homological group H_1 because the RNA folding space can be characterised by the value of Betti number 1 (β_1), which counts the size of H_1. In general, the value of β_1 indicates the number of 1- dimensional holes, which in this settings correspond to the loops of the

RNA secondary structure. These loops are defined as cycles of 4 or more edges with nucleotides as vertices. Nowadays, it is still not clear the role played by higher dimensional topological invariants in the context of the RNA folding space, thus we focus our attention to the homological generators of β_1 and, possibly in future work, to the ones of β_n for higher n. By exploiting the persistent loops, we can extract new insights characterizing the **helices**, **hairpin loops**, **bulges**, **internal loops** and **multi-branched loops**, as shown in Figure 5.

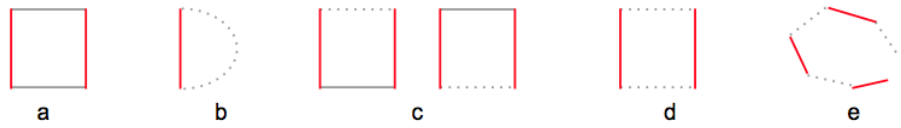

Figure 5: RNA structural elements. Labels from $(a - e)$ show helix, hairpin, bulge, internal and multi-branched loops, respectively. The broken lines, from $b - d$, indicate the existence of one or more unpaired nucleotides. In case of hairpin (b) the number of unpaired nucleotides should be ≥ 3 and in case of multi-branched loops (e) the number of base pairs should be ≥ 3 while the unpaired nucleotides can be zero or more.

4 Result and Discussion

In this study we aim at understanding the topological features of the RNA folding space by using persistent homology. To demonstrate the validity of our approach, we perform Vietoris-Rips filtration (using JavaPlex [27]) and clique weight rank persistent homology filtration (using jHoles [2]) on a short (32 nucleotides long) Homo sapiens ncRNA (URS00001625D1) sequence, taken from the RNAcenteral database [28]. From the analysis of these generators we can deduce that persistent loops are classified into six equivalences classes, namely, the five RNA structural elements and irrelevant loops. The persistent loop is called irrelevant in two cases (1) when one nucleotide is involved in more than one base pair (no more than two edges can depart from the same vertex) on a single persistent loop; and (2) whenever there is a cross-serial interaction in a loop.

To analyse the folding space using Vietoris-Rips persistent homology filtration, we construct a half matrix from the *RNAstructure* output. The half matrix entry at (i, j) indicates the base pair probability (pairwise distances) of nucleotides i and j. Additionally, for all backbone interactions the highest base pair probability is assigned. This matrix representation suffices to realize the Vietoris-Rips filtration. Differently, in the case of clique weight rank persistent homology, the filtration depends on the weight of each interaction.

Table 1 shows the results obtained by using JavaPlex and jHoles. Using both tools a total of 63 persistent loops are generated from the folding space; we found 14 and 24 irrelevant loops in jHoles and JavaPlex, respectively. The number of irrelevant loops in jHoles are less than those in JavaPlex. From our results we also understood that jHoles is more faithfully while analysing the folding space than JavaPlex. This is due to the fact that jHoles gives more emphasis to persistent loops that are generated from high probable base pairs and persist forever than JavaPlex. Persistent loops that are generated as *helices* are 2 in both cases; however, only 1 loop is similar to each other. Concerning *bulges* and *hairpins*, jHoles generates 4 bulges and 43 hairpins while JavaPlex generates 5 bulges and 32 hairpins. All the 4 *bulges* generated by jHoles are also generated by JavaPlex and 27 hairpins are structurally similar each other. The majority of the irrelevant loops in the two cases are due to the existence of cross-serial interactions in a single loop. The RNA primary structure contains many structural elements that have the potential to form the possible RNA secondary structures; however, in the predicted Minimum Free Energy (MFE) structure, highly probable base pairs are shown to be the pairs most likely present in the known structures contained in a database of diverse RNA sequences [20]. In order to get persistent loops generated from high probable base paris, we can further perform persistent homology filtration, using jHoles, over the folding space by ignoring low probable base pairs and adjusting the weight of the backbone interactions. By ignoring the base-pairs with probability less than 0.20 and adjusting the

Table 1: Persistent homology filtration over the Homo sapiens ncRNA (URS00001625D1) sequence using jHoles and JavaPlex.

	Helicies	Bulges	Hairpins	Irrelevant Loops	Total
jHoles	2	4	43	14	63
JavaPlex	2	5	32	24	63

backbone interactions weight to 0.40, we identified 12 persistent loops, i.e., 6 helices, 2 hairpins and 4 irrelevant loops. To prove the existence of persistent loops on the MFE structure, we generated the minimum free energy secondary structure of the given sequence using the Fold web server [1]. As shown in Figure 6, all the structural elements (5 helices and 2 hairpins) that make up the MFE secondary structure are identified by topologically characterising the reduced RNA folding space.

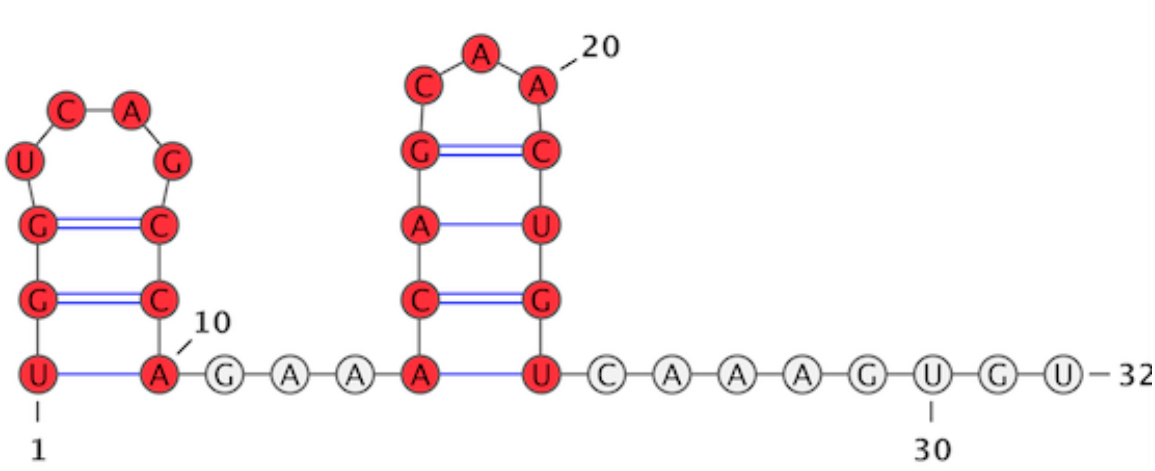

Figure 6: Persistent structural elements: all the red coloured substructures (5 helices and 2 hairpins) are identified by the homology filtration of the reduced partition function landscape.

To generate multi-branched loops the effect of such systematic adjustment is visible on longer RNA sequences; for instance, a 78 nucleotides long sequence of Homo sapiens 5S rRNA (URS0000690F72), taken from the RNAcenteral database, is analysed by ignoring base pair probabilities less than 0.01 and by adjusting the backbone interaction probability to 0.40. This results in 40 persistent loops, namely 18 helices, 2 hairpins, 1 bulge, 1 internal-loop, 1 multi-branched loop and 17 irrelevant loops, as shown in Table 2. This adjustment helps identifying all the structural elements that are the basic components of the MFE structure of the given sequence. 18 helices, 2 hairpins, 1 internal-loop and 1 multi branched persistent loops match with the structural elements of the MFE structure.

Table 2: Persistent homology filtration over Homo sapiens 5S rRNA (URS0000690F72) sequence using jHoles.

	Helicies	Bulges	Hairpins	Internal-loop	Multi-branched loop	Irrelevant Loops	Total
jHoles	18	1	2	1	1	17	40

Furthermore, 4 randomly selected sequences from the RNAcenteral database are analysed using our proposed approach with the base pair probability threshold of 0.20, and backbone weight threshold of 0.40, see Table 3. The adjusted values are good enough to identify all the structural elements of 64 and 121 nucleotide sequences, but 1, 6 and 30 loops are missed from sequences of length 152, 252 and 707, respectively. This

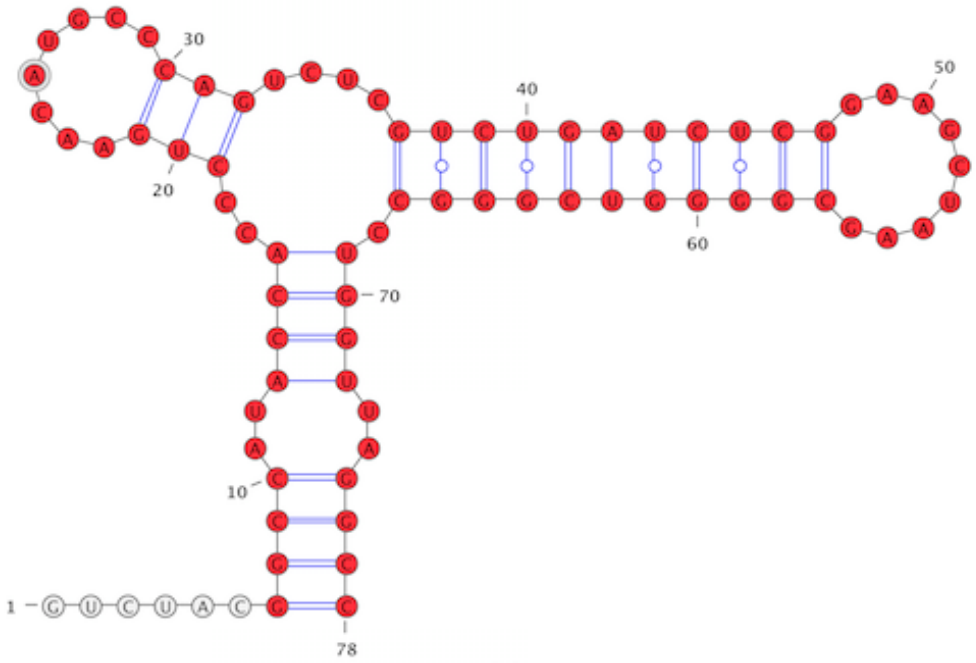

Figure 7: Persistent structural elements: all the red coloured substructures (18 helices, 2 hairpins, 1 internal-loop and 1 multi-branched loops) are identified by persistent homology filtration.

suggests that the accuracy of the topological analysis of the folding space is affected by the length of the given nucleotide sequence.

Table 3: Persistent homology filtration over randomly selected sequences.

Name	Number of nucleotides	B_1	Loops missed
URS0000637866	64	17	0
URS000071AD88	121	51	0
URS00006848FC	152	55	1
URS000006CFF3	252	101	6
URS00006B81E3_9606	707	254	30

We also checked the correlation between the number of persistent loops missing from the optimal structure and the sequence length. For this task we analysed RNA sequences using the web servers Fold [1] and MaxExpect [20]. The selected 40 sequences range from a length of 20 to 400 nucleotides with an increasing step of length 10. They are analysed by ignoring the base pairs with probability less than 0.20 and adjusting the backbone interactions weight to 0.40. The result, depicted in Figure 8, show that when the sequence length increases the number of structural elements missed by TDA on the MFE structures, as predicted by Fold, increases as well, but with discrepancies.

We also used the MaxExpect web server that provides a better estimation for an RNA secondary structure by maximizing the expected accuracy of base pair probabilities [1, 17]. It is an alternative method for structure prediction that may have higher fidelity in structure prediction. It outperforms free energy minimisation since maximum expected accuracy structures include similar number of base-pairs with big pairing probability but fewer base-pairs with low pairing probability than MFE structure [20]. We compared the persistent homology analysis results of the same dataset with the secondary structures predicted by MaxExpect. The analysis result, shown in Figure 8, is better than the previous one. The comparison of the Fold and MaxExpect web servers with the TDA persistent loop finding is presented in Figure 9. It can be observed that the

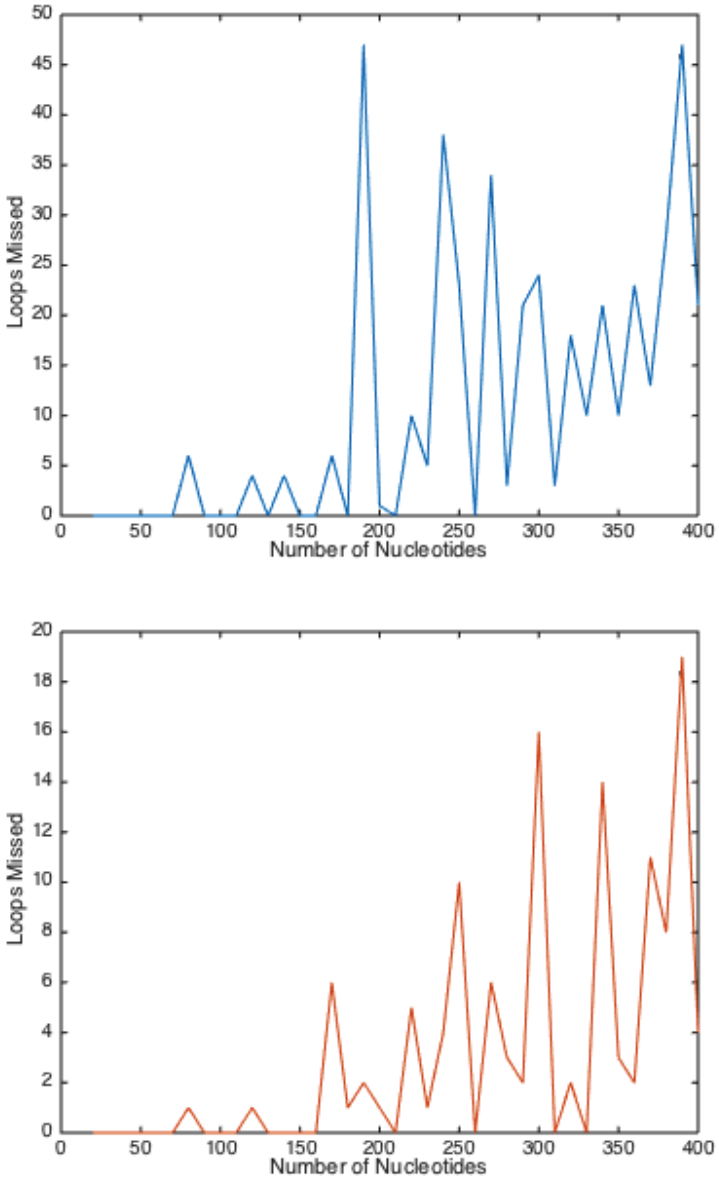

Figure 8: The TDA analysis result over the MFE structures (right) and maximum expected accuracy of the selected sequences (left). The maximum number of loops missed on the MFE structures is 50 while 20 loops are missed in the maximum expected accuracy structures.

proposed TDA approach provides structural elements of optimal secondary structure (for shorter sequences) or near optimal secondary structure (for longer sequences) more similar to those generated by MaxExpect than those generated by Fold.

Additionally, we estimated the underlying linear relationship between the RNA sequence length and the number of persistent loops missed, by adopting regression techniques. For the sake of clarity, we used a linear regression model namely the *least squares method*, as shown in Figure 10. The relationship between loops missed and RNA sequence length is described by a best fit regression line. To check how well the model can predict the data, we calculate the coefficient of determination, $\mathcal{R}^2$. The higher the value of the $\mathcal{R}^2$ the better is the model in predicting the near optimal RNA secondary structure. According to Figure 10, the slop and y-intercept regression lines coefficient of determinations between the RNA sequence length and the loops

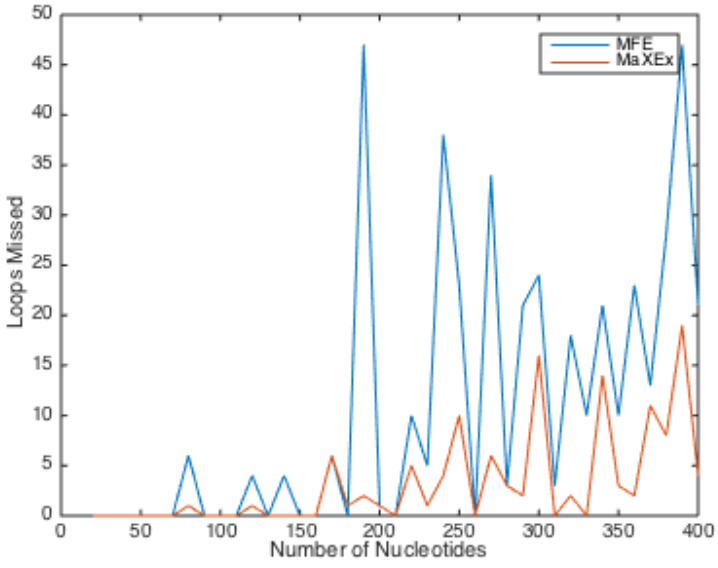

Figure 9: Comparison of the secondary structure prediction performed by TDA approach with respect to the structures present in the Fold and the MaxExpect web servers.

missed are 0.30211 and 0.3452, respectively. This implies that the second fit ($\Re^2 = 0.3452$) that includes a y-intercept yields a better coefficient of determination than the first one ($\Re^2 = 0.30211$). Additionally, the correlation coefficient, i.e. 0.5875, suggests that there is a moderate positive linear relationship between the RNA sequence length and the number of loops missed by TDA. The probability P=0.001 is smaller than the default significant level (0.05); thus, the corresponding correlation is considered as significant. Moreover, the correlation coefficient between the Fold and MaxExpect analysis result is 0.6562.

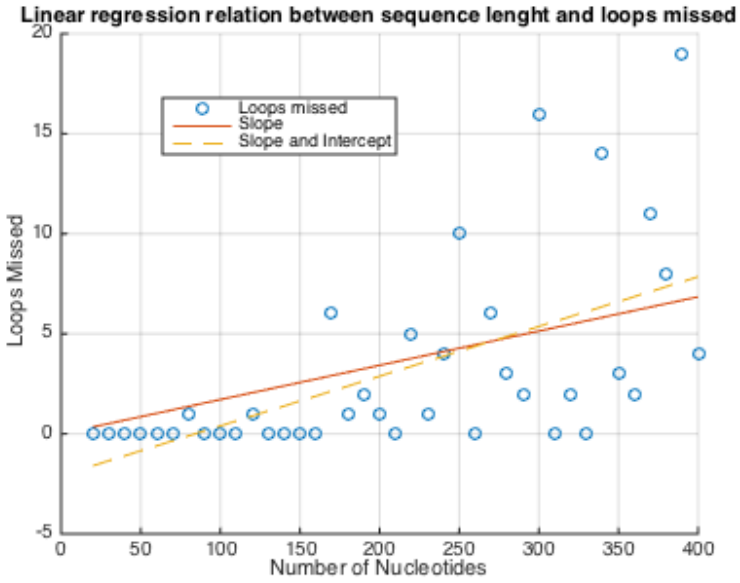

Figure 10: The slope and y-intercept regression lines of the TDA analysis and the MaxExpect optimal structure.

The resulted linear regression line, $y = 0.02485x - 2.09$ where 0.02485 is the slope and -2.09 is the intercept, can predict the number of loops that may be missed by our TDA approach on a given RNA sequence.

5 Conclusion

In addition to protein encoding, diverse classes of non-protein-coding RNAs participate in several roles. Understanding the RNA folding space is vital to detect insight on the RNA roles. The recent application of algebraic topology to the analysis of multidimensional biological complex systems motivated us to apply persistent homology to the exploration of the RNA data and in detail to the RNA folding space. Since the RNA secondary structures in the folding space grow exponentially with its sequence length, the resulting folding space corresponds to a multidimensional space. Thus, we propose persistent homology filtration as an important step towards obtaining the persistent structural elements, encoded by the generators of one dimensional holes, of optimal or near-optimal RNA secondary structures. The proposed approach is tested on RNA sequence data of different lengths and showed promising results. In general, our results clearly indicate that persistent homology extracts the essential structural elements that make up the optimal or near-optimal structure of the given RNA sequences, but it does not replace MFE structure prediction. However, we argue that the approach proposed in this contribution opens a new way to predict RNA secondary structure through the application of computational topology. Besides, in previous work, we introduced the graph rewriting formalism for modeling the RNA folding evolution [19]. Thus, there is the possibility of combining our TDA approach with graph rewriting; the structural elements generated by our TDA approach are sub-graphs that are constructed over high-probable base-pairs. This has the potential to reduce the state space of the RNA graph rewriting system.

Acknowledgement: We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM (www.topdrim.eu), number FP7-ICT- 318121.

We also thank Christian Reidys for the fruitful discussions and concise criticism on the idea presented during the TOPDRIM meetings.

References

[1] S. Bellaousov, J. S. Reuter, M. G. Seetin, D. H. Mathews, RNAstructure: Web Servers for RNA Secondary Structure Prediction and Analysis. Nucleic acids research, 41(2013).

[2] J. Binchi, M. Rucco, E. Merelli, G. Petri, F. Vaccarino, jHoles: a Tool for Understanding Biological Complex Networks via Clique Weight Rank Persistent Homology. Electronic Notes in Theoretical Computer Science 306(2014), 5-18.

[3] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological Classification of RNA Structures. Journal of Molecular Biology 379(2008), 900-911.

[4] G. R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, V. S. Pande, V.S., Structural Insight into RNA Hairpin Folding Intermediates. Journal of the American Chemical Society 130(30)(2008), 9676-9678.

[5] Z. Cang, L. Mu, K. Wu, K. Opron, K. Xia, G. W. Wei, A Topological Approach for Protein Classification. Mol. Based Math. Biol. (3) 2015, 140-162.

[6] G. Carlsson, Topology and Data. Am. Math. Soc, 46(2)(2009),255-308.

[7] G. Carlsson, A. Zomorodian, A. Collins, L. J. Guibas, Persistence Barcodes for Shapes. International Journal of Shape Modeling, 11(2)(2005),149-187.

[8] G. Carlsson, T. Ishkhanov, V. Silva, A. Zomorodian. On the Local Behavior of Spaces of Natural Images. International Journal of Computer Vision, 76(1)(2008),1-12.

[9] J. M. Chan, G. Carlsson, R. Rabadan. Topology of Viral Evolution. *Proceedings of the National Academy of Sciences, 110(46)*,18566-18571,Epub, 2013.

[10] R. Dirks, N. Pierce, A Partition Function Algorithm for Nucleic Acid Secondary Structure including Pseudoknots. Journal of Computational Chemistry 24(2003), 1664-1667.

[11] H. Edelsbrunner, J. Harer, Persistent Homology a Survey. Contemporary mathematics 453(2008), 257-282.

[12] C. Flamm, I. L. Hofacker, Beyond Energy Minimization: Approaches to the Kinetic Folding of RNA. Monatshefte fur Chemie-Chemical Monthly 139(4)(2008),447-457.
[13] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. Topological Measurement of Protein Compressibility via Persistence Diagrams. Japan Journal of Industrial and Applied Mathematics, 32(2014),1-17.
[14] J. Huang, R. Backofen, B. Voβ, Abstract Folding Space Analysis Based on Helices. RNA 18(12)(2012), 2135-2147.
[15] I. L. Hofacker, F. S. Peter, RNA Secondary Structures. Bioinformatics: From Genomes to Therapies 1(2007), 439-489.
[16] H. Jabbari, A. Condon, A Fast and Robust Iterative Algorithm for Prediction of RNA Pseudoknotted Secondary Structures. BMC Bioinformatics 15.1(2014).
[17] Z. J. Lu, J.W. Gloor, D. H. Mathews, Improved RNA Secondary Structure Prediction by Maximizing Expected Pair Accuracy. RNA, 15(10)(2009), 1805-1813.
[18] A. Mamuye, M. Rucco. Persistent Homology on RNA Secondary Structure Space. *Proc.of the 9th EAI Int. Con. on Bio-inspired Information and Communications Technologies*, 189-192,New York, USA, 2016.
[19] A. L. Mamuye, E. Merelli, L. Tesei, A Graph Grammar for Modeling RNA Folding. Proceedings of GaM 2016, Eindhoven, the Netherlands, April 2-3, 2016. Electronic Proceedings in Theoretical Computer Science, to appear.
[20] D. H. Mathews, Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization. RNA 10(8)(2004), 1178-1190.
[21] E. Merelli, M. Rucco, P. Sloot, L. Tesei (2015): Topological Characterization of Complex Systems: Using Persistent Entropy. Entropy 17(10), pp. 6872-6892.
[22] K. V. Morris, J. S. Mattick, The Rise of Regulatory RNA. Nat. Rev. Genet 15(2014), 423-437.
[23] R. Penner, M. W. Waterman, Spaces of RNA Secondary Structures. Advances in Mathematics 101(1)(1993), 31-49.
[24] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, F. Vaccarino, Homology Scaffolds of Brain Functional Networks. J. R. Soc. Inteerface 11(2014).
[25] C. M. Reidys, F. Huang, J. E. Adersen, R. C. Penner, P. F Stadler, M.E. Nebel, Topology and Prediction of RNA Pseudoknots. Bioinformatics 27(2011), 1076-1085.
[26] V. D. Silva and R. Ghrist, Blind Swarms for Coverage in 2-D. *In Proceedings of Robotics: Science and Systems*, page 01, 2005.
[27] A. Tausz, M. Vejdemo-Johansson, H. Adams, JavaPlex: A Research Software Package for Persistent (co)homology. Software available at http://code. google. com/javaplex (2011).
[28] The RNAcentral Consortium. RNAcentral: an International Database of ncRNA Sequences. Nucleic acids research, 991(2014).
[29] M. S. Waterman, F. F. Smith, RNA Secondary Structure: A Complete Mathematical Analysis. Mathematical Biosciences 42(3)(1978), 257-266.
[30] K. Xia, Z. Zhao, G. W. Wei, Multiresolution Persistent Homology for Excessively Large Biomolecular Datasets. J Chem Phys, 143(13)(2015),134103.
[31] K. Xia, Z. Zhixiong, and G. W. Wei, Multiresolution Topological Simplification. Journal of Computational Biology 22(9) (2015): 887-891.
[32] Y. Yao et al. Topological Methods for Exploring Low-density States in Biomolecular Folding Pathways. J. Chem. Phys. 130(2009), 144115.