Cilt 12, Sayı 2, Sayfa 32-39, 2023 Volume 12, Issue 2, Page 32-39, 2023
Araştırma Makalesi https://doi.org/10.46810/tdfd.1240075 Research Article

BİNGÖL ÜNİVERSİTESİ
Türk Doğa ve Fen Dergisi
Turkish Journal of Nature and Science
www.dergipark.gov.tr/tdfd

Classifying RNA Strands with A Novel Graph Representation Based on the Sequence Free Energy

Enes ALGÜL[1*]

[1] Bingol University, Engineering and Architecture Faculty, Department of Computer Engineering, Bingöl, Türkiye
Enes ALGÜL ORCID No: 0000-0001-6597-4242

**Corresponding author: ealgul@bingol.edu.tr*

(Received: 20.01.2023, Accepted: 05.05.2023, Online Publication: 22.06.2023)

Keywords
Graph representations,
RNA,
Graph Kernel,
Machine Learning

Abstract: Ribonucleic acids (RNA) are macromolecules found in all living cells and act as mediators between DNA and protein**.** Structurally, RNAs are more similar to the DNA. In this paper, we introduce a compact graph representation utilizes the Minimum Free Energy (MFE) of RNA molecules' secondary structure. This representation represents structural components of secondary RNAs as edges of the graphs, and MFE of these components represents their edge weights. The labeling process is used to determine these weights by considering both the MFE of the 2D RNA structures, and the specific settings in the RNA structures. This encoding makes the representation more compact by providing a unique graph representation for the secondary structural elements in the graph. Armed with the representation, we apply graph-based algorithms to categorize RNA molecules. We also present the result of the cutting-edge graph-based methods (All Paths Cycle Embeddings (APC), Shortest Paths Kernel/Embedding (SP), and Weisfeiler - Lehman and Optimal Assignment Kernel (WLOA)) on our dataset [1] using this new graph representation. Finally, we compare the results of the graph-based algorithms to a standard bioinformatics algorithm (Needleman-Wunsch) used for DNA and RNA comparison.

Serbest Enerji Serisine Dayalı Yeni Bir Graf Temsili ile RNA Zincirlerini Sınıflandırma

Anahtar Kelimeler
Graf gösterimleri,
RNA,
Graf çekirdek yöntemleri,
Makine öğrenmesi

Öz: Ribonükleik asitler (RNA), tüm canlı hücrelerde bulunan makromoleküllerdir ve DNA ile protein arasında aracı görevi görürler. Yapısal olarak daha çok DNA'ya benzerler. Bu makalede, RNA moleküllerinin ikincil yapısının Minimum Serbest Enerjisini (MSE) kullanan kompakt bir grafik gösterimi sunuyoruz. Bu temsilde, RNA moleküllerinin yapısal bileşenleri grafların kenarlarını ve bu bileşenlerin MSE'si, kenar ağırlıklarını temsil eder. Etiketleme işlemi, hem iki boyutlu RNA yapılarının MSE'sini hem de RNA yapılarındaki belirli ayarları dikkate alarak kenar ağırlıkları belirlenir. Bu kodlama, graftaki ikincil yapı elemanlarına benzersiz bir graf temsili vererek gösterimi daha kompakt hale getirmek için kullanılır. Temsil ile donanmış olarak, RNA moleküllerini kategorilere ayırmak için graf tabanlı algoritmalar uyguluyoruz. Ayrıca, veri kümemizde [1] bu yeni graf temsilini kullanarak en son graf tabanlı yöntemlerinin (All Paths Cycle Embeddings (APC), Shortest Paths Kernel/Embedding (SP) ve Weisfeiler - Lehman and Optimal Assignment Kernel (WLOA)) sonuçlarını sunuyoruz. Son olarak, graf tabanlı algoritmaların sonuçlarını DNA ve RNA karşılaştırması için kullanılan standart bir biyoinformatik algoritma (Needleman-Wunsch) ile karşılaştırıyoruz.

1. INTRODUCTION

DNA is a double-stranded nucleic acid that holds the blueprint for the development and function of all living beings, in the form of genetic code [29]. The code is a sequence of three base letters (’AAA’, ’ATC’, ’GTT’, ’GTA’, …). RNA is a DNA-like single-stranded polymer of nucleotides made up of various kinds of nucleobase building blocks. One of the functions of RNA is to copy the blueprints of the DNA and translate them to the protein via ribosomes in the cell. Therefore, a new RNA is a complementary part of the DNA chain. Both molecules include bases (nucleotides).

The RNA's primary structure refers to the order of its nucleobases. The nucleotide sequences then fold upon themselves to build topological structures, known as 2D RNA structures. These 2D RNA structures composed of loops (hairpin loop, bulge, junction, and interior loop) and base pairs (stem/helix) [30]. The 2D RNA structure, which is more complex than the primary structure, holds more information than just the order of its nucleobases. Representing the primary RNA structures in graph data is relatively straightforward by using vertices for each nucleobase and edges for connections between adjacent nucleobases. However, representing complex secondary RNAs in graph data is a more challenging task. These structures can also fold onto themselves to build $3D$ complex shapes, which are more similar to protein shapes, but are chemically more similar to DNA. Therefore, RNA contains sequence, structural, and shape information.

RNA's biological functions are closely linked to its structure [2]. The 2D structures of RNA are predicted and may differ from their actual shapes. Therefore, it is essential to utilize a robust tool for efficiently generating secondary structures. The prediction of the biological functions of RNA is related to obtaining useful information contained in the structure of RNA. Currently, the research community primarily employs two methods for generating secondary RNA structures. One approach focuses on minimizing free energy, while the other approach involves determining the association of nucleotides through the distance between their atoms. The following challenges are encountered during the encoding of secondary RNA structures.

- How to transform secondary RNA structures into a graph data form, as well as extract valuable structural features of 2D structures?
- How reduce the size of 2D RNA graph data?
- How to make pairwise comparison of RNA strands using their 2D structures to predict their biological tasks?
- How to solve multi-class classification problem of 2D RNA representations?

This article presents novel and advanced techniques to tackle these problems by encoding RNA secondary structural motifs as edges and connections between them as vertices. Additionally, to gain more features for encoding 2D RNAs in the graph data form, the total MFE is encoded as edge weights.

2. RELATED WORK

A Graph $G = (V, E)$ can be defined as set of nodes and edges (V, $E \subseteq V \times V$), and their labels. An edge (E) connects two nodes. A path is a sequence of vertices in a given graph**.**

RNA can be represented as a graph by encoding each RNA nucleobase as a node and placing edges between neighboring nodes. In this representation, X3DNA [3] is used to extract bound base pairs. In the graph representation of RNA, each node is labelled with its associated base. This representation was first used in the 1970s by Waterman [4] and has been used in various studies since then.

In 2004, the RAG [5] introduced a graph representation using existing and hypothetical secondary structural components of RNAs. The RAG web resource enumerates the number of nodes in graphs of RNA motifs according to their 2D complexity and classifies them based on functional types. RAG introduces a tree graph representation for hypothetical RNA tree shapes and a dual graph representation for other RNA structures that include both trees and pseudoknots. RAG uses labelled vertices and directed edges in graphs, and in the RAG representation, it assigns stems as edges and loops as nodes.

In 2012, Knisley et al [6] developed a dual graph RNA representation that can encode all structural types of 2D RNAs. They employed a multi-graph secondary RNA representation by using existing graph-theoretic descriptors to categorize all likely secondary RNA topologies with stems encoded as nodes and other structural motifs encoded as edges.

In 2016, a novel method was introduced by Huang et al [7], where RNA molecules were encoded by transforming graphs into topological spectrums. The subgraphs of the RNA strands are defined as their topological fingerprints and are classified the RNA strands by comparing the fingerprints.

In 2018, a variety of RNA graph-based representations were introduced, including 1D/2D/3D RNA structures [8]. In the 1D RNA graph representation, nucleotides were encoded as nodes; in the 2D RNA graph representation, X3DNA was used to generate structures. In this representation, matched base-pair is indicated by adding cross-link edges. The (x, y, z) coordinate information of the $C3$ atom is used to provide 3D RNA graph representation.

3. A NEW GRAPH REPRESENTATION

All 2D RNA structures are predicted structures and closely related to their structural components [2, 8, 10-14]. 2D RNA structural components can be predicted by free energy minimization. Many techniques have been developed to predict 2D RNA structures. First, a dynamic algorithm [15] was developed by Zuker and Stiegler in the 1980s for generating secondary RNA structures utilizing MFE in loops. Subsequently, widely used methods like the Mfold Web Server, Vienna RNA Software, and many other studies [8, 10, 11, 13, 16, 18] have used the MFE approach to predict secondary RNA structures.

X3DNA [3] can generate and visualize 2D/3D RNA structures from PDB files using the distance between atoms of base pairs. X3DNA considers the orientation and relative position of atoms $(C3, O2)$ of two bases to predict 2D RNA structures [17].

3.1. Sequence Free Energy

In this work, our approach is to receive all MFEs of 2D RNA structural components and transform them into a graph-structured data form.

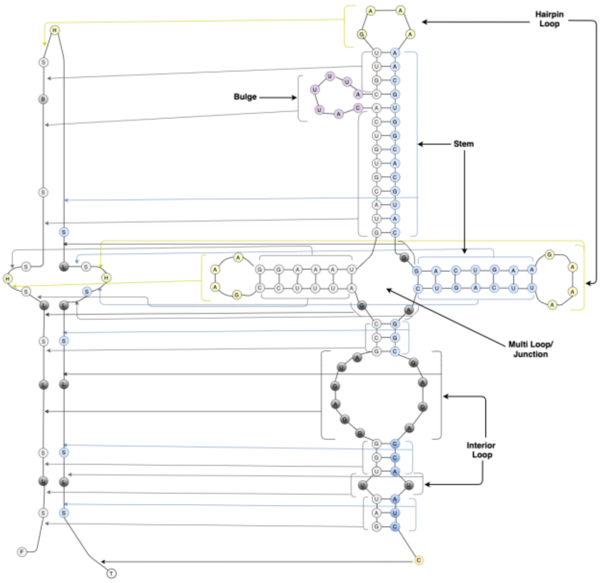


Figure 1. The structure of the Escherichia coli Riboswitch is represented in 2D using the MFE information obtained from the 4Y1M.pdb file.

Our secondary RNA graph representation is divided into three categories: stems/helixes, loops, and unpaired parts of the chain. We construct a graph using the MFE for the stems/helixes shown in Table 1 produced by Vendeix et al [19, 20] and loops shown in Figure 2 produced by Tinoco et al. [21], and a regular tetrahedral approach for unpaired part of the RNA chain as follows.

Table 1. MFE between base pairs received from [19].

Base pair	kcal/mol
$U - G$	− 4.45
$U - U$	− 5.82
$C - G$	− 5.53
$U - A$	− 4.42
$U - C$	− 0.37
$U - A$	− 4.42

The paired parts of RNAs are called stems or helixes. The highest negative energy arises among the most sturdy RNA base pairs. X3DNA [3] is utilised to generate base pair information.

The edge weight of loops in RNA molecules is determined by using the MFE on hairpin, interior, and bulge loops as seen in Figure 2 [21]. This method, although being one of the oldest techniques, is still used today for computing MFE [22]. However, we have no information of MFE on junctions available, and they are utilised similarly to the interior loops when calculating edge weights.

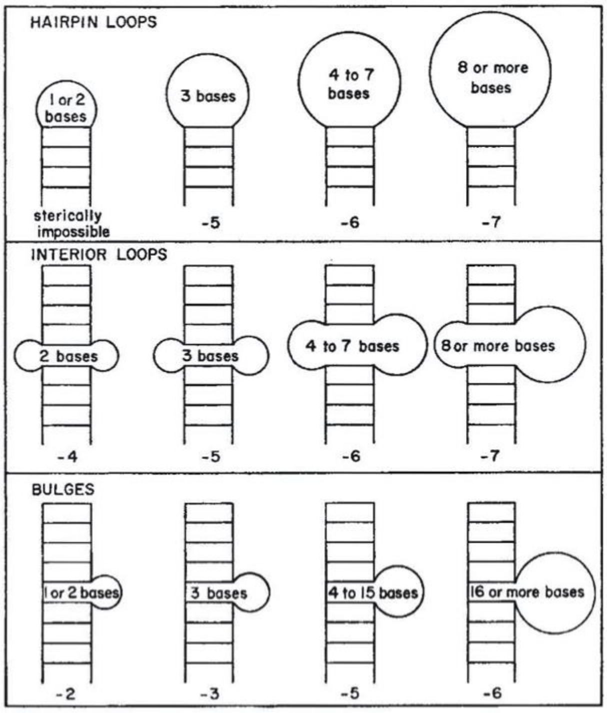


Figure 2. This figure is received from [21]. The MFE information on loops.

Pretty much, 24 percent of RNA molecules in the York RNA Dataset [1] lack 2D structural components and are unpaired. This affects correctly categorizing RNA molecules using learning methods. To solve this problem, we developed an advanced approach for encoding these unpaired RNA strands based solely on their base sequences. We used 3-base codes as edge weights to make the graph smaller (see Figure 3). Each of 3-base codes $A(1,1,1)$, $G(-1,-1,1)$, $U(1,-1,-1)$, and $C(1,-1,-1)$ is encoded as a regular tetrahedron with equal distance between any paired bases $(2\sqrt{2})$ (see Figure 4). We used K-medoids to reduce the group of possible 64 sequences to 4 for generating the code-book's code-words. Then, Learning Vector Quantization (LVQ) utilized these code-words to cluster edge weights into four classes. We made this improvement because some graph kernels solely apply to a finite number of discrete node labels.

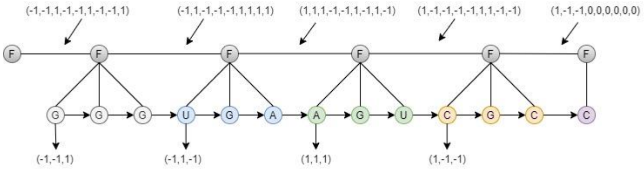


Figure 3. Unpaired part of the strands represented into weighted graph structured data using base sequences obtained from 1a3m.pdb (chain B) file.

RNA strands with 2D structures also have unpaired sequences, known as tails and queues, from the 5′ to 3′ of RNA strands. The same method discussed previously is applied to these tails and queues.

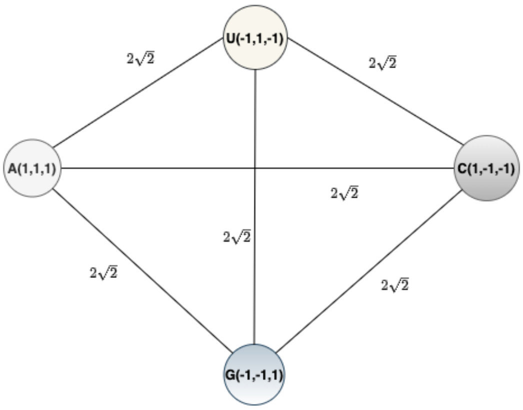


Figure 4. A regular tetrahedron represents base position.

Some graph kernels, such as the WL kernel, operate on discrete labels. Therefore, edge weights are refined as follows: the edge labels of 3-base codes are integers ranging from (-1 to 4). For other edge weights, we divided them by 10 if their absolute value was greater than 5, and rounded all numbers to tenths to reduce the number of decimal places to one.

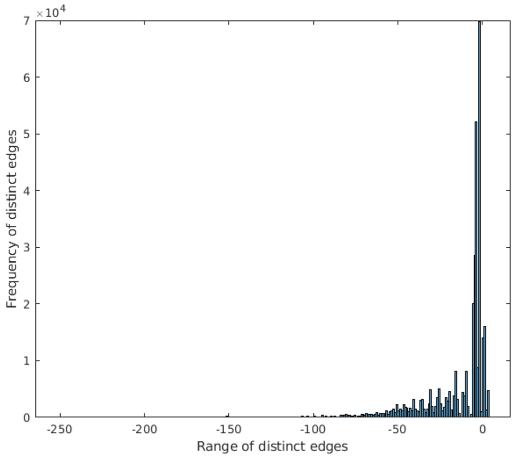


Figure 5. The histogram illustrates the range of edge labels prior to edge modifications [-252.0 to 4.0]

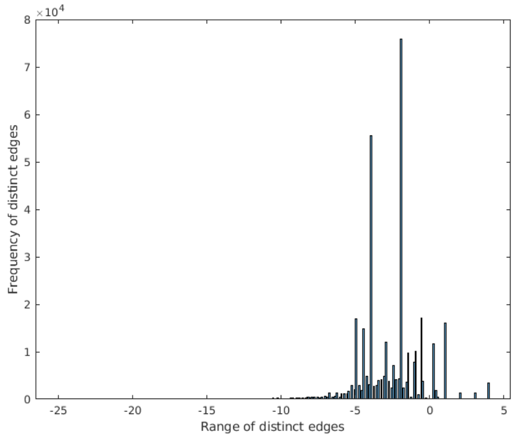


Figure 6. The histogram illustrates the range of edge labels after modifications [-26.0 to 4.0].

As illustrated in Figure 5 and Figure 6, the range of edge weights significantly reduced. This representation allows for relevant information to be retained while decreasing the number of vertices and edges by 75.4%.

5. DATA

The RNA Graph Classification Dataset used in this work was compiled by the University of York [1]. The dataset was labelled by using the biological function of the RNA molecules and includes 3178 RNA chains in the PDB files. The sequence and (x, y, z) coordinates information of the nucleobases were extracted from the pdb files.

This dataset is the largest dataset based on the categorization of RNA molecules according to their biological functions. The RNA classes are RIBONUCLEASE (14), RIBOSWITCH (227), MRNA (179), RIBOZYME (259), RRNA (1135), SRP (57), TRNA (581), and OTHER (726). The amount of each class of RNA is represented in parentheses. Detailed information about the dataset is available in [1], and the dataset is available for download at the https://www.cs.york.ac.uk/cvpr/RNA.html web site.

5. CLASSIFICATION METHODS

We apply pairwise graph comparison methods to categorize RNA structures according to their 8 types of biological function. In particular, we use graph kernels previously used on the other RNA datasets.

5.1. Weisfeiler-Lehman Optimal Assignment Kernel (WLOA)

The Weisfeiler-Lehman (WL) is a graph kernel function that compares the structural similarity of labelled graph pairs [31]. This cutting-edge kernel method treats each graph as a sequence of its labels. The WL method recursively relabels the nodes of graphs according to their neighbourhood node labels and makes a comparison of the resulting node and edge labels at each iteration [23]. For a fixed number of iterations, the edge refinement process is repeated. At each iteration, the WL kernel sorts the edge and node labels, uses a global hash function for compression, and counts common labels [25].

Optimal Assignment (OA) [24] is a kernel function base on the optimal assignment problems between the nodes of graph pairs [32]. OA count the similarity score of the perfect match between the node labels of graph pairs for classification problems. WLOA receives an OA kernel where the labels are obtained by the WL method [25], with the starting label corresponding to the encoded RNA graph's vertex labels [1]. This method provides the highest classification accuracy in applying classifier methods to our dataset.

5.2. Shortest Path Kernel/ Embedding (SPK)

A walk kernel calculates the similarity of two graphs by determining the number of shared walks of a specific type in both graphs. The Shortest path kernel (SPK) [26] is a type of walk-based kernel that computes the shortest walks between graph pairs to measure the similarities between two vertices within a graph.

$$K_{SP}(G,H) = \sum_{p_i \in SP(G)} \sum_{p_j \in SP(H)} K_B(p_i, p_j) \tag{1}$$

Here, $SP(.)$ represents the set of shortest paths, and K_B is a delta kernel that compares all shortest distances in the graph. Each path is represented by a label sequence, and in our application, we compute the SP length for each RNA strand and represent it as a histogram of these paths in a feature space. The edge lengths are determined by the RNA edge weights, and the paths are labelled by their start and end vertex labels.

5.3. All Paths and Cycles Kernel /Embedding (APC)

The APC is a kernel function [27] based on counting all possible paths, including cycles, within a graph, unlike SP kernel which only computes the shortest path.

$$K_{APC}(G_1, G_2) = \sum_{p_i \in PC(G_1)} \sum_{p_j \in PC(G_2)} K_B(p_i, p_j) \tag{2}$$

Here, $K_B(.)$ is a base kernel, and $PC(.)$ represents a group of all possible paths, including cycles, on G . The maximum length of the path must be limited, and the amount of the discrete node/edge labels must be less than 4. To apply this method to the primary and secondary RNA structure, we label the bases of RNA strands with three labels *A/U*, *G/C,* and *other.* Similarly, we label the nodes with three labels *F*, *N/Q/H,* and *B/L/*other to apply this method to our introduced representation. With these limitations, this kernel is computationally very efficient. This kernel has the same explicit embedding as the SP, such as the histogram of discrete labelled paths [1].

5.4. Needleman - Wunsch Algorithm

The Needleman-Wunsch algortihm [28] is a populer sequence-based method in application of DNA and protein for comparison. Since RNA nucleobase sequences are similar to that of DNA, Needleman-Wunsch algortihm can also be operated on RNA sequences for comparison. The RNA nucleotides are denoted as a sequence of 4 letters (A, C, U, G) and the Needleman-Wunsch algortihm [28] used to align these strings. The Jukes-Cantor method is employed to count the distances between RNA sequences.

$$d = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right) \tag{3}$$

Where, the distances between RNA sequences denotes p with the range of, $0 \le p \le 1$, for the portion of sets that are distinct. A distance matrix, D, is then populated with these distances, and multidimensional scaling is applied to map these distances into a feature space.

3. RESULTS

We presented the results from the introduced graph representations of secondary RNAs, utilizing the cutting-edge graph kernels algorithms to categorise RNA strands from the York RNA Graph dataset.

This analysis aims to determine if the introduced representation is effective in classifying RNA strands by reducing complexity and increasing accuracy**.** Various structural information from RNA, including topology, sequence, and the introduced representations, were also evaluated to classify RNA molecules. The RNA sequence representation from [1] consists of base labels and edges located between adjacent bases, while the 2D RNA representation from [1] includes edges of the graph but no labels of bases. The NR-W representation is a weighted secondary graph representation of RNA introduced in the Section 3.1. On the other hand, the NR-UW is an unweighted representation of the same graph, in which the edge weights have been eliminated.

Table 2. The prediction accuracy of various methods and representations, such as APC, WLOA, SP, and SA on the York RNA Graph Dataset. *These classification results are taken from our previous study [1].

	Seq*	Top*	Seq+Top*	NR-W	NR-UW
WLOA	92.0	73.1	**92.4**	86.6	77.5
SP	**91.3**	79.5	91.1	79.6	75.4
APC	**90.3**	85.4	89.9	80.8	79.4
SA	89.2				

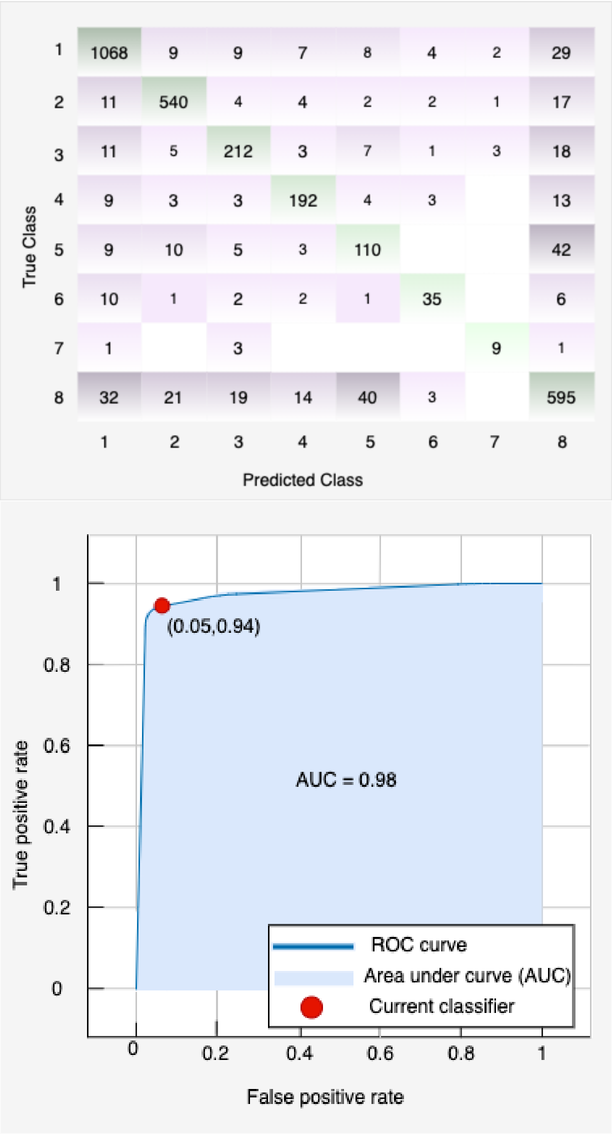
Table 3: The classification time for the York RNA Dataset utilising various methods and representations, such as APC, WLOA, SP, and SA, expressed in units of seconds (s), minutes (m), and hours (h).

	Seq		Top		Seq + Top		NR-W		NR-UW	
	Acc.	speeds	Acc.	speeds	Acc.	speeds	Acc.	speeds	Acc.	speeds
WLOA	**92.0**	$16m\ 32s$	73.1	$5m\ 32s$	**92.4**	$3m\ 13s$	86.9	$2h\ 59m\ 44s$	77.5	$19s$
SP	91.3	$10h\ 36m\ 14s$	79.5	$10h\ 40m\ 10s$	91.1	$10h\ 43m\ 15s$	80.9	$13m\ 1s$	75.4	$10m\ 19s$
APC	90.3	$17m\ 57s$	**85.4**	$16m\ 33s$	89.9	$24m\ 20s$	79.5	$1m\ 46s$	79.4	$1m\ 9s$

Then, we used graph kernel methods and machine learning classification algorithms to evaluate the outcome of various techniques on various sources of RNA information. Bagged Trees performed the best with the SPK method utilising NR-UW, while Quadratic SVM and Cubic SVM produced the most advanced results with WLOA on the same representation when PCA was applied. Furthermore, Subspace KNN achieved the most favourable outcomes with the APC method when utilising all representations, as well as with the SPK using NR-W, Top, Seq, and with WLOA using NR-W.

Moreover, we analyzed the outcome of classifications using ROC curves and found that Subspace KNN had a TPR of 94% and an AUC of 98%. Our experiments revealed that using reduced graph representations resulted in lower performance but faster execution time.

The highest performance was achieved when applying WLOA to the Seq + Top representation of RNA, with an outcome of 92.4% accuracy. The introduced NR-W representation had an accuracy of 86.6% and performed better than the topology graph representation of 2D RNA. The 2D RNA graph representation revealed the highest accuracy (85.4%) when using the APC method. Our introduced NR-W representation had a higher accuracy of 86.6% when using the WLOA kernel, as it includes both sequence and edge labels of reduced graph representation, whereas the topology representation solely consists of edge information of the exact graph representation. As shown in Table 2 and Table 3, incorporating Minimum Free Energy as edge weights on structural components produced better outcomes. However, combining the topology with the sequence (Seq + Top) provided superior performance than NR-W.



	1	2	3	4	5	6	7	8
1	1068	9	9	7	8	4	2	29
2	11	540	4	4	2	2	1	17
3	11	5	212	3	7	1	3	18
4	9	3	3	192	4	3		13
5	9	10	5	3	110			42
6	10	1	2	2	1	35		6
7	1		3				9	1
8	32	21	19	14	40	3		595

Figure 7. Evaluation of the NR-W using the ROC Curve and Confusion Matrix using WL-OA kernel.

Eliminating the weights from the edge of the full graphs decreased the accuracy (79.4%) when applying APC. The outcomes indicate that using weighted RNA graph representations improves performance compared to unweighted RNA graph representations when using WLOA, SP, and APC kernel methods.

Overall, our introduced representation effectively reduced the exact graph size and improved classification accuracy for 2D RNA structures. Despite this, the primary RNA representation performed better than all 2D RNA representations. For achieving greater classification accuracy, more advanced graph representations are needed for encoding 2D RNA structures.

7. CONCLUSION

In this research article, we addressed the problem of representing 2D RNA structures. We investigated existing representations and introduced a novel RNA graph representation using various methods, including investigating the impact of minimum free energy (MFE) on predicting RNA function. We also considered infrequent parts of the RNA structure and applied alternative approaches to encoding them, ultimately building a compact graph representation.

We successfully transformed MFE into a graph-based representation and demonstrated that representing MFE as weigths of graph edges, along with refinements on infrequent parts of the RNA graph representation, improves classification accuracy.

We applied a standard sequence-alignment method and the most advanced graph kernel methods to the graph representations for producing a Gramian matrix. Then, we utilised machine learning classification algorithms to this matrix to categorise it into high-level classes. The highest result from [1] using WLOA on the Seq had an accuracy of 92.0%. However, our representation, using the WLOA method on the 2D RNA structure and with 75% fewer vertices, achieved the best result with an accuracy of 86.6% and performed better than the highest results from [1] using the same graph kernels. The unweighted graph representation of the 2D RNA can also classify RNA molecules but with a trade-off of increased speed but lower accuracy.

Despite this success, the results from all 2D RNA graph representations did not perform as well as the primary RNA graph representation due to the loss of information and unpredictability of actual 2D RNA shapes.

REFERENCES

[1] E. Algul and R. C. Wilson, “A database and evaluation for classification of RNA molecules using graph methods,” in Graph-Based Representations in Pattern Recognition, D. Conte, J.-Y. Ramel, and P. Foggia, Eds. Cham: Springer International Publishing, 2019, pp. 78–87.

[2] D. Bechhofer and M. Deutscher, "Bacterial ribonucleases and their roles in rna metabolism," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 54, pp. 242–300, 05 2019.

[3] "3dna: a suite of software programs for the analysis, rebuilding and visualization of 3-dimensional nucleic acid structures," x3dna.org. [Online]. Available: http://x3dna.org/

[4] M. S. WATERMAN, "Secondary structure of singlestranded nucleic acids," Studies in Foundations and Combinatorics Advances in Mathematics Supplementary Studies, vol. 1, pp. 167–212, 1978. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.4425rep=rep1type=pdf

[5] D. Fera, N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick, "Rag: Rna-as-graphs web resource," BMC Bioinformatic, vol. 5, 07 2004. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471- 2105-5-88

[6] D. Knisley, J. Knisley, C. Ross, and A. Rockney, "Classifying multigraph models of secondary rna structure using graph-theoretic descriptors," ISRN Bioinformatics, International Scholarly Research Network, 11 2012. [Online]. Available: https://doi.org/10.5402/2012/157135

[7] J. Huang, K. Li, and M. Gribskov, "Accurate classification of rna structures using topological fingerprints," PLOS ONE, vol. 11, no. 10, pp. 1–19, 10 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0164726

[8] R. C. Wilson and E. Algul, "Categorization of rna molecules using graph methods," in Structural, Syntactic, and Statistical Pattern Recognition, X. Bai, E. R. Hancock, T. K. Ho, R. C. Wilson, B. Biggio, and A. Robles-Kelly, Eds. Cham: Springer International Publishing, 2018, pp. 439–448.

[9] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph kernels," Journal of Machine Learning Research, vol. 11, pp. 1201–1242, 2010.

[10] G. M. Blackburn, M. J. Gait, D. Loakes, D. M. Williams, J. A. Grasby, M. Egli, A. Flavell, S. Allen, J. Fisher, A. M. Pyle, *et al.*, *Nucleic acids in chemistry and biology*. Royal Society of Chemistry, 2006.

[11] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," Nucleic Acids Research, vol. 31, no. 13, pp. 3406–3415, 07 2003. [Online]. Available: https://doi.org/10.1093/nar/gkg595

[12] H. Jabbari, I. Wark, and C. Montemagno, "Rna secondary structure prediction with pseudoknots: Contribution of algorithm versus energy model," *PLOS ONE*, vol. 13, pp. 1–21, 04 2018.

[13] Y. Wu, B. Shi, X. Ding, T. Liu, X. Hu, K. Y. Yip, Z. R. Yang, D. H. Mathews, and Z. J. Lu, "Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data," *Nucleic Acids Research*, vol. 43, pp. 7247–7259, 07 2015.

[14] K. Doshi, J. Cannone, C. Cobaugh, and R. Gutell, "Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction," *BMC bioinformatics*, vol. 5, p. 105, 09 2004.

[15] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," Nucleic Acids Research, vol. 9, no. 1, pp. 133–148, 01 1981. [Online]. Available: https://doi.org/10.1093/nar/9.1.133

[16] I. L. Hofacker, "Vienna RNA secondary structure server," Nucleic Acids Research, vol. 31, no. 13, pp. 3429–3431, 07 2003. [Online]. Available: https://doi.org/10.1093/nar/gkg599

[17] L. Wang, Y. Liu, X. Zhong, H. Liu, C. Lu, C. Li, and H. Zhang, "Dmfold: A novel method to predict rna secondary structure with pseudoknots based on deep learning and improved base pair maximization principle," *Frontiers in Genetics*, vol. 10, p. 143, 2019.

[18] P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner, "SCOR: a Structural Classification of RNA database," *Nucleic Acids Research*, vol. 30, pp. 392–394, 01 2002.

[19] X. Lu and W. K. Olson, "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures," *Nucleic Acids Research*, vol. 31, pp. 5108–5121, 09 2003.

[20] F. Vendeix, A. Munoz, and P. Agris, "Free energy calculation of modified base-pair formation in explicit solvent: A predictive model," RNA (New York, N.Y.), vol. 15, pp. 2278–87, 10 2009.

[21] I. TINOCO, O. C. UHLENBECK, and M. D. LEVINE, "Estimation of Secondary Structure in Ribonucleic Acids," Nature, vol. 230, pp. 362–367, 04 1971. [Online]. Available: https://doi.org/10.1038/230362a0

[22] N. Nicolo, "Learning with kernels on graphs: Dag-based kernels, data streams and rna function prediction," Alma Mater Studiorum-Universita di Bologna ´, 2014. [Online]. Available: https://pdfs.semanticscholar.org/313b/7d182e81e021faed1cf650f480fdeaeeb3d6.pdf

[23] G. K. D. de Vries, "A fast approximation of the weisfeiler-lehman graph kernel for rdf data," in Machine Learning and Knowledge Discovery in Databases, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezn ˇ y, Eds. ´ Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 606–621.

[24] N. M. Kriege, P.-L. Giscard, and R. C. Wilson, "On valid optimal assignment kernels and applications to graph classification," in Advances in Neural Information Processing Systems, 2016, pp. 1615–1623.

[25] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," Journal of Machine Learning Research, vol. 12, pp. 2539–2561, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2078187

[26] K. M. Borgwardt and H. Kriegel, “Shortest-path kernels on graphs,” in Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA, 2005, pp. 74–81. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2005.132

[27] P.-L. Giscard and R. C. Wilson, “The all-paths and cycles graph kernel,” arXiv preprint arXiv:1708.01410, 2017.

[28] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” Journal of Molecular Biology, vol. 43, no. 3, pp. 443–453, 1970.

[29] Schmidt, Marco F. "DNA: Blueprint of the Proteins." *Chemical Biology: and Drug Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2022. 33-47.

[30] Ou, Xiujuan, et al. "Advances in RNA 3D Structure Prediction." *Journal of Chemical Information and Modeling* 62.23 (2022): 5862-5874.

[31] Schulz, Till Hendrik, et al. "A generalized weisfeiler-lehman graph kernel." *Machine Learning* 111.7 (2022): 2601-2629.

[32] Salim, Asif, S. S. Shiju, and S. Sumitra. "Graph kernels based on optimal node assignment." *Knowledge-Based Systems* 244 (2022): 108519.