
Book Section:
Algul, Enes and Wilson, Richard Charles orcid.org/0000-0001-7265-3033 (2019) A Database and Evaluation for Classification of RNA Molecules Using Graph Methods. In: Graph-Based Representations in Pattern Recognition - 12th IAPR-TC-15 International Workshop, GbRPR 2019, Tours, France, June 19-21, 2019, Proceedings. Springer , pp. 78-87.

A Database and Evaluation for Classification of RNA Molecules using Graph Methods

Enes Algul and Richard C. Wilson

University of York, York, UK
ea918,richard.wilson@york.ac.uk

Abstract. In this paper, we introduce a new graph dataset based on the representation of RNA. The RNA dataset includes 3178 RNA chains which are labelled in 8 classes according to their reported biological functions. The goal of this database is to provide a platform for investigating the classification of RNA using graph-based methods. The molecules are represented by graphs representing the sequence and base-pairs of the RNA, with a number of labelling schemes using base labels and local shape. We report the results of a number of state-of-the-art graph based methods on this dataset as a baseline comparison and investigate how these methods can be used to categorise RNA molecules on their type and functions. The methods applied are Weisfeiler Lehman and optimal assignment kernels, shortest paths kernel and the all paths and cycle methods. We also compare to the standard Needleman-Wunsch algorithm used in bioinformatics for DNA and RNA comparison, and demonstrate the superiority of graph kernels even on a string representation. The highest classification rate is obtained by the WL-OA algorithm using base labels and base-pair connections.

1 Introduction

Ribonucleic acid (RNA) is chemically very similar to DNA in their polymer of nucleotides [1]. These nucleotides have sequences that can encode genetic information [2]. DNA stores genetic information while RNA, copied from DNA, carries and provides this genetic information to other biological process. RNAs are also well known to play important regulatory and catalytic roles [3]. These roles including transcriptional regulation, RNA splicing, and RNA modification and maturation [3]. RNA also very important for treatment of diseases including viral and bacterial infections, and cancer [4]. RNA is therefore crucial to all life and it is important to understand its function.

The primary structure of the RNA consists of nucleotide sequences, this nucleotide sequences can fold onto itself to create secondary and tertiary structure of the RNA. Unlike DNA, RNA is single strand and is encouraged to fold into complex shapes, like proteins, by the matching of base-pairs from the same strand. The secondary structure is formed by both Watson-Crick base pairs [5], (A-U, C-G) and non-standard pairs. The base pairs between A-U, C-G, and the wobble pair between G-U are referred to as canonical base pairs while base pairs

between other base pairs are called non-canonical base pairs [6] . The canonical base pairs are more stable and important than non-canonical bases in the structure of the RNA [6]. The secondary structures is the topology of the RNA folding, and it consist of five main structural components: called internal loops, hairpins, bulges, junctions, and stems. The geometric shape of the RNA is its tertiary (3D) structure.

The objective of this paper is to present a new, large, graph RNA dataset which can be used to investigate graph-based methods for RNA classification and discovery. We also investigate the performance of some standard methods on the dataset and the role of different elements of the RNA representation, particularly the labelling, topology and geometry of the RNA.

Outline of the paper: In section 2 we will explain related works. In section 3 we will demonstrate our dataset. In section 4 we will represent RNA molecules. In section 5 we will explain sequence alignment, Weisfeiler-Lehman optimal assignment kernel, all paths and cycle embedding, and shortest path embedding methods. In section 6 we will show our experimental results. In section 7 we will discuss on our experiments and conclude the paper.

2 Related Work

DNA and RNA have chemical and structural similarities. Both molecules are nucleic acid composed of nucleobases, although the sugar backbone of the polymer is different. The structure of the molecule is determined by the nucleotide sequence. Because of this, sequence alignment is commonly used to determine the biological function of the DNA, such as the Needleman Wunch algorithm [7]. This is essentially a string edit distance between the strings of base-labels. The nucleotide sequence is the *primary structure* of the RNA. Because RNA is single strand, it can fold on itself, and the folds can be held in place by base-pair bonding between bases at different points on the RNA strand. This creates the potential for a more complex topology than RNA. This structure is called the *secondary structure* and can naturally be represented by a (labelled) graph.

Graph Theory is a branch of mathematics which has been used in various areas, such as road systems, neurosciences, irrigation networks, chemical processes and structures, computer science, and bioinformatics [1]. Graph-based data is becoming more abundant in chemical pathways and protein structures, protein or gene regulation networks, and social networks [8]. Graph Kernels allow the application of kernel methods to graph data [10] and allow using a range of algorithms for pattern recognition [9]. Graph kernels bridge the gap between graph-structured data and a large spectrum of the machine learning algorithms such as SVM kernel regression, kernel PCA [8], KNN and ensemble classifiers (Subspace KNN, Subspace Discriminant, Bagged Trees, and Boosted Trees). In this work, the goal of applying graph kernels is to measure similarities between two patterns, while the goal of the Machine learning is to classify these similarities. Kernel methods are widely used in the field of the bioinformatics, such as in

Lodhi and Huma [10] where the spectrum kernel, marginalise kernel and fisher kernel were applied for sequence analysis.

3 Database

There exist large databases of DNA, RNA and protein structures. In the reviewed literature, most of the dataset are in fasta and protein data bank (pdb) file formats. The fasta files include the basic sequence of macro-molecules(protein, DNA, RNA) [11][12]. The pdb files include the information of each atom of the macro-molecules, their sequence information, and their atomic coordinates. The data which was used in recent studies are in the pdb format which is standardised according to the atomic coordinates [13]. The pdb files provided by the many organisations. The three largest of them are the Protein Data Bank Japan (PDBj)[14], Nucleic Acid Database (NDB) [15], and The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [16].

The pdb files consist of more than one chain of information of the RNA molecules. For instance, in the same pdb files there might be different kinds of chains belong to a different type of RNA or DNA or RNA-Protein interaction. For instance, in 1XNR.pdb file, chain A is 16S ribosomal RNA, chain X is anticodon TRNA, chain M is MRNA, and other chains such as chain B, C, D, ... , T,V are 16S ribosomal protein. It is not straightforward to extract single RNA chains from this type of data. The largest classified database of RNA structure (i.e. RNA strands with functional labels) known to us is that of Klosterman et al [17] which contains 419 RNA strands.

RNA Bricks [18] downloaded pdb files from the World Wide Protein Data Bank web site and split each pdb files by their chains. Their dataset is publicly available. We have extracted 3178 RNA chains from this dataset. For each of this molecules, we have investigated the literature to classify them into one of 8 possible RNA classes. The first step is to look at the MOL-ID field in the pdb files which include information of the type of RNA molecules. In some pdb files, the type of the RNA is not very clear from this field. For these, we undertook a more elaborate investigation using information derived from the HEADER, the TITTLE, the KEYWDS, and HD-RNA [19] in order to determine the type of the RNA chains. We then removed any chains where we were still unsure of the type. The result is a curated database of 3178 RNA molecules with 8 possible class labels which is available for download [20].

The RNA classes from the Table 1, ribozyme is a type of the RNA which catalyses chemical reactions, riboswitches behave like ribozymes and participate gene regulation, ribonucleases are very important enzymes in RNA degradation and maturation pathways, signal recognition particle (SRP) RNA, a part of ribonucleo-protein (protein-RNA complex), involves in targeting translocation of membrane proteins and secretory proteins. We labelled all other RNA classes, which the number of classes too small, in the OTHER section. We did not labelled our dataset according to the source of the organism.

CLASSES	KEYWORDS/ DESCRIPTION
RIBONUCLEASE (14)	RIBONUCLEASE P, RNASE P
RIBOSWITCH (227)	APTAMER
MRNA (179)	UTR, EXON
RIBOZYME (259)	S-TURN, CATALYTIC RNA, HAMMERHEAD, GLMS
RRNA (1135)	4.8S, 5S, 5.8S,16S, 18S, 23S, 25S, 26S, 28S, 30S, 40S, 50S, 60S, 70S, 80S, A-SITE OF HUMAN RIBOSOME, A-SITE OF HUMAN MITOCHONDRIAL RIBOSOME, A-SITE OF BACTERIAL RIBOSOME, SARCIN/RICIN 28S RRNA
SRP (57)	4.5S, 7S, 7SL
TRNA (581)	A-site, P-site, tRNA X-MER, FMET, FME, INITIATOR, INI, PRIMER CODON, ANTICODON,ACCEPTOR, tRNA-fragment,
OTHER (726)	viral RNA, miRNA, snoRNA, IRES RNA, and some undefined RNAs such as 5' RNA, 16-MER, 192-MER, 28-MER, 119-MER, 97-MER etc.

Table 1. The labelled classification of the RNAs and their descriptions. The number of each type of RNAs represents in the brackets.

In our data set, 332 RNA chains' nucleotide lengths are from 6 to 9, 1798 RNA chain's nucleotide sizes are from 10 to 100, 469 RNA chains nucleotide sizes are from 101 to 500, 277 RNA chain's nucleotide sizes are from 1326 to 1861, 286 RNA chain's nucleotide sizes are from 2227 to 2912, 15 RNA chain's nucleotide sizes are from 3119 to 3662, and nucleotide size of one chain is 4298. The percentage of the each type of the RNA molecules are 35.71% (RRNA), 22.84% (OTHER), 18.28% (TRNA), 8.15% (RIBOZYME), 7.14% (RIBOSWITCH), 5.63% (MRNA), 1.79% (SRP), and 0.44% (RIBONUCLEASE).

4 RNA Representation

The main component of Nucleic Acid is 5 carbon sugar (2-deoxyribose or ribose), a phosphate group and a base (one of four molecules = adenine, guanine, cytosine, and uracil/thymine). There is two macro nucleic acids called as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Ribonucleic Acid (RNA) consist of nucleotides. The nucleotides are composed of purine nucleobases, namely Adenine (A), Guanine (G), and pyrimidine nucleobases calling Cytosine (C) and Uracil (U).

The base sequences represents the primary sturcture of the RNA such as the following sequence is single stranded (strand B) of an Escherichica coli Riboswitch '4Y1M':

"GAUUUGGGGAGUAGCCGAUUUCCGAAAGGAAAUGUACGUGUCAA
CAUUUUCGUUGAAAAACGUGGCACGUACGGACUGAAGAAAUUCAGU
CAGGCGAGACCAUAUCC"

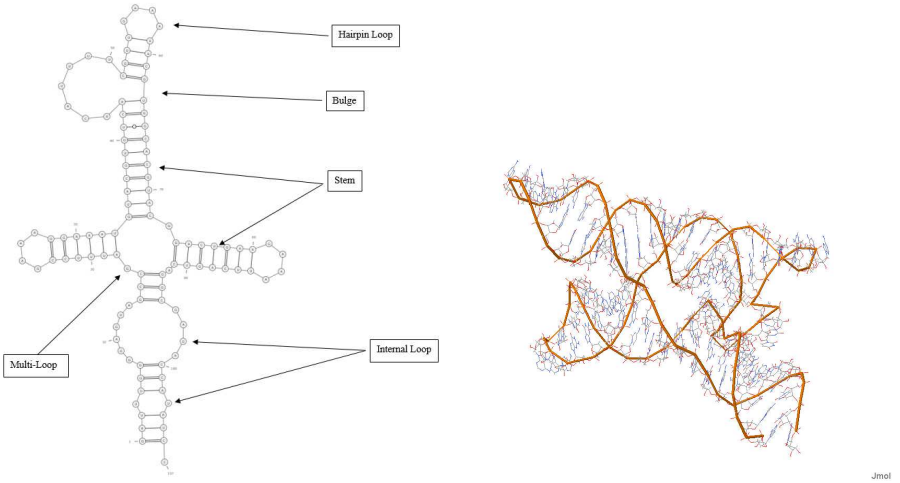

Fig. 1. 3D and secondary structure of the Escherichica coli Riboswitch derived from 4Y1M.pdb. The secondary structure's image produced using RNApdbee 2.0 [21] and 3D structure's image produced using Matlab molviewer.

The primary structure of the RNA folds on to itself to build secondary structure and 3D structure of the RNA as represented in Fig 2. 2D structure of the RNA is widely used to classify and define RNA molecules with their functions. The main components of the 2D RNAs are hairpin loop, bulge, internal loop, multi-loop, and stem. As represented on Fig 2; the uninterrupted base-paired portion of the RNA molecules is called a stem, the hairpin loops are sets of unpaired bases which connect to only one stem and occur at the end of the some sections. The multi-loop occurs among more than two stems while internal loops exist between only two stems. The internal loop is defined as two strands of unpaired bases occuring between two stems. The bulges are similar to internal loop it occurs between two stems, the only difference is that one strand's bases are unpaired.

The RNA is encoded as a graph using a straightforward representation. Each vertex represents an RNA residue and is labelled with the base code (A,G,C,U or a rare non-standard base). An adjacency matrix encodes the graph edges, which join all residues in sequence and any base pairs. Base-pairs are detected using the X3DNA program 'find-pairs' [22]. The vertices are also labelled with a geometric 'type' label; 1 if it is part of a base pair, 2 if it is unpaired but within 6.5Å of another base, and 3 otherwise. Two sets of 3D coordinates are also provided, firstly for the backbone position of the residue (the position of atom C3') and secondly the centroid position of the base. From this data, information about the secondary and tertiary structure can be inferred. A set of classification labels from the 8 classes in Table 1 is also provided.

5 Classification Methods

In this section, we present the results of applying some standard classification algorithms on our database. These are sequence alignment (SA), the shortest path kernel (SP) , all-paths and cycles kernel (APC), and the Weisfeiler Lehman optimal assignment method (WL-OA). These methods are described briefly in the following sections. SP and APC have an explicit embedding in vector space, which is used. WL-OA produces a kernel matrix which is embedded using kernel embedding. Sequence alignment produces a distance measure which is embedded using multi-dimensional scaling (negative dimensions use the absolute value [23]).

5.1 Sequence-Based Methods

Sequence-based methods mostly use in the study of the DNA. We have applied this method on the primary structure of the RNA for comparison. The sequence of the RNA is the primary structure of the RNA. The nucleotides are represented as strings such as adenine (A), guanine (G), cytosine (C), and uracil (U). The Needleman - Wunsch algorithm [7] to align the strings (A,G,C,U), and Jukes - Cantor score used to calculate the distances between RNA sequences, which are significantly larger than amino-acid sequences. The running time of this algorithm is therefore quite high. The equation of the Jukes - Cantor score:

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

Here, p denotes the distance between them in terms of the fraction of letters which differ.

5.2 Weisfeiler-Lehman Optimal Assignment (WL-OA)

The Weisfeiler-Lehman optimal assignment (WL-OA) graph kernel [24] is a state-of-the-art method for labelled graph comparison which utilises an optimal assignment kernel with the labelling generated by the WL method [8]. The method generated vertex labels using the WL label refinement process, with initial labelling corresponding to the RNA vertex labels. An implicit optimal assignment is sought which minimises the labelling difference, and the kernel value is the count of label differences for this optimal assignment. WL-OA performs favourably compared to state-of-art graph kernel in a wide range of datasets.

5.3 Shortest Path Embedding

The walk kernel counts the similar walk in graph pairs. The shortest path kernel (SPK) [25] is a type of walk kernel which counts only the shortest walks between each pair of nodes in a graph:

$$K_{SP}(G,H) = \sum_{p_i \in SP(G)} \sum_{p_j \in SP(H)} K_B(p_i, p_j) \tag{2}$$

where $SP(.)$ is the set of shortest paths in a graph and K_B is a base kernel which compares paths. In the case that the base kernel is the delta kernel, this has an explicit embedding as the histogram of the shortest paths, where each path is denoted by its sequence of labels. This method also called shortest path embedding. Each labelled shortest path in the molecule is generated, and the embedding is a histogram of these paths.

5.4 All Paths and Cycles Embedding(APC)

The all-paths graph kernel is a recently proposed extension to the shortest path kernel, which counts all paths (up to a maximum length).

$$K_{APC}(G,H) = \sum_{p_i \in PC(G)} \sum_{p_j \in PC(H)} K_B(p_i, p_j) \tag{3}$$

Here, $PC(G)$ denotes a set of all paths and simple cycles (a cycle is $v_1 v_2 ... v_n v_1, v_1 < 1 < v_n$ is distinct) on G and $K_B(.,.)$ is a base kernel for paths [26]. In order to evaluate this kernel in a computationally efficient way, the maximum path length and the number of distinct labels must be limited. We therefore label bases with three labels G/C, A/U and other. This is embedded in the same way as the shortest path kernel, as a histogram of distinct labelled paths.

6 Results

In this work, we presented our graph-based RNA dataset. We classified this dataset in 8 type of RNA categories as listed in Table 1. We also demonstrate here that graph based methods can be used to classify RNA molecules. To evaluate the effect of different types of structure (broadly corresponding to the primary, secondary and tertiary structure), we include information from the topology, sequence, and the geometry of shape of the RNA. *Seq* includes the graph edges corresponding to the sequence only, and the base labels. *Top* includes the graph edges (including the sequence and base-pair edges) but no base labels. *Geo* adds additional labels to the bases corresponding to the geometry type labels described earlier. The combinations are the union of these sources of information.

Then, we have applied graph-based embedding methods and a classifier to find the most effective methods on our RNA dataset to determine the accuracy. We tried a variety of classifiers; Subspace KNN, Subspace Discriminant, Linear discriminant, Boosted Trees, Cosine KNN and Bagged Trees. In our experiments, Subspace KNN outperform the results with APC methods, and SP methods on three source of information of the RNA. Bagged Trees outperform best results with SA method. On the other hand, WL-OA performed its best results from variety methods on a variety of representations. Subspace KNN performs best result on Geometry Label, Geometry Shape of the RNA + Topology, Geometry + Residue Label, and All Label; Subspace Discriminant and Linear Discriminant demonstrate best results on Residue Label, and Residue + Topology Label; Boosted Trees performs best result on Topology Label.

	Seq only,	Top only,	Geo only,	Seq +Top,	Geo + Top,	Geo +Seq,	All
WL-OA	**92.0**	73.1	**86.8**	**92.4**	**87.1**	89.5	90.2
SP	91.3	79.5	86.7	91.1	86.7	**91.1**	**90.8**
APC	90.3	**85.4**	84.3	89.9	85.5	...	...
SA	89.2	...	...	...	...	...	...

Table 2. Classification accuracies for the RNA dataset using Weisfeiler Lehman Optimal Assignment (WL-OA), Shortest Path (SP), All-Paths and Cycle methods (APC), Sequence Alignment (SA), and a variety representations

The experimental results on the RNA dataset represents that, Weisfeiler-Lehman Optimal Assignment (WL-OA) methods outperform the Shortest Path (SP) method, All Path Cycle (APC) method, and Sequence Alignment (SA) method. The accuracy with WL-OA increased up to 92.4 percent, which is the best performance in all methods. The results here largely support the results of [2] in that all graph based methods outperform sequence alignment even when only the sequence data is available. It is clear that the base labels are important to the classification, with a drop-off of between 4–10% when they are not included. The all-paths kernel can only be evaluated with three labels, which means that it cannot be used on the experiments with rich label sets and explains the lower performance overall.

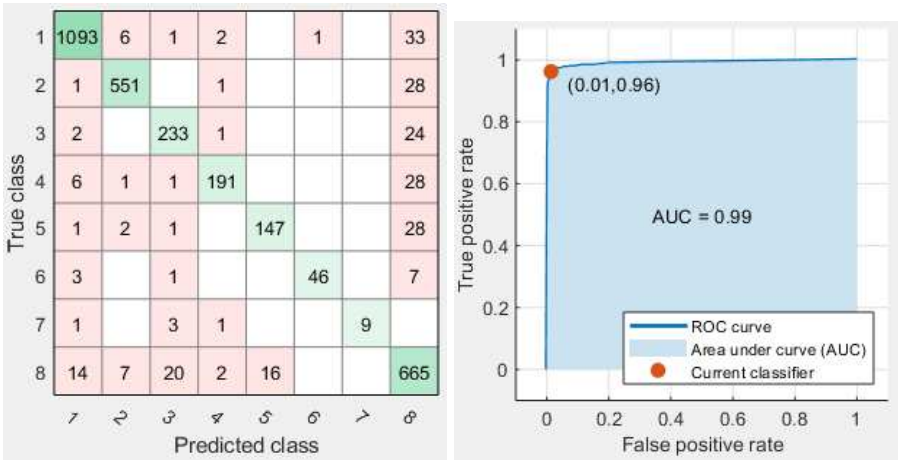

Fig. 2. Confusion Matrix and ROC Curve on WL-OA method on Sequence + Topology Label

7 Conclusion

In this work, we have described graph based methods and kernel based methods for encoding RNA molecules. Then, we applied these methods on our RNA dataset using MATLAB classifiers. We have demonstrated that, graph kernels can be used to classify RNA in high accuracy. We received the best results on the Residue Label (sequence information) and Topology Label using WL-OA with an accuracy of 92.4 percent. We received the worst results on nucleotide sequences using general Sequence Alignment method with an accuracy of 89.2 percent.

References

1. B. Shabash and K. C. Wiese, "Rna visualization: Relevance and the current state-of-the-art focusing on pseudoknots," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 14, no. 3, pp. 696–712, May 2017. [Online]. Available: https://doi.org/10.1109/TCBB.2016.2522421
2. 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.
3. J. Huang, K. Li, and M. Gribskov, "Accurate classification of rna structures using topological fingerprints," *PLOS ONE*, vol. 11, p. e0164726, 10 2016.
4. L. Chen, G. A. Calin, and S. Zhang, "Novel insights of structure-based modeling for rna-targeted drug discovery," *Journal of Chemical Information and Modeling*, vol. 52, no. 10, pp. 2741–2753, 2012, pMID: 22947071. [Online]. Available: https://doi.org/10.1021/ci300320t
5. Z. Miao and E. Westhof, "Rna structure: Advances and assessment of 3d structure prediction," *Annual Review of Biophysics*, vol. 46, no. 1, pp. 483–503, 2017, pMID: 28375730. [Online]. Available: https://doi.org/10.1146/annurev-biophys-070816-034125
6. A. M. e. a. Rybarczyk A, Szostak N, "New in silico approach to assessing rna secondary structures with non-canonical base pairs," *BMC Bioinformatics*, vol. 16,1 276, 2015.
7. 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.
8. 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
9. J.-P. Vert, "The optimal assignment kernel is not positive definite," *ArXiv e-prints*, 2008. [Online]. Available: http://adsabs.harvard.edu/abs/2008arXiv0801.4061V
10. H. Lodhi, "Computational biology perspective: kernel methods and deep learning," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 5, pp. 455–465. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1223
11. "What is fasta format?" [Online]. Available: https://zhanglab.ccmb.med.umich.edu/FASTA/

12. J. M. Shelton and S. J. Brown, "Fasta-o-matic: a tool to sanity check and if needed reformat fasta files," *bioRxiv*, 2015. [Online]. Available: https://www.biorxiv.org/content/early/2015/08/21/024448
13. "Protein data bank contents guide: Atomic coordinate entry format description," *Wwpdb.org*. [Online]. Available: http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html
14. "Protein data bank japan," *Pdbj.org*. [Online]. Available: https://pdbj.org
15. "Nucleic acid database (ndb)," *Ndbserver.rutgers.edu*. [Online]. Available: http://ndbserver.rutgers.edu/
16. "Rcsb pdb," *Rcsb.org*. [Online]. Available: https://www.rcsb.org
17. P. Klosterman, M. Tamura, S. Holbrook, and S. Brenner, "Scor: a structural classification of rna database," *Nucleic Acids Res.*, vol. 30, pp. 392–394, 2002.
18. G. Chojnowski, T. Walen, and J. M. Bujnicki, "Rna bricks - a database of rna 3d motifs and their interactions," *Nucleic Acids Research*, vol. 42, no. D1, pp. D123–D131, 2014. [Online]. Available: http://dx.doi.org/10.1093/nar/gkt1084
19. S. S. Ray, S. Halder, S. Kaypee, and D. Bhattacharyya, "Hd-rnas: An automated hierarchical database of rna structures," *Frontiers in Genetics*, vol. 3, p. 59, 2012. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2012.00059
20. "York RNA Graph Dataset." [Online]. Available: https://www.cs.york.ac.uk/cvpr/RNA.html
21. T. Zok, M. Antczak, M. Zurkowski, M. Popenda, J. Blazewicz, R. W. Adamiak, and M. Szachniuk, "Rna secondary structure annotation based on pdb/mmcif files." [Online]. Available: http://rnapdbee.cs.put.poznan.pl/
22. "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/
23. R. P. W. Duin, E. Pękalska, A. Harol, W.-J. Lee, and H. Bunke, "On euclidean corrections for non-euclidean dissimilarities," in *Structural, Syntactic, and Statistical Pattern Recognition*, N. da Vitoria Lobo, T. Kasparis, F. Roli, J. T. Kwok, M. Georgiopoulos, G. C. Anagnostopoulos, and M. Loog, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 551–561.
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. 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
26. P.-L. Giscard and R. C. Wilson, "The all-paths and cycles graph kernel," *arXiv preprint arXiv:1708.01410*, 2017.