# Classification of RNAs with Pseudoknots using k-mer Occurrences Count as Attributes

Kwan-Yau Cheung[#], Kwok-Kit Tong[#], Kin-Hong Lee, Kwong-Sak Leung
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin N.T. Hong Kong
Email: {kycheung, kktong, khlee, ksleung}@cse.cuhk.edu.hk

***Abstract*— RNAs are functionally important in many biological processes. Predicting secondary structures of RNAs can help understanding 3D structures and functions of RNAs. However, RNA secondary structure prediction with pseudoknots is NP-complete. Predicting whether the RNAs contain pseudoknots in advance can save computation time as secondary structure prediction without pseudoknots is much faster. In this paper, we use k-mer occurrences as attributes to predict whether the RNAs have pseudoknots in the secondary structure. The results show two classifiers can predict 90% of the instance correctly.**

***Index Terms*— pseudoknots, classification, k-mer**

## I. Introduction

RNAs are very important in organisms. They are not only responsible for proteins formation, but also for gene regulation and catalyzing biological reactions [1], [2]. In Bioinformatics, predicting RNA secondary structure with pseudoknots is a very important problem, because it can reveal the 3D structure and functions of an RNA [3], [4]. Pseudoknot is a special motif of RNA secondary structure. They can be found in ribosomal RNAs, telomerase RNAs and viral RNAs [5], [6]. Moreover, they play key roles in many biological processes, like splicing, ribosomal frameshifting, rival genome replication and regulation of translation[7], [8], [9], [10]. Figure 1 shows an example of pseudoknot.

The motivations of the paper are as follow. First, predicting RNA secondary structure with pseudoknots is NP-complete [11] while the problem can be solved in $O(n^3)$ time if pseudoknots are neglected, where $n$ is the sequence length. A lot of computational time can be saved if fast pseudoknots classification is done before structure prediction. Second, there is no previous work related to RNA pseudoknots classification. Third, pseudoknots are RNA secondary structure and RNA secondary structure depends on its sequences. K-mer of RNA sequences should contain the information to predict pseudoknots in RNAs.

In this paper, we use the k-mer occurrences of RNAs as attributes to classify whether pseudoknots exists in RNAs. We first collect RNA sequences from a database called RNA STRAND [12]. Then we perform data preprocessing to generate k-mer occurrences files for the RNAs sequences. Three classifiers from a machine learning software called WEKA [13] is used for classification. We then perform a 10-fold cross-validation test for evaluation.

[#]Equal Contributors

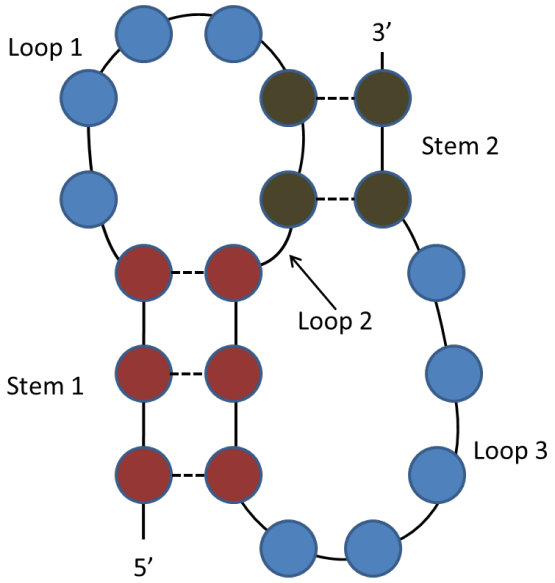



Fig. 1: An example of RNA pseudoknot

The outline of the paper is as follows. Section II describes the problem definition. Section III describes the materials and methods. Section IV describes the results. Section V concludes the paper.

## II. Problem Definition

The problem definition of pseudoknots classification is described here. The input of the problem is a RNA sequence. The output of the problem is to predict whether the input RNA sequence have pseudoknots in the secondary structure.

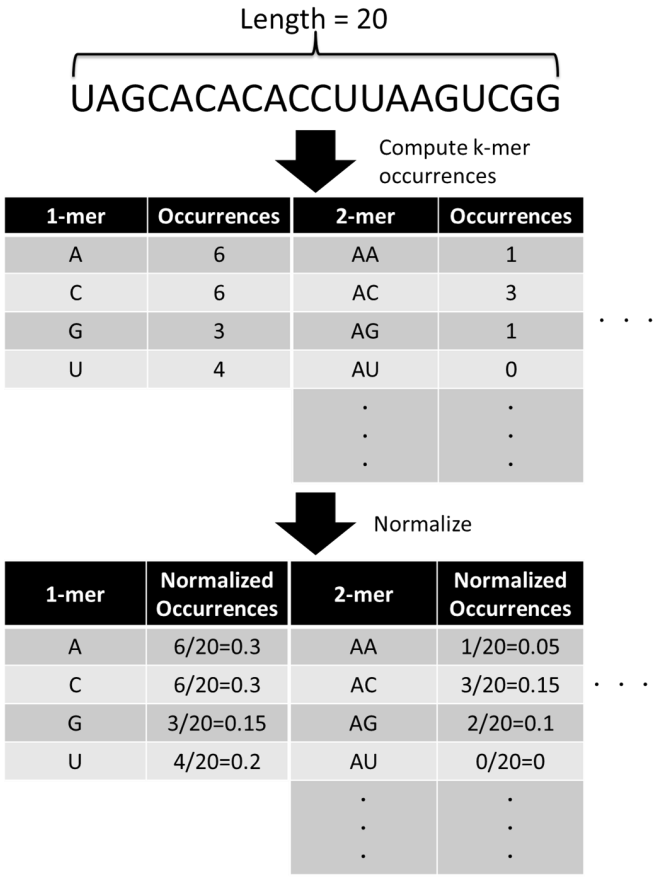
## III. Materials and Methods

In this section, we will describe the data we used for the pseudoknots classification, how the data preprocessing is done and how the classification is done.

### A. Data preparation and preprocessing

We get the RNA sequences with and without pseudoknots from a database called RNA STRAND [12]. The dataset

978-1-4799-3163-7/13/$31.00 ©2013 IEEE




| 1-mer | Occurrences |
|---|---|
| A | 6 |
| C | 6 |
| G | 3 |
| U | 4 |
| . | . |
| . | . |
| . | . |

| 2-mer | Occurrences |
|---|---|
| AA | 1 |
| AC | 3 |
| AG | 1 |
| AU | 0 |
| . | . |
| . | . |
| . | . |

| 1-mer | Normalized Occurrences |
|---|---|
| A | 6/20=0.3 |
| C | 6/20=0.3 |
| G | 3/20=0.15 |
| U | 4/20=0.2 |
| . | . |
| . | . |
| . | . |

| 2-mer | Normalized Occurrences |
|---|---|
| AA | 1/20=0.05 |
| AC | 3/20=0.15 |
| AG | 2/20=0.1 |
| AU | 0/20=0 |
| . | . |
| . | . |
| . | . |

Fig. 2: An example showing the occurrences computation and normalization for a RNA sequence

contains 2333 RNAs sequences with pseudoknots and 2333 RNAs sequences without pseudoknots. A filtering process is carried out to filter out those sequences that contain notations besides "ACGU". After the filtering, there are 1624 RNAs sequences with pseudoknots and 1615 RNAs sequences without pseudoknots remaining. Then, the k-mer occurrences of each RNA sequence are computed. We compute the k-mer occurrences from 1-mer to 6-mer. The k-mer occurrences are then normalized using the length of the RNA sequence. Figure 2 shows an example of the k-mer occurrences counting and the normalization. Finally, six occurrences files with class labels are generated from 1-mer to 6-mer. An example is shown in Figure 3.

### *B. Data Mining*

An open sourced machine learning software WEKA [13] is used for classification. The WEKA version is 3.7.9. Three different classification methods have been used separately for classification. The three classifiers are: a multinomial Naive Bayes classifier (NVM) [14], a multinomial logistic regression model with a ridge estimator (Logistic) [15] and C4.5 decision tree (J48) [16]. We will test all occurrences files on all three classifiers. For each testing, one occurrences file is inputted to one classifiers and a 10-fold cross-validation test on the input data is performed.

1-mer occurrences count file

| 1-mer | A | C | G | U | Class (With Pseudoknots?) |
|---|---|---|---|---|---|
| Sequence1 | 0.3 | 0.3 | 0.1 | 0.3 | Yes |
| Sequence2 | 0.2 | 0.4 | 0.2 | 0.2 | No |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |

2-mer occurrences count file

| 2-mer | AA | AC | ... | UU | Class (With Pseudoknots?) |
|---|---|---|---|---|---|
| Sequence1 | 0.1 | 0.05 | ... | 0.5 | Yes |
| Sequence2 | 0.1 | 0.05 | ... | 0.15 | No |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |

Fig. 3: An example showing the occurrences files for 1-mer (top) and 2-mer (bottom) with class labels

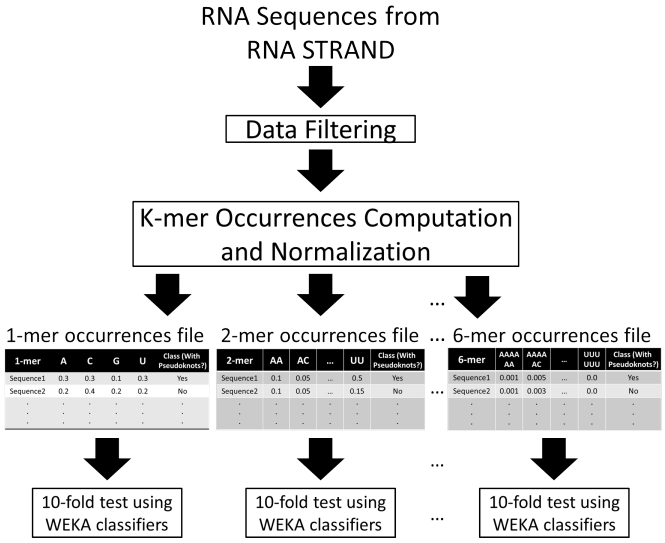



Fig. 4: An complete workflow of the classification process

Before the classification, an attribute selection process is carried out by the attribute selection function of WEKA. The attribute evaluator used is "CorrelationAttributeEval" which "evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class" as mentioned in the WEKA document. In the classification process, we have tested different threshold for the number of attributes selected, which will be shown in Section IV.

Here, we will describe the settings of the three classifiers. For the multinomial Naive Bayes classifier (NVM) and the multinomial logistic regression model with a ridge estimator (Logistic), the settings are both kept as default. For C4.5 decision tree (J48), all settings are kept as default except the minimum numbers of instances per leaf to reduce the tree size. We have tested several thresholds of minimum numbers

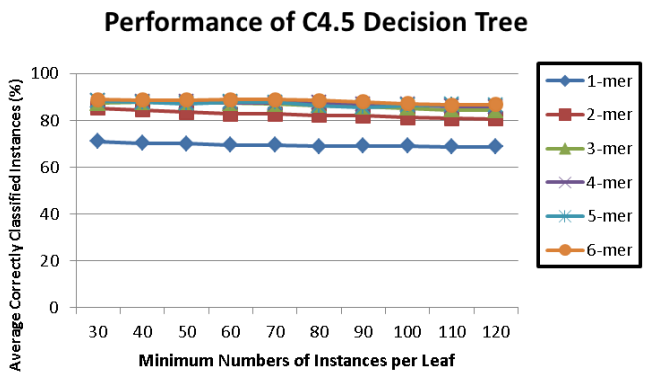



Fig. 5: Performance of C4.5 decision tree. X-axis is the average correctly classified instances in percent. Y-axis is the minimum numbers of instances per leaf in C4.5 decision tree. The graph shows the performance of C4.5 decision trees using different k-mer occurrences files with different thresholds.

| 4-mer | 5-mer | 6-mer |
|---|---|---|
| GUAA | GUAAA | GCAAAC |
| UAAA | UAAAC | AAACCC |
| CAAA | ACAGA | GGUAAA |
| UAGA | CGACA | AAGGUG |
| AAAG | GCAAA | AAGAGC |
| UAAC | AUAGU | CGUAAA |
| AAUA | GGUAA | CUUAAU |
| CGAC | AAACC | ACAGAA |
| UAAU | UGCAA | GACGGG |
| AAUG | ACCAA | CUCCAC |

Fig. 6: First ten k-mers selected by attribute selection process

of instances per leaf. The complete workflow is shown in Figure 4.

## IV. Results

In this section, the performance of three classifiers with different k-mer occurrences as attributes will be described. To evaluate the performance of the classifiers, the number of correctly classified instances is used.

We have tested the performance of C4.5 decision tree (J48) using several thresholds of minimum numbers of instances per leaf. The result is shown in Figure 5. The thresholds are chosen from 30 to 120, which are around 1% to 4% of the total number of instances. The average correctly classified instances only slightly decrease when the minimum numbers of instances per leaf increases.

Figure 7 shows the performance of three classifiers with 1-mer to 6-mer occurrences. Performance of three classifiers increase as the number of attributes selected increases and then converges quickly after 10. C4.5 decision tree and logistic regression perform very well. They can both predict around 90% instances correctly using 3-mer, 4-mer, 5-mer and 6-mer occurrences as attributes. Multinomial Naive Bayes classifier always performs worse than the other classifiers.

The convergences speeds of the classifiers are different. The performance of C4.5 decision tree can converge using around 10 attributes for 3-mer, 4-mer, 5-mer and 6-mer occurrences. Logistic regression needs more number of attributes to converge. The fast convergence of C4.5 decision tree may suggest some k-mers are more important for the formation of pseudoknots in secondary structure. The top ten k-mer selected by the attribute selection process are listed in Figure 6.

## V. Discussion and Conclusion

In this paper, we have used k-mer occurrences of RNA sequences to predict whether the RNA sequences have pseudoknots in the secondary structures. Both C4.5 decision tree and multinomial logistic regression model with a ridge estimator can predict 90% of the instances correctly. This result may further help secondary structure prediction in the future. The fast convergence of the C4.5 decision tree suggests that some k-mers may be critical to the formation of pseudoknots. We will look into the relative position of the k-mers with respect to the known pseudoknots in the future.

## VI. Acknowledgments

This research is partially supported by the Direct Grant of CUHK and the General Research Fund (Project Number: LU310111) of Hong Kong SAR, China.

## References

[1] G. Storz, "An expanding universe of noncoding rnas," *Science*, vol. 296, no. 5571, pp. 1260–1263, 2002. [Online]. Available: http://www.sciencemag.org/content/296/5571/1260.abstract

[2] C. Mello and D. Conte, "Revealing the world of rna interference," *Nature*, vol. 431, no. 7006, pp. 338–342, 2004.

[3] I. T. Jr and C. Bustamante, "How rna folds," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 271 – 281, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283699930012

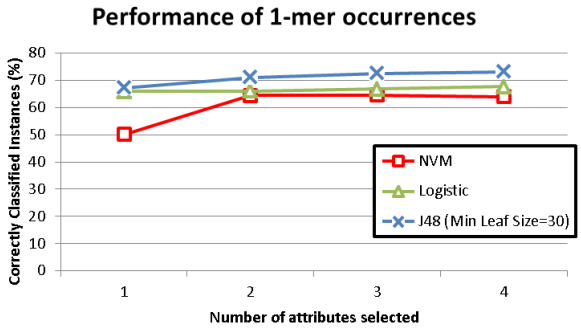
[4] B. A. Shapiro, Y. G. Yingling, W. Kasprzak, and E. Bindewald, "Bridging the gap in rna structure prediction," *Current Opinion in Structural Biology*, vol. 17, no. 2, pp. 157 – 165, 2007, theory and simulation / Macromolecular assemblages. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959440X07000310

[5] F. H. D. van Batenburg, A. P. Gultyaev, and C. W. A. Pleij, "Pseudobase: structural information on rna pseudoknots," *Nucleic Acids Research*, vol. 29, no. 1, pp. 194–195, 2001. [Online]. Available: http://nar.oxfordjournals.org/content/29/1/194.abstract

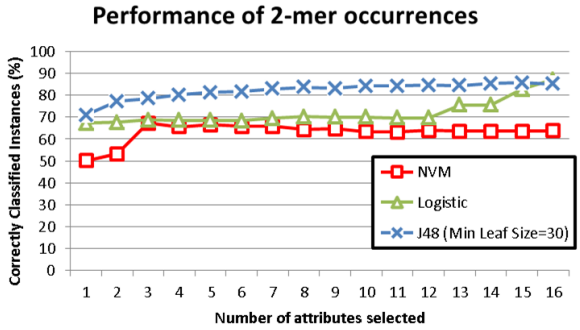
[6] J.-L. Chen and C. W. Greider, "Functional analysis of the pseudoknot structure in human telomerase rna," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 23, pp. 8080–8085, 2005. [Online]. Available: http://www.pnas.org/content/102/23/8080.abstract

[7] I. Brierley, R. Gilbert, and S. Pennell, "Rna pseudoknots and the regulation of protein synthesis," *Biochemical Society Transactions*, vol. 36, no. 4, pp. 684–689, 2008.

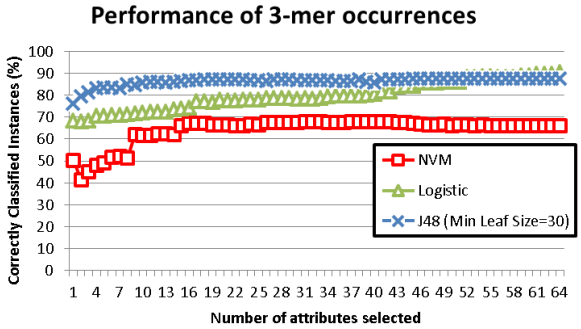
[8] D. W. Staple and S. E. Butcher, "Pseudoknots: Rna structures with diverse functions," *PLoS Biol*, vol. 3, no. 6, p. e213, 06 2005. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.0030213



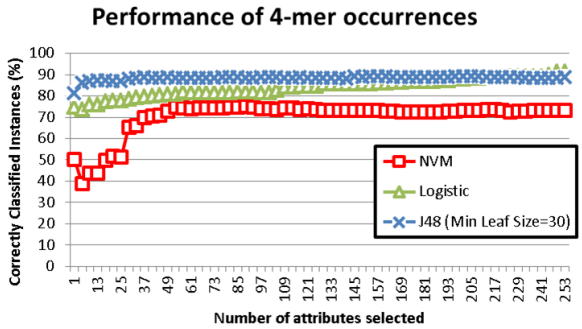

(a) Performance of 1-mer occurrences



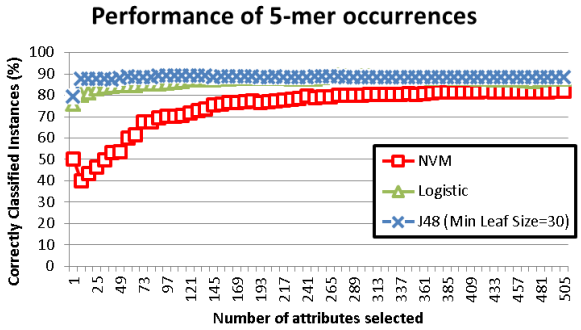

(b) Performance of 2-mer occurrences



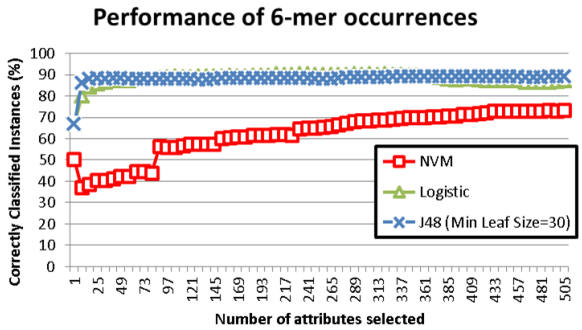

(c) Performance of 3-mer occurrences




(d) Performance of 4-mer occurrences




(e) Performance of 5-mer occurrences




(f) Performance of 6-mer occurrences

Fig. 7: Performance of three classifiers with different k-mer occurrences. X-axis is the average correctly classified instances in percent. Y-axis is the number of attributes selected, which is the threshold of the attributes selection process. Red line with square markers shows the performance of the multinomial Naive Bayes classifier (NVM). Green line with triangle markers shows the performance of multinomial logistic regression model with a ridge estimator (Logistic). Blue line with cross markers shows the performance of C4.5 decision trees (J48) with minimum numbers of instances per leaf = 30.


[9] D. P. Giedroc and P. V. Cornish, "Frameshifting rna pseudoknots: Structure and mechanism," *Virus Research*, vol. 139, no. 2, pp. 193 – 208, 2009, structural motifs controlling the replication cycle of positive strand RNA viruses. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168170208002323

[10] D. P. Giedroc, C. A. Theimer, and P. L. Nixon, "Structure, stability and function of rna pseudoknots involved in stimulating ribosomal frameshifting," *Journal of Molecular Biology*, vol. 298, no. 2, pp. 167 – 185, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283600936684

[11] R. B. Lyngsø and C. N. Pedersen, "Rna pseudoknot prediction in energy-based models," *Journal of computational biology*, vol. 7, no. 3-4, pp. 409–427, 2000.

[12] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon, "Rna strand: the rna secondary structure and statistical analysis database," *BMC bioinformatics*, vol. 9, no. 1, p. 340, 2008.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[14] A. Mccallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 Workshop on 'Learning for Text Categorization'*, 1998.

[15] S. le Cessie and J. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[16] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993, vol. 1.