In Silico Validation of ncRNA-ncRNA Interaction Sites with ncRNAs Represented by k-mers Features

Malik Yousef[1], Walid Khaleifa[2] and Tugba Onal-Suzek[3,4]
[1]Department of Community Information Systems, Zefat Academic College, Zefat, Israel
[2]Computer Science, The College of Sakhnin, Sakhnin, Israel
[3]Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, Turkey
[4]Bioinformatics Graduate Program, Mugla Sitki Kocman University, Mugla, Turkey

Keywords: ncRNA, Machine Learning, Differentiate Reliable ncRNA-ncRNA Interactions, k-mer ncRNA Categorization.

Abstract: A recent catalogue of human transcriptome, namely CHESS database, assembled from RNA sequencing experiments as a part of the Genotype-Tissue Expression (GTEx) Project reported more non-coding RNA genes (21,856) than protein-coding (21,306), revealing an unexpectedly vast amount of transcriptional noise (Pertea *et al*, 2018). In this study, we introduce a workflow coded in KNIME that computationally distinguishes the ncRNA-ncRNA interaction sites with less reliable interaction sites containing less experimentally validated binding sites than the interaction sites with more experimental validation. Duplex structure and k-mer features of the ncRNA-ncRNA binding sites with experimental verification were used as input to the classification workflow. In our analysis, we observed that although duplex structure features had no positive effect on the success rate of the classification, using just the k-mer features, ~80% success could be achieved in categorization of the confidence of the ncRNA-ncRNA binding sites. Our result verified the classification performance of miRNA-mRNA targets using only k-mer features from our previous study (Yousef *et al*, 2018).

1 BACKGROUND

Post transcriptional gene regulation influences protein abundance and its dysregulation is a hallmark for many diseases. With the recent analysis of transcriptional data (Pertea *et al*, 2018), more of the transcriptome is revealed to be originating from non-coding RNAs (ncRNAs). There is a wide variety of ncRNAs such as microRNAs(miRNAs), silencing RNA(siRNAs), Piwi-interacting RNA(piRNAs), small nucleolar RNAs(snoRNAs), small nuclear RNAs(snRNAs), exRNAs, scaRNAs and the long noncoding RNAs (lncRNAs) , differing from each other by nucleotide sequence length, folding and function. All these non-coding RNAs are predicted to be involved in targeting post transcriptional regulation. Although the importance of the miRNAs role in gene regulation is well known to an extent, genetic mechanism of the other types of non-coding RNAs, especially, the role of long non-coding RNAs (lncRNAs) is not that well characterized.

With the recent pan-cancer analysis of the lncRNAs, the dysregulation of the lncRNAs is considered to be widely involved in transcriptional perturbation in various cancer tumor contexts [49]. This recent surge in our knowledge of lncRNAs' role in cancer is mostly attributable to high-throughput sequencing, especially CLIP-seq technology. With the assistance of the high-throughput CLIP-seq technology, functional annotation of the large functional interaction network of lncRNAs with other noncoding RNAs is getting clarified further revealing the previously unknown involvement of ncRNA interactions in cancer transcriptome.

Experimental determination of ncRNA – ncRNA interactions is highly involved, error-prone and naturally dependent on tissue of expression, which is why computational methods have become important. One important problem in the identification of lncRNA-ncRNA interaction network is due to the nonspecific binding of RNA polymerase to random or very weak sites in the genome, causing the 95% transcriptional noise Several automated classification

Yousef, M., Khaleifa, W. and Onal-Suzek, T.
In Silico Validation of ncRNA-ncRNA Interaction Sites with ncRNAs Represented by k-mers Features.
DOI: 10.5220/0007361801680173
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 168-173
ISBN: 978-989-758-353-7

methods have been applied to distinguish the plethora of ncRNA classes such as RNA-CODE (Yuan and Sun, 2013) based on the alignment of short reads, or RNAcon (Panwar *et al.*, 2014) and (Childs *et al.*, 2009) based on the multi-feature extraction and full-sequence analysis. The last two of the listed tools RNAcon and GraPPle incorporate the secondary RNA structure into the machine learning algorithms. Another recent tool, spongeScan, uses a k-mer-complementarity based algorithm to predict the miRNA response elements in lncRNAs(Furió-Tarí *et al,* 2016).

In this study, similar to RNAcon and GraPPle, we tested incorporating both the k-mers features and the duplex structure information to a KNIME machine learning workflow. But rather than classifying the ncRNA classes, we tried to classify the ncRNA-ncRNA binding data and compared the accuracy of the k-mer based and structure features. In addition, we introduced three new set of features which are extracted from the sequence and summarize the distance between k-mers. These new set of features named inter k-mer distance, k-mer location distance and k-mer first-last distance were compared to k-mer and all other published features describing an ncRNA-ncRNA interaction.

The remainder of the paper is organized as follows: In Data section, the three types of data used by the classification algorithm are described. In Methods section, feature selection, classification and performance evaluation of the proposed algorithms are explained. In the Results and Discussion section, a performance summary of each categorization approach on the three set of data are presented. Finally, the limitations and future work are outlined in the Conclusions section.

2 DATA

The first set of data was obtained from StarBase(Li *et al.*, 2014) where positive lncRNA-ncRNA interaction data is the subset with at least 5 supporting experiments confirming the interaction, labelled the “highest stringency” by Starbase. Different pools of positive classes were generated by considering different number of supporting experiments. We have 1950 examples with the 5 supporting experiments (very high stringency), 3280 with 3 (high stringency) and 8143 with 2 (medium stringency).

Negative lncRNA-ncRNA interaction data is the subset of Starbase with only 1 supporting experiment with an lncRNA not listed in the CHESS 2.0[44] catalog. This filter yield in 2043 examples to form the negative class. CHESS is the most comprehensive and recent catalog of coding and non-coding transcripts, therefore our assumption was if an lncRNA is not documented in this vast catalog and contributes to an interaction that is observed only once experimentally, the interaction data might not be as reliable. The second set of data was downloaded and parsed from StarBase(Li *et al.*, 2014) and mirTarbase (Chou *et al.*, 2018) for the organism human via in-house Perl scripts.

Table 1: Description of the data sets used in our study. The data sets were downloaded from StarBase[30] and mirTarbase [31]. Each entry has the interaction name of the data, the name of the source, number of examples and the number of unique miRNA involved.

	Source	**Number of examples**	**Number of unique miRNA**
miRNA-lncRNA interactions	StarBase	10199	278
miRNA-circRNA interactions	StarBase	9997	24
miRNA-pseudogene interactions	StarBase	16133	277
miRNA-sncRNA interactions	StarBase	3293	273
miRNA-mRNA	mirTarBase	3121	52

The third set of data was generated on cancer implication of the lncRNA-miRNA interactions on StarBase. Positive data 603 examples contain the lncRNA-miRNA interactions involved in at least 1 cancer type by pan-cancer grouping derived from TCGA for the 14 cancer types (>6000 samples). Negative data with 9513 examples contain all the lncRNA-miRNA interactions which are not detected in any of those 14 cancer types.

3 METHODS

3.1 Parameterization of ncRNAs

The first step in applying machine learning to the current data is to represent the data in vector space $v=(v_1,v_2,\ldots,v_n)$, where each component v relates to a specific feature and where n is the number of features. One simple way of representing sequences that consist of 4 nucleotide letters is by employing k-mers. In a recent study, we have shown that k-mers are sufficient to allow to categorize pre-miRNAs into species (Yousef *et al.*, 2006).

3.2 K-mer Features

Many studies performing ncRNA analysis used sequence-based features. Sequence-based features can be words or short sequence of nucleotides {A,U,C,G,} with the length k (so called k-mers or n-grams). For example, 1-mers are the 'words' A, U, C, and G; 2-mers are the words AA, AC, …, UU, and 3-mers lead to 64 (4^3) short nucleotide sequences ranging from AAA to UUU. In general, the number of all k-mers up to and including length k is$\sum_0^k 4^i$.

There were other studies with longer k (Cakir and Allmer, 2010), but in this study 1-, 2-, and 3-mers were used as features. The k-mer frequencies are calculated by dividing raw counts by the length of the sequence (i.e., len(sequence) - k + 1). Hence, for k-mers with $k = \{1, 2, 3\}$, 84 features were calculated per example. The k-mer frequency ranges between 0 (if the k-mer is not present in the sequence) and 1 (if the sequence is a repeat of a mononucleotide; no such case was present in the data).

3.3 Secondary Features

Following the study of (Yousef *et al.*, 2006) we generated features considering the secondary structure of ncRNA-ncRNA interactions:

1. Number of Base Pairs
2. Number of Bulges
3. Number of Loops
4. Number of bulges with length *i* , *i=1 to 6*
5. Number of bulges with length greater than 6
6. Number of loops with length *I*, *i=1 to 6* (odd number capture asymmetric loops)
7. Number of loops with length greater than 6.

A KNIME workflow (Berthold *et al.*, 2008) was created to extract those features using the secondary structure obtained from the mirBase (Griffiths-Jones, 2010) and StarBase(Li *et al.*, 2014).

3.4 Feature Vector and Feature Selection

In this study we considered different kinds of features, k-mer-based features, the novel k-mer distance features, and 831 features that were used in the study of ('Sacar MD, Allmer J. Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. 2013 8th Int. Symp. Heal. Informatics Bioinforma.IEEE; 2013 p. 1--6.', no date). We also used information gain measurement (Shaltout *et al.*, 2014) as implemented in KNIME (version 3.1.2) (Berthold *et al.*, 2008) for feature selection when we combined different kind of features.

3.5 Classification

A random forest (RF) classifier was implemented in KNIME (Berthold *et al.*, 2008). The classifier was trained and tested with a split of 80% training and 20% testing data. Equal number of negative and positive examples were forced to during the 100-fold Monte Carlo cross-validation (MCCV) (Xu and Liang, 2001) for model establishment.

3.6 Model Performance Evaluation

For each established model, we calculated a number of statistical measures like the Matthews's correlation coefficient (MCC) (Matthews, 1975), sensitivity, specificity, and accuracy for evaluation of model performance. The following formulations were used to calculate the statistics (with TP: true positive, FP: false positive, TN: true negative, and FN referring to false negative classifications):

Sensitivity (SE, Recall) = TP / (TP + FN)
Specificity (SP) = TN / (TN + FP)
Precision = TP / (TP + FP)
F-Measure = 2 * (precision * recall) / (precision + recall)
Accuracy (ACC) = (TP + TN) / (TP + TN + FP + FN); ACC

$$\text{MCC} = \frac{(\text{TP/TN}-\text{FP/ FN})}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FN})(\text{TN+FP})}}$$

Average of 100-fold MCCVs were reported as performance.

4 RESULTS

Several different datasets were tested using k-mers and duplex structures. Table 1 shows that just using the duplex structure was not sufficient to distinguish between the two classes. Interestingly using just, the non-coding sequence k-mers features was able to give a good signature for separating different datasets.

With the two sets of data; miRNA-lncRNA (Stringency 5) vs miRNA-lncRNA (Stringency 1); e have compared the performance of the several different kinds of features (First row in Table 1). Duplex features yielded a low accuracy of 0.57 which means that the structure alone does not contribute to the information to distinguish these two classes. However, ncRNA k-mer features yielded an overall accuracy of 0.82 (AUC is 0.91) which indicates that the ncRNA sequence coding information was sufficient to distinguish between two classes. When the difference between the different types on non-coding miRNA types of the two classes are tested, we failed to see any positive results.

We have tested the influence of the degree of the stringency and observed that with the presence of more experimental data supporting the interaction (higher stringency), the performance of the method improved (See Table rows 2-4).

Additionally, we have tested the differences between the interactions of miRNA-ncRNA from the interaction of lncRNA-miRNA. Results in row 5 and 6 show that based on k-mer feature generated from non-coding region we are able to distinguish between those classes with accuracy of 0.8, however, the duplex structure alone is not able to provide a sufficient separation.

Row 7 shows the results of classification between two classes that we have generated as a sub-data that we call the third set (See Data Section) that related to the pan-cancer grouping rate of lncRNA-miRNA interaction. The accuracy indicates that although there is a difference between the two classes, still the difference is not as remarkable as other cases in this study.

Rows 8 to 13 shows the results of all combination of the two-classes of the 1) miRNA-lncRNA interactions, 2) miRNA-circRNA interactions, 3) miRNA-pseudogene interactions and miRNA-sncRNA interactions. The high classification accuracy rates of all these three combinations indicate a good separation.

Table 2: Average performance for 100-fold MCCV using a random forest classifier and a split of 80% training and 20% testing employing on different datasets. AUC: area under curve.

Dataset used for classification	**Sub-Type**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
miRNA-lncRNA (Stringency 5) vs miRNA-lncRNA (Stringency 1) and in CHESS (#ncRNA=1950 , #chess=2043)	duplex features	0.61	0.52	0.57	0.59
	ncRNA k-mer features	0.87	0.77	0.82	0.91
	duplex + ncRNA k-mer features	0.86	0.77	0.81	0.90
	miRNA k-mer features	0.53	0.60	0.56	0.58
miRNA-lncRNA (Stringency 2) vs miRNA-lncRNA (Stringency 1) and in CHESS #ncRNA=8143 , #chess=2043	ncRNA k-mer features	0.80	0.65	0.72	0.82
miRNA-lncRNA (Stringency 3) vs miRNA-lncRNA (Stringency 1) and in CHESS (#ncRNA= 3280 , #chess=2043)		0.84	0.72	0.78	0.87
miRNA-lncRNA (Stringency 5) vs miRNA-lncRNA (Stringency 1) and in CHESS #ncRNA=1950 , #chess=2043		0.87	0.77	0.82	0.91
mirTarBase(has- miRNA:mRNA) vs miRNA-lncRNA (Stringency 5) #miRNA:mRNA= 3121 , #ncRNA=1950	duplex features	0.64	0.65	0.64	0.68
	ncRNA k-mer features	0.85	0.75	0.80	0.87
	ncRNA k-mer features + duplex	0.85	0.74	0.80	0.87
mirTarBase(has- miRNA:mRNA) vs miRNA-lncRNA (Stringency 3) #miRNA:mRNA= 3121 , #ncRNA=3280	ncRNA k-mer features	0.84	0.75	0.79	0.87
ncRNA Cancer vs ncRNA non cancer (#pos = 603, #neg = 9513)	ncRNA k-mer features	0.75	0.62	0.68	0.75
lncRNA vs circRNA (#lcRNA =10199 , #circRNA = 9997)	ncRNA k-mer features	0.91	0.78	0.85	0.92
lncRNA vs sncRNA (#lcRNA =10199 , #sncrRNA = 3293)		0.83	0.76	0.79	0.89
lncRNA vs psuedoGene (#lcRNA =10199 , #psuedogene=16133)		0.85	0.76	0.80	0.90
circRNA vs sncRNA (#circRNA = 9997, #sncrRNA = 3293)		0.87	0.83	0.85	0.93
circRNA vs psuedoGene (#circRNA = 9997, #psuedogene=16133)		0.86	0.81	0.83	0.92
sncRNA vs psuedoGene (#sncrRNA = 3293,, #psuedogene=16133)		0.88	0.76	0.82	0.90

5 DISCUSSION

The increased accuracy of the predictions by using data with more experimental evidence suggest that the ncRNA-ncRNA interaction data with scarce experimental support is not reliable enough to avoid misclassifications. Concluding, therefore, in spite of the overall good performances of our classification approach, supplementing the ncRNA-ncRNA interaction data with more experimental evidence will aid in increasing the accuracy of the classification workflow.

6 CONCLUSIONS

In this study we have tested different datasets to study different types of non-coding RNA interactions and the differences between those interactions. In our experiments we have tested two main kind of features, k-mers features and duplex features. Interestingly we have discovered that using the k-mers features is sufficient to distinguish between different types of noncoding RNA interactions. We didn't observe any positive contribution of the duplex features.

7 AVAILABILITY OF DATA AND MATERIALS

All of the ncRNA data was obtained from www.mirbase.org and starbase.sysu.edu.cn.

8 FUNDING

The work was supported by Zefat Academic College to MY.

ACKNOWLEDGEMENTS

The work was supported by Zefat Academic College to MY.

REFERENCES

Berthold, M. R. *et al.* (2008) 'KNIME: The Konstanz Information Miner', in *SIGKDD Explorations*, pp. 319–326. doi: 10.1007/978-3-540-78246-9_38.

Cakir, M. V. and Allmer, J. (2010) 'Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii', in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*. Ankara, Turkey: IEEE, pp. 31–38. doi: 10.1109/HIBIT.2010.5478909.

Childs L, Nikoloski Z, May P, Walther D. Identification and classification of ncRNA molecules using graph properties. Nucleic Acids Res. 2009;37(9):66.

Chiu HS, Somvanshi S, Patel E, Chen TW, Singh VP, Zorman B, Patil SL, Pan Y, Chatterjee SS; Cancer Genome Atlas Research Network, Sood AK, Gunaratne PH, Sumazin P. Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. Cell Rep. 2018 Apr 3;23(1):297-312.e12. doi: 10.1016/j.celrep.2018.03.064.

Chou, C.-H. *et al.* (2018) 'miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions.', *Nucleic acids research*, 46(D1), pp. D296–D302. doi: 10.1093/nar/gkx1067.

Furió-Tarí P, Tarazona S, Gabaldón T, Enright AJ, Conesa A. spongeScan: A web for detecting microRNA binding elements in lncRNA sequences. Nucleic Acids Res. 2016 Jul 8;44(W1):W176-80.

Griffiths-Jones, S. (2010) 'miRBase: microRNA sequences and annotation.', *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 12, p. Unit 12.9.1-10. doi: 10.1002/0471250953.bi1209s29.

Li, J. H. *et al.* (2014) 'StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data', *Nucleic Acids Research*. doi: 10.1093/nar/gkt1248.

Malik Yousef, Dalit Levy, Jens Allmer: Species Categorization via MicroRNAs - Based on 3'UTR Target Sites using Sequence Features. BIOINFORMATICS 2018: 112-118

Matthews, B. W. (1975) 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme', *BBA - Protein Structure*, 405(2), pp. 442–451. doi: 10.1016/0005-2795(75)90109-9.

Panwar B, Arora A, Raghava GP. Prediction and classification of ncRNAs using structural information. BMC Genomics. 2014;15(1):127.

Pertea M, Shumate A, Pertea G, Varabyou A, Chang Y, Madugundu AK, Pandey A, Salzberg S (2018) 'Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise'. bioRxiv

'Sacar MD, Allmer J. Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. 2013 8th Int. Symp. Heal. Informatics Bioinforma.IEEE; 2013 p. 1--6.' (no date).

Shaltout, N. A. N. *et al.* (2014) 'Information gain as a feature selection method for the efficient classification of Influenza-A based on viral hosts', in *Proceedings of the World Congress on Engineering*. Newswood Limited, pp. 625–631.

Xu, Q.-S. and Liang, Y.-Z. (2001) 'Monte Carlo cross validation', *Chemometrics and Intelligent Laboratory Systems*, 56(1), pp. 1–11. doi: 10.1016/S0169-7439(00)00122-2.

Yousef, M. *et al.* (2006) 'Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier', *Bioinformatics*, 22(11), pp. 1325–1334. doi: 10.1093/bioinformatics/btl094.

Yuan C, Sun Y. RNA-code: a noncoding RNA classification tool for short reads in NGS data lacking reference genomes. PloS one. 2013;8(10):77596.