Viral MicroRNA Precursors Discovery with Self-containment Index and Derivative Features using Back-propagation Neural Network

Songtham Anuntakarun, Supawadee Ingsriswang, Warin Wattanapornprom and Supatcha Lertampaiporn[1]
Information Systems Laboratory (ISL),
National Center for Genetic Engineering and Biotechnology (BIOTEC)
Pathumthani, Thailand
e-mail: supatcha.ler@biotec.or.th[1]

Abstract — **This article presents efficiency of self-containment index and its derivative features in back-propagation neural network in order to discover new viral miRNA precursors. This tool mainly discriminates the viral miRNA precursors from coding sequences and pseudo miRNA precursors. It was trained with known dataset from miRBase, pseudo miRNA precursors and coding sequences from ViralmiR and NCBI database respectively. Top 20 features were selected using GainRatio out of totally 115 features including sequence-based features, secondary structure, base-pair, triplet sequence-structure and structural robustness features. In particular, 6 out of top 20 features ranked by GainRatio are found to be the derivatives of self-containment index feature. The results show that the back-propagation neural network achieves up to 97% accuracy and only 1% false positive rate on the test set. It has been compared with ViralMiR and achieved higher in geometric mean which reflects compromising balance between sensitivity and sensitivity performances.**

Keywords - Virus miRNA precursors; Structural robustness features; Back-propagation neural network;

I. INTRODUCTION

Recently, non-coding RNAs (ncRNA) are found to play important role in regulation of gene expression, immunity pathway, posttranscriptional process, oncogenesis, and development of organisms [1]. MicroRNA (miRNA) is a major class of short ncRNAs which contains about 22 nucleotide sequences. They function in inhibition of translational process or degradation of coding sequence of plants [2] and mammals [3]. Viruses also encode miRNAs that contribute to the complex interaction with their hosts. In addition, the first report of virus encoded miRNAs was found in herpesvirus family [4]. Viral miRNAs can function in both cis and trans by regulating accumulation of viral products and down regulating of host transcripts [5] such as evading immunity system of the host by indirectly decreasing levels of viral proteins and consequent antigenicity or suppressing components of immune systems to respond virus directly [6], regulating of virus replication, cell cycling, controlling of apoptosis pathway by targeting host genes related pro-apoptotic and maintaining viral latency in the host [7]. Moreover, many researches have reported that human diseases associated with virus-encoded miRNA such as gastric cancer [8].

Computational methods play an important role in identification of novel miRNAs. The hairpin structure of miRNA precursor (pre-miRNA) is an important characteristic usually used in the computational identification of pre-miRNAs. However, there is evidence that the viral pre-miRNAs display rapid sequence evolution [11]. They also show low degree of conservation among the viral pre-miRNAs [12]. Accordingly, computational comparative methods which relied only on sequence or structure alignment may not be sufficient.

In addition to comparative methods which relied on conservation of alignment analysis [13], non-comparative or machine learning approaches are also employed to identify the pre-miRNAs. These techniques have been developed to discriminate pre-miRNAs from pseudogene, pseudo pre-miRNAs, snoRNA and others using base-pair, sequence and structure features. Among these features, minimum free energy (MFE) or thermodynamic score of stem-loop hairpin pre-miRNA structure which based on calculation from MFOLD software [14] have been widely used as a main feature since pre-miRNAs have been reported to possess lower folding free energy than the random sequences [15]. However, the MFE feature is not sufficient to describe the characteristic of the pre-miRNAs. Therefore, other significant features should be taken in to account to improve the prediction performance. Currently, many machine learning tools that applicable to predict the pre-miRNAs including Triplet-SVM [16], MiPred [17], miPred [18], miR-KDE [19], microPred [20], MirenSVM [21], miR-BAG [22] and HeteromiRPred [23] have been constantly released. Recently, in 2015, a web based miRNAs identification tool specifically trained for viral pre-miRNAs using support vector machine called ViralmiR [24] has been proposed. ViralmiR uses 54 features including triplet element features, sequential features, thermodynamic features, base-pair-related features and RNAfold-related features to predict the viral pre-miRNAs. In addition to these features, this research proposes to use additional structural robustness feature called self-containment index (SCI) [25] as well as its derivative features [23] to identify viral pre-miRNAs using Back-propagation Neural Network (BNN). With the high mutation rates of RNA viruses, these additional SCI features were extracted by simulating of the intrinsic robustness characteristic of viral

pre-miRNAs to improve the efficiency viral pre-miRNAs identification.

The remaining sections of this paper are organized as follows. Research methodology, datasets and features are introduced in Section II. The results are discussed in Section III. Finally, Section IV concludes this work.

II. METHODOLOGY

A. Research Methodology

The experiments are divided into two main experiments including tool research and development and tool verification. For the tool research and development, six classifier algorithms including Naïve Bayes (NB), Random forest (RF), K-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Tree (C4.5) and Back-propagation Neural Networks (BNN) were selected using 5-fold cross validation. These classifiers are provided in the open source data mining software package celled Weka [26]. Each of them was trained and tested using the default parameters. In this process, the datasets are divided into training set and test set. The process begins with feature extraction and then feature selection, 115 features from 5 group of features were selected using GainRatio [27]. Then, 5-fold cross validation were performed using the training set in order to select the best classifier in terms of accuracy (ACC), sensitivity (Sn), specificity (Sp), false positive rate (FPR) and receiver operating characteristics (ROC) area. The selected classifier was later used in verification process. It would be benchmarked on the test set against new 2015 released tool called ViralmiR. The work flow of these processes are illustrated in Figure 1.

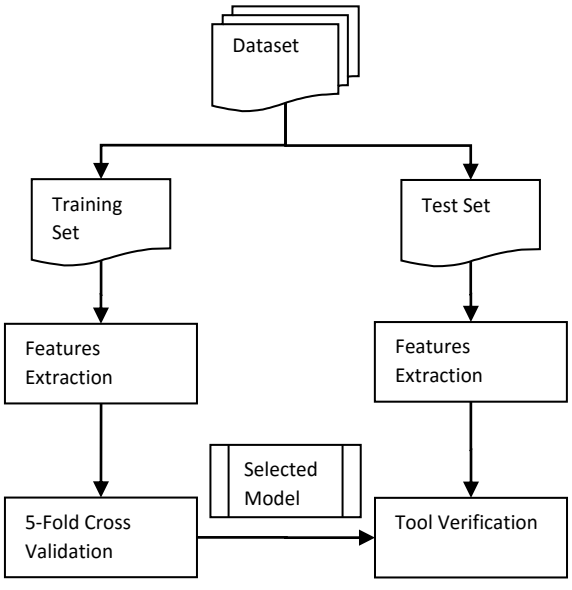

Figure 1. Work flow of tool research development and verification processes

B. Datasets

The datasets in this research can be divided into positive and negative datasets. The positive dataset is the viral pre-miRNAs sequences obtained from miRBase version 19 [28], while the negative datasets were obtained from two data sources including non-viral pre-miRNAs and non- miRNA sequences from ViralmiR and coding sequence of virus from NCBI database. The training set includes 211 miRNA sequences randomly selected from the positive dataset and 8,875 coding sequences randomly selected from the negative dataset. This covers 80% of each of the datasets. The remaining 20% data were reserved as the test data.

C. Features and Feature Selection

In this research, features are obtained from many ncRNA identification researches [16][17][18][19][20][21][22][23]. They can be grouped into 5 categories including sequence-based features, secondary structure features, base-pair features, triplet sequence-structure features and structural robustness features as summarized in Table 1. To improve the performance of the classifiers, the feature selection was applied in order to filter the irrelevant and redundant features. The filtering method called GainRatio or the ratio of information gain and intrinsic value was used to select the top 20 features and are reported in the following section. Detailed of all 115 features categorized in to 5 groups are described below.

1. Sequence-based features

This group of features include mononucleotide and dinucleotide calculated more detailed in [29].

2. Secondary structure features

The secondary structure features include minimum free energy (mfe), the ratio of mfe to the length of the sequence (dG), the ratio of dG to GC content ratio (mfe1), the ratio of dG to Stem (mfe2 - see base-pair features section below for the definition of Stem) , the ratio of dG to number of loops (mfe3), the ratio of mfe to total number of base pairs (mfe4), the ratio of mfe and GC content to total number of base-pair in structure (mfe5) [18], the ratio of Shannon entropy to length (dQ), the compactness degree of tree-graph represent in the sequence (dF), structure entropy (dS), the ratio of structure entropy to length (dSL), the ratio of structure entropy to loop (dS/loop), the structure enthalpy (dH), the ratio of structure enthalpy to length (dHL), the ratio of structure enthalpy to loop (dH/loop), the melting temperature of the structure (Tm), the ratio of melting temperature of the structure to length (TmL), the melting energy of the structure to loop (Tm/loop) [20], ensemble free energy (efe), the ratio of efe to length (nefe), z score of X, where X is number of standard deviation in which X differs from the mean of all shuffled (X) sequences with same dinucleotide composition (zX), the probability of mfe given RNA sequence is different from a distribution computed with random sequences (Prob) [14].

TABLE I. LIST OF 115 FEATURES USED IN THIS RESEARCH

Feature groups	No of features	Feature symbols
Sequence based – features	20	%A, %U, %G, %C, %AA, %AU, %AG, %AC, %UA, %UU, %UG, %UC, %GA, %GU, %GG, %GC, %CA, %CU, %CG, %CC
Secondary structure features	26	MFE, efe, MFE1, MFEI2, MFE3,MFE4, MFE5,dG, dQ,dF, Prob, zG,zQ, nefe, Freq, diff, Tm, TmL, Tm/loop, dS, dSL, dS/loop, dH, dHL, dH/loop
Base-pair features	26	dP, zP, div, tot_bp/loop, A-U/L, G-U/L, G-C/L, %A-U/stem, %G-C/stem, %G-U/stem, Probpair1–10, Avg_PP, NonBP A, NonBP C, NonBP G, NonBP U, Non BPP
Triplet-sequence structure features	32	A(((, A((., A(.., A(.(,A.((,A.(.,A..(, A. . . , C(((, C((., C(.., C(.(, C.((, C.(., C..(, C. . . , G(((, G((., G(.., G(.(, G.((, G.(., G..(, G. . . , U(((, U((., U(.., U(.(, U.((, U.(., U..(, U...
Structural robustness features	11	SCI, SC/non_tot, SCxMFE/Mean, SCxabsZG, SCxdP, SC/len, SC/auLgcLguL, SC/nonauLgcLguL, SC/1dP, SC/nonA, SC/nonU

3. Base-pair features

The base-pair features include the ratio of number of base pairs in secondary structure to loop (tot_bp/loop), the ratio of tot_bp to length (dP), the number of standard deviation from the mean of all shuffled dP with identical dinucleotide composition (zP), the distance of structural base-pair (Div), the number of base X paired with base Y in the secondary structure (X-Y), the percentage of X-Y (%X-Y), the number of all loops in secondary structures (Loop), the number of all stems found in the secondary structure (Stem), the sum of pairing probabilities for each pair of AA, AC, AG, AU, CC, CG, CU, GG, GU, and UU respectively (Probpair1-10), the non-pairing probability for nucleotide X, where X can be A, C, G or U (NonBp_X), and the overall non base-pairing probability (Non_BPP) [30].

4. Triplet sequence-structure features

The triplet sequence-structure features are the local contiguous structure obtained by repeatedly counted the matched patterns found in every three adjacency nucleotides. The eight possible compositions are "(((", "((.", "(..", "(.((", ".(((", ".(.", "..(" and "…", where an opening parenthesis "(" denoted paired nucleotide while a dot "." denoted unpaired nucleotide. Counting all the possible middle nucleotide, there are 32 possible structure sequence combinations, such as "G(((", "A.((", etc [21].

5. Structural robustness features

The features in this group include the self-containment index score (SC) and its derivative features. SC is calculated as followed: for each RNA sequence *w* of length *L* folding into a particular minimum free energy secondary structure, the original sequence then embedded in between randomly generated sequence *x* and *y* of equal length, forming a concatenated molecule *xwy* of length 3*L*. The embedded larger sequence is folded and measured the proportion of the original structure preserved in the larger structure. This process repeats 1000 times to generate a single statistic value ranging from 0.0 to 1.0, where 1.0 indicates a maximal degree of self-containment. In particular, the characteristic of SC score of pre-miRNAs shows value is in the range between 0.85 – 0.98 [25]. Derivative features of SC scores included SC/non_tot, SCxMFE/Mean, SCxabsZG, SC/1dP, SCxdP, SC/len, SC/auLgcLguL, SC/nonauLgcLguL, SC/nonA and SC/nonU [23].

The triplet-sequence structures features are extracted using and Triplet-SVM [16] and structural robustness features are extracted using self-containment feature extraction program [25]. These programs require a secondary structural folding program called RNAFold [31] in the Vienna RNA package. Such features extraction programs were embedded in the classification tool using Java and Perl script [23].

D. Performance Measurement

In this research, the performance of classifier model were assessed using Accuracy (ACC), Sensitivity (Sn), Specificity (Sp) False positive Rate (FPR) and Geometric mean (GM). The performance matrices are calculated as followed:

$$\mathrm{ACC} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$$

$$\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$$

$$\mathrm{Sp} = \mathrm{TN} / (\mathrm{TN}+\mathrm{FP})$$

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{TN}+\mathrm{FP})$$

$$\mathrm{GM}=\sqrt{(\mathrm{SN}\times\mathrm{SP})}$$

,where TP, TN, FP, and FN are the numbers of true positive prediction, true negative prediction, false positive prediction and false negative prediction, respectively. In addition, receiver operating characteristic (ROC) curves were plotted in order to visualize the trade-off between the true positive and false positive rates for every possible cut off and area under the ROC curve (AUC) were also calculated to compare the performance of the classifiers.

III. RESULTS

When there are large number of features in the feature space, feature selection is an important step to select a smaller subset of informative and relevant features to build a better effective classification model. Among 115 features, top 20 features obtained from GainRatio method are used in the model selection process including 7 secondary-based features, 4 base-pairing features and 9 structural robustness features as summarized in Table 2. According to the GainRatio, it can be seen that the features of the SC derivative features have achieved up to six of the top ranks. This implies that these set of features significantly contribute to identify the viral pre-miRNAs in the sequences.

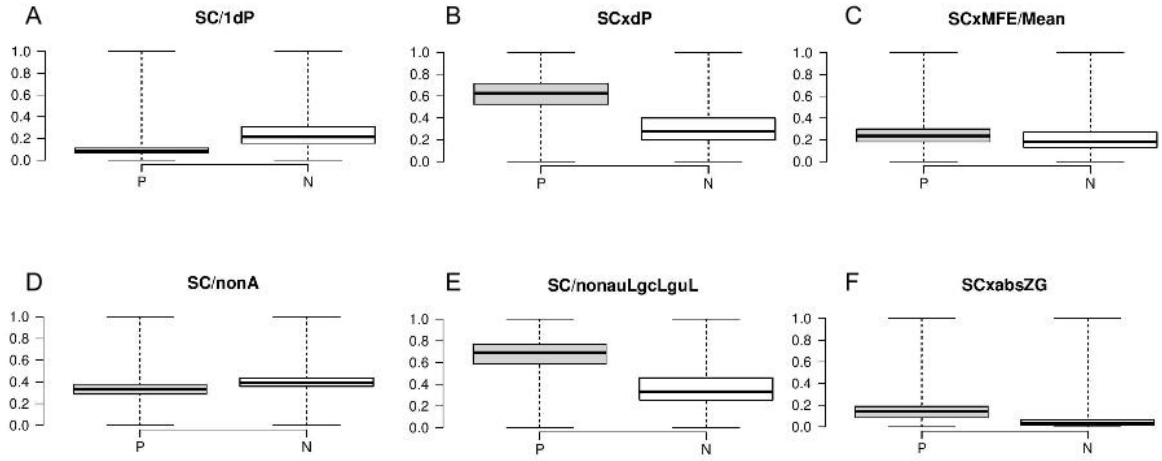

Figure2. Box plots of six SC-derivative features. P and N represent positive and negative dataset, respectively. The Y axis shows the score.

TABLE II. TOP 20 FEATURES FROM GAINRATIO METHOD

Features	GainRatio
SC/1dP	0.21434
SCxdP	0.15199
SCxMFE/Mean	0.1425
SC/nonA	0.14123
SC/nonauLgcLguL	0.12453
SCxabsZG	0.11443
Non_BPP	0.09831
dP	0.09492
Mfe1	0.09297
zG	0.07895
SC/non_tot	0.07606
SCI	0.07595
SC/len	0.07514
zP	0.07149
Prob	0.06733
Mfe3	0.05394
Tm/Loop	0.05314
Mfe2	0.05044
dG	0.04968
MeanBP	0.04765

The top six ranks features based on gain ratio are SC/1dP, SCxdP, SCxMFE/Mean, SC/nonA, SC/nonauLgcLguL and ScxabsZG. To gain further insight into theses features, the score of the features were normalized such that the values range from 0 to 1 and then plotted using box plots. Figure 2 illustrates the discriminative power of these six features in discriminating between the positive and negative examples of viral pre-miRNAs.

As shown in Figure 2, it is observed that the positive data of SCxdP, SCxMFE/Mean, SC/nonauLgcLguL and ACxabsZG have higher score than the negative data while the positive data of SC/1dP and SC/nonA has lower score than the negative data. These observations confirm the hypothesis that structural robustness features are informative features for classifying the viral pre-miRNAs.

The results of 5-fold cross validation are shown in Table 3. According to the results, RF and BNN algorithms show high accuracies, specificities and low false positive rates compared to the others. Moreover, both models offer comparable performance in ROC area. Thereby, to select between RF and BNN model, their performances were verified using another independent test set as shown in Table 4.

This test set has totally 359 sequences including non-viral pre-miRNAs, non-miRNA precursor and coding sequences in virus. The results of the performance of RF and BNN demonstrate that BNN has better performance in accuracy and sensitivity than RF while RF offers lower false positive rate. In order to measure the balance between classification performance on positive and negative class, the geometric mean was also calculated. BNN has greater geometric mean than RF, therefore the BNN can better detect the viral pre-miRNAs and, at the same time, can better recognize negative data such as coding genes, pseudo genes and miRNA-like negative hairpins [32]. Therefore, BNN was selected as the proposed model according to its performance in accuracy, sensitivity and geometric mean. The proposed BNN adopted SC score and its derivative features to identify the viral pre-miRNAs, while ViralMir does not include these features in its model. The results of the proposed BNN compared with ViralMir are shown in Table 5. With the less number of selected features, BNN shows higher accuracy, sensitivity, specificity, geometric mean values and less false positive rate compared to ViralMir.

TABLE III. PERFORMANCES OF C4.5, KNN, NB, BNN, RF AND SVM USING 5 FOLD CROSS VALIDATION

Classifiers	Performance Measurement				
	ACC	***Sn***	***Sp***	***FPR***	***ROC area***
C4.5	0.989	0.681	0.996	0.004	0.872
KNN	0.985	0.676	0.992	0.008	0.834
NB	0.955	0.91	0.957	0.043	0.975
BNN	0.989	**0.71**	0.996	0.004	**0.98**
RF	**0.99**	**0.71**	**0.997**	**0.003**	**0.98**
SVM	0.988	0.571	0.999	0.001	0.785

TABLE IV. PERFORMANCES OF RF AND BNN ON THE TEST SET

Classifiers	Performance Measurement				
	ACC	***Sn***	***Sp***	***FPR***	***GM***
RF	0.973	0.792	**1**	**0**	0.889
BNN	**0.978**	**0.830**	0.99	0.01	**0.906**

TABLE V. PERFORMANCE FOR DISCRIMINATING VIRAL MIRNA PRECURSOR FROM NON-VIRAL MIRNA PRECURSOR, NON MIRNA PRECURSOR SEQUENCES AND CODING SEQUENCE OF VIRUS

Classifiers	Performance Measurement				
	ACC	***Sn***	***Sp***	***FPR***	***GM***
ViralMir	0.80	0.69	0.82	0.18	0.752
BNN	0.978	0.830	0.99	0.01	0.906

IV. CONCLUSION

In this research, a classification tool for viral pre-miRNAs has been developed based on the back-propagation neural network algorithm. The feature subset was selected specifically for virus-encoded pre-miRNAs in order to enhance the performance of the machine learning model. The results revealed that SC and derivative of SC scores features in associated with secondary structure features are informative features that can reflect the unique intrinsic genetic robustness characteristic of viral pre-miRNAs. As a consequence, the model achieved up to 0.96 accuracy and as low as 0.01 false positive rate. Moreover, this machine learning model could be a useful tool for researchers who study in the field of non-coding RNAs for discovering novel viral pre-miRNAs based on robustness features that simulate the intrinsic characteristics of viral pre-miRNAs. The feasible future work could be a simulation of a folding kinetic of viral-encoded pre-miRNAs to reflect the underlying biological properties and utilize as a feature to distinguish biological function related information in viral pre-miRNAs subclasses.

ACKNOWLEDGMENT

The authors would like to extent their deep gratitude to National Center for Genetic Engineering and Biotechnology (BIOTEC) for research funding.

REFERENCES

[1] J. S. Mattick, and I. V. Makunin. "Non-coding RNA," Hum. Mol. Genet. Vol. 15, Apr. 2006, pp. R17-R29, doi: 10.1093/hmg/ddl046.

[2] P. Brodersen, L. Sakvarelidze-Achard, M. Bruun Rasmussen, F. Dunoyer, Y.Y. Yamamoto, L. Sieburth, and O. Voinnet. "Widespread translational inhibition by plant miRNAs and siRNAs," Science. Vol. 320, May. 2008, pp. 1185-1190, doi: 10.1126/science.1159151.

[3] S. Gu, L. Jin, F. Zhang, P. Sarnow, and M.A. Kay. "Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs." Nat. Struct. Mol. Biol. Vol. 16, Feb. 2009, pp. 144-150, doi: 10.1038/nsmb.1552.

[4] S. Pfeffer, M. Zavolan, F.A. Grasser, M. Chien, J.J. Russo, J. Ju, B. John, A.J. Enright, D. Marks, C. Sander, and T. Tuschl. "Identification of virus-encoded microRNAs." Science. Vol. 304, Apr. 2004, pp. 734-736, *doi:* 10.1126/science.1096781

[5] Pfeffer, S., and O. Voinnet. "Viruses, microRNAs and cancer." Oncogene Vol. 25, 2006 , pp. 6211-6219.

[6] Sullivan, Christopher S. "New roles for large and small viral RNAs in evading host defences." Nature Reviews Genetics Vol. 9, Jul. 2008, pp. 503-507.

[7] J. W. Carl, J. Trgovcich, and S. Hannenhalli. "Widespread evidence of viral miRNAs targeting host pathways." BMC bioinformatics. Vol. 14, Jan. 2013, pp. S3, doi: 10.1186/1471-2105-14-S2-S3.

[8] D. N. Kim, H. Chae, S.T. Oh, J. Kang, C.H. Park, W. S. Park, K. Takada, J. M. Lee, W, Lee, and S. K. Lee. "Expression of viral microRNAs in Epstein-Barr virus-associated gastric carcinoma." J. Virol. Vol. 81. Jan. 2007, pp. 1033-1036, doi: 10.1128/JVI.02271-06.

[9] Kincaid, Rodney P., and Christopher S. Sullivan. "Virus-encoded microRNAs: an overview and a look to the future." PLoS Pathog . Vol 8, Dec 2012, e1003018

[10] Zisoulis, Dimitrios G., et al. "Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans." Nature structural & molecular biology. Vol. 17,.2 2010, pp. 173-179.

[11] S. Kumar, F. A. Ansari, and V. Scaria. "Prediction of viral microRNA precursors based on human microRNA precursor sequence and structural features." Virol. J. Vol. 6, Aug. 2009, pp.129, doi: 10.1186/1743-422X-6-129.

[12] A. Grundhoff, and C. S. Sullivan. "Virus-encoded microRNAs." Virology. Vol. 411, Mar. 2011, pp. 325–343, doi: 10.1016/j.virol.2011.01.002.

[13] Lowe, Todd M., and Sean R. Eddy. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic acids research Vol. 25, May 1997, pp. 955-964.

[14] Zuker, Michael. "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic acids research Vol31, Apr 2003, pp. 3406-3415.

[15] E. Bonnet, J. Wuyts, P. Rouze, and Y. Peer. "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences." Bioinformatics. Vol. 20, Jan. 2004, pp. 2911-2917, doi: 10.1093/bioinformatics/bth374.

[16] C. Xue, F. Li, T. He, G. Liu, Y, Li, and X. Zhang. "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine." BMC bioinformatics. Vol. 6, Dec. 2005, pp. 310, doi: 10.1186/1471-2105-6-310.

[17] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z, Lu. "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features." Nucl. Acids. Res. Vol. 35, Jun. 2007, pp. W339-W344, doi: 10.1093/nar/gkm368.

[18] K. L. Ng, and S. K. Mishra. "De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures." Bioinformatics. Vol. 23, Jun. 2007, pp. 1321-1330, doi: 10.1093/bioinformatics/btm026.

[19] D. T. Chang, C. Wang, and J. Chen. "Using a kernel density estimation based classifier to predict species-specific microRNA

precursors." BMC bioinformatics. Vol. 9, Dec. 2008, pp. S2, doi: 10.1186/1471-2105-9-S12-S2.

[20] R. Batuwita, and V. Palade. "microPred: effective classification of pre-miRNAs for human miRNA gene prediction." Bioinformatics. Vol. 25, Apr. 2009, pp. 989-995, doi: 10.1093/bioinformatics/btp107.

[21] J. Ding, S. Zhou, and J. Guan. "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features." BMC bioinformatics. Vol. 11, Dec. 2010, pp. S11, doi: 10.1186/1471-2105-11-S11-S11.

[22] A. Jha, R. Chauhan, M, Mehra, H. R. Singh, and R. Shankar. "miR-BAG: bagging based identification of microRNA precursors." PLos ONE. Vol. 7, Aug. 2012, pp. e45782, doi:10.1371/journal.pone.0045782.

[23] S. Lertampaiporn, C. Thammarongtham, C, Nukoolkiy, B. Kaewkamnerdpong, and M. Ruengjitchatchawalya. "Heterogeneous ensemble approach with discriminative features and modified-SMOTE bagging for pre-miRNA classification." Nucl. Acids. Res. Vol. 41, Sep. 2013, pp. e21, doi: 10.1093/nar/gks878.

[24] K. Y. Huang, T. Y. Lee, Y. C. Teng, and T. H. Chang. "ViralmiR: a support-vector-machine-based method for predicting viral microRNA precursors." BMC bioinformatics. Vol. 16, Jan. 2015, pp. S9, doi: 10.1186/1471-2105-16-S1-S9.

[25] M. T. Lee, and J. Kim. "Self containment, a property of modular RNA structures, distinguishes microRNAs." PLoS Comput. Biol. Vol. 4, Jul. 2008, pp. e1000150, doi:10.1371/journal.pcbi.1000150.

[26] 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, Jun. 2009, pp. 10-18, doi: 10.1145/1656274.1656278.

[27] Quinlan, Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann (1993).

[28] A. Kozomara, and S. Griffiths-Jones. "miRBase: integrating microRNA annotation and deep-sequencing data." Nucl. Acids. Res. Jan. 2010, pp. D152-7, doi: 10.1093/nar/gkq1027.

[29] L. Kong, Y. Zhang, Z. Q. Liu, S. Q. Zhao, L. Wei, and G. Gao. "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucl. Acids. Res. Vol. 35, Jul. 2007, pp. W345-W349, doi: 10.1093/nar/gkm391.

[30] S. Lertampaiporn, C. Thammarongtham, C, Nukoolkiy, B. Kaewkamnerdpong, and M. Ruengjitchatchawalya. "Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm." Nucl. Acids. Res. Vol. 42, Apr. 2014, pp. e93, doi: 10.1093/nar/gku325.

[31] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. “ViennaRNA Package 2.0.” Algorithms. Mol. Biol. Vol. 6, Nov. 2011, pp. 1748-7188, doi: 10.1186/1748-7188-6-26.

[32] Batuwita, Rukshan, and Vasile Palade. "Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning." Journal of bioinformatics and computational biology. Vol 10, Apr. 2012 : 1250003.