

Received February 26, 2020, accepted March 11, 2020, date of publication March 16, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980897

# An Improved Method for Identification of Pre-miRNA in *Drosophila*

TIEYING YU<sup>1,2,3</sup>, MIN CHEN<sup>1,2</sup>, AND CHUNDE WANG<sup>1,2,4</sup>

<sup>1</sup>Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China

<sup>2</sup>Center for Ocean Mega-Science, Chinese Academy of Sciences, Yantai 264003, China

<sup>3</sup>College of Life Sciences, Shandong Agricultural University, Tai'an 271018, China

<sup>4</sup>Marine Science and Engineering College, Qingdao Agricultural University, Qingdao 266109, China

Corresponding author: Chunde Wang (chundewang2007@163.com)


The work of Chunde Wang was supported by the Natural Science Foundation of China under Grant 31572618 and Grant 31972791.

**ABSTRACT** Identification of microRNAs is important in studies of regulation of gene expression in many biological processes. In this study, we developed an improved method for identification of microRNAs in *Drosophila*. We used the iLearn, PyFeat, and Pse-in-One methods to extract the features and then used Max-Relevance-Max-Distance (MRMD2.0) and t-Distributed Stochastic Neighbour Embedding (t-SNE) to reduce dimension of the features and the random forest classifier in Weka to identify miRNAs. With this method, we found that the discriminative features for identification of pre-miRNAs were, in *Drosophila melanogaster*, the occurrences of G\_GUG and C\_AGU when the value of the feature vector was greater than 2, and in *Drosophila pseudoobscura*, the 4-tuple nucleotide composition and the occurrence of 4-length neighbouring nucleic acids when the value of the feature vector was less than 0.02. These vectors covered all compositional information or the frequency of bases. Classification results showed the classification accuracy was 95.7% and 93.6%, the precision rate was 95.8% and 93.6%, and the recall rate was 95.7% and 93.6% in *Drosophila melanogaster* and *Drosophila pseudoobscura*, respectively, which are higher than those reported in previous studies.

**INDEX TERMS** microRNA, iLearn, PyFeat, Pse-in-One, MRMD2.0, t-SNE, accuracy, random forest, discriminative features.

## I. INTRODUCTION

MicroRNAs (miRNAs) are endogenous small RNAs of approximately 20-24 nucleotides in length [1]–[3]. The expression of a gene can be fine-tuned via a combination of several miRNAs [4]–[6]. MiRNAs are thought to regulate one-third of human genes [7]. More than 50% of human miRNAs are situated in cancer-related gene fragment areas, including breast cancer, nerve-cell carcinomas and so on [6], [8]–[10]. Moreover, miRNAs are closely related to many common diseases [11]–[23]. Currently, gene silencing and RNA interference technologies play vital roles in the development of anticancer medicines and crop improvements [24], [25]. Therefore, given the important biological function of miRNAs, the accurate detection and prediction of miRNA sequences is a significant issue, and may provide new solutions for many biological problems [26]–[28].

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou .

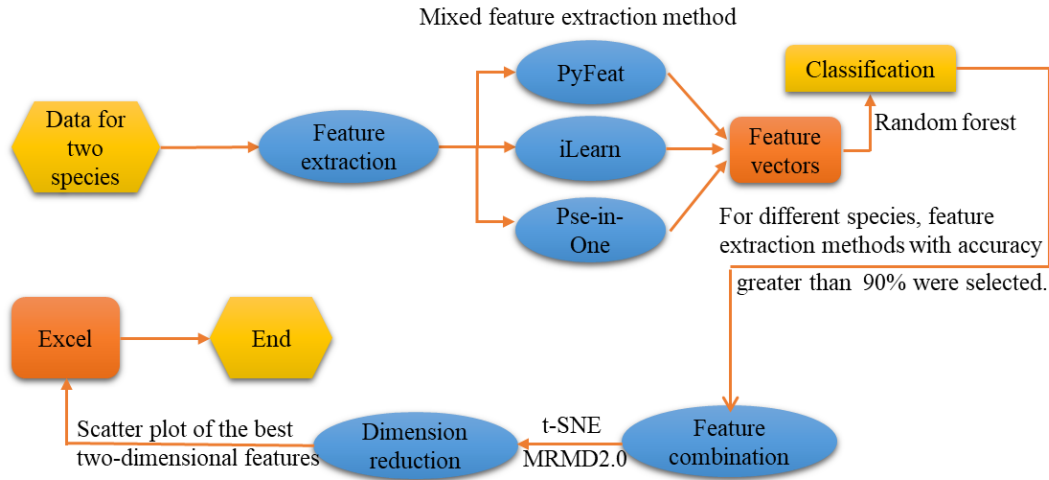
MicroRNAs can be routinely identified by sequencing methods, such as RNA-Seq [29], [30], but these methods have many disadvantages, including high costs [31], [32], sensitivity to environmental influences, and difficulty in distinguishing pre-miRNA from other RNAs [33]. In addition, RNA-Seq-based approaches have been shown to be biased against miRNAs with higher copy number or expression levels, and may exclude transient or cellular or developmental stage-specific miRNAs [34].

Computational prediction based on discriminative features provides new means for identification of microRNAs. Good discriminative features should represent the major features of the whole sequence and can be found by feature extraction and feature selection. These operations seek to find the most effective features for classification and recognition of miRNA from a large number of features to decrease the feature dimension, that is, to obtain a “fewer but better” group of classification characteristics with a low probability of classification error. A good method should be able to

**TABLE 1.** Summary of the miRNA datasets in two *Drosophila* species.

Datasets	P	N
<i>Drosophila melanogaster</i>	238	443
<i>Drosophila pseudoobscura</i>	691	2094

Note: Note that P represents positive samples (true miRNAs) and N represents negative samples (sequences that are not miRNAs).



**FIGURE 1.** Main flow chart for the identification of pre-miRNAs.

find the discriminative features in a shorter time and yet produce a better classification effect. Researchers have used many methods to extract the features for identification of miRNAs. For example, Xue *et al.* suggested 32D novel triplet features, but the method requires a long time to run [35]. Some methods may generate multidimensional features, such as those proposed Wei *et al.* [36]. The method developed by Jiang *et al.* (2016) increased the classification accuracy by extracting 98 dimensional features, but it took a long time to run, and none of the most discriminative features were able to identify whether a sequence was a miRNA [37].

To obtain a good classification effect, in this study we developed a new method which combined several feature extraction methods including iLearn [38], PyFeat [39], and Pse-in-One [40] and used the random forest [41]–[50] as the classifier. With this improved method, we were able to identify the key features for identification of pre-miRNAs in two *Drosophila* species. Cross-validation tests [51]–[58] and classification results showed that the classification accuracy, precision rate and recall rate obtained with this method were increased compared with those of previous reports.

## II. MATERIALS AND METHODS

### A. DATASETS

The miRNA sequence data of the two *Drasophila* species, *D. melanogaster* and *D. pseudoobscura*, used in this study were taken from the published article “A framework for improving microRNA prediction in non-human genomes” [34]. The details of the datasets are summarized in Table 1.

As seen in Table 1, datasets for each species included both positive data, which are true miRNAs, and negative data, which are not miRNAs.

### B. THE SCHEME OF THE ANALYSES

As shown in Fig. 1, in this study, firstly, we converted the miRNA sequence data into feature vectors using the mixed feature extraction method, and then used the random forest classifier for classification. With the n-fold cross-validation of the model built in this study, accuracy, precision, and recall rates were tested. Thirdly, we combined the features with classification accuracy greater than 90% obtained by iLearn, PyFeat and Pse-in-One, and then reduced the dimensions with Max-Relevance-Max-Distance (MRMD2.0) and t-Distributed Stochastic Neighbour Embedding (t-SNE). Finally, we identified the lowest number of dimensional features and the most discriminative features that could distinguish miRNAs from non-miRNAs and using Excel to draw scatter plots to find the most discriminative features.

### C. FEATURE EXTRACTION

iLearn, PyFeat, and Pse-in-One were used to extract features. The pseudoKNC, z-Curve [59], gcContent, cumulativeSkew, atgcRatio, monoMonoKGap, monoDiKGap, monoTriKGap, diMonoKGap, diDiKGap, diTriKGap, triMonoKGap and triDiKGap algorithms from PyFeat produced 2971 dimensional features and automatically optimized the selection of more discriminative features. Additionally, the Pseudo k-tupler Composition (PseKNC) [60]–[65] and the other 12 algorithms from iLearn and the Mismatch, Kmer and

PC-PseDNC-General algorithms from Pse-in-One were used to convert the miRNA sequences into feature vectors. The FyFeat feature encoding algorithms are described in details in the Supplementary Material.

#### D. CLASSIFIER

In this study, Waikato Environment for Knowledge Analysis (Weka) was used as a classifier. For the 'Classify' panel choice, we used the random forest [41], [66] which has been successfully employed in many studies [67]–[72] for identification and classification. Moreover, in Weka, the n-fold ( $n \in 5, 9, 10, 11, 12$ ) cross-validation test was adopted to perform the predictions.

#### E. FEATURE COMBINATION

After feature extraction, the features with classification accuracies higher than 90 % were chosen and combined. For *Drosophila melanogaster*, we combined the results of eight feature extraction methods. For *Drosophila pseudoobscura*, we combined the results of twelve feature extraction methods.

#### F. DIMENSION REDUCTION

Max-Relevance-Max-Distance (MRMD2.0), proposed by Qu *et al.* [73], was used to reduce dimension. The features were ranked according to their classification scores so that the more discriminative features will have higher ranks. We then determined the lowest number of dimensional features that could distinguish between positive and negative examples and the most discriminative features to distinguish whether a sequence was a miRNA.

In addition to MRMD2.0, after feature combination, we further used t-Distributed Stochastic Neighbour Embedding (t-SNE) [74] to reduce dimension and visualized the combined features in a 2D feature space, and then the positive and negative samples were easily separated by two distinct background areas of SVM [75]. The points of the positive cases could be allocated in the red area, and the points of the negative cases could be allocated in the blue area. So, the positive and negative cases could be separated easily.

After dimension reduction with both MRMD2.0 and t-SNE, we found the best two-feature pairs that could discriminate positive cases and negative cases for the two species of *Drosophila*. They were the first two features in the dimension reduction accuracy list of each species. Then, we used Excel to draw a scatter plot of the best two-dimensional features to determine whether the two features were discriminative.

#### G. MEASUREMENT

To evaluate the performance of the newly developed method, we calculated the classification accuracy (ACC), precision and recall and compared them with those from previous studies. These indices have been widely used in several bioinformatics studies [14], [20], [38], [76]–[88].

To calculate these indices, the data were divided into 4 categories including true positive data (TP), true negative

data (TN), false positive data (FP) and false negative data (FN).

Accuracy(ACC) rate is the proportion of the dataset that is correctly classified, and can be calculated by the below formula [89].

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

Precision rate is the proportion of the positive dataset that is correctly classified and can be calculated following the formula [89].

$$precision = \frac{TP}{TP + FP} \times 100\%$$

Recall rate is the proportion of the positive examples that are correctly classified and can be calculated by the formula [89].

$$recall = \frac{TP}{TP + FN} \times 100\%$$

### III. RESULTS

#### A. FEATURE EXTRACTION

Tables 2a-2b compare the classification results for two *Drosophila* species with the previous study [37]. The best feature extraction method is a combination of the first 13 methods from PyFeat. For *Drosophila melanogaster*, a total of 312 dimensional features were extracted. The accuracy rate, precision rate and recall rate were 95.7%, 95.8% and 95.7%, respectively, which were all higher than those reported in the previous study [37]. For *Drosophila pseudoobscura*, a total of 413 dimensional features were extracted. The accuracy rate, precision rate and recall rate were 93.6%, 93.6% and 93.6%, respectively, which were also higher than those the previous study [37]. Thus our feature extraction methods are able to extract features for discrimination of positive and negative cases and the accuracy, precision and recall are all improved compared with previous methods [37].

#### B. DIMENSION REDUCTION

Figures 2a-2b show the dimension reduction results for the two species of *Drosophila*. The details of the results are described in Supplementary Material Tables S1-S2.

In *Drosophila melanogaster*, after feature extraction with iLearn, PyFeat, and Pse-in-one, a total of 1761 dimensional features with a classification accuracy of more than 90% were combined. After using MRMD2.0 to reduce the dimension, the lowest number of dimensional features that could distinguish positive and negative cases was found to be a total of 12 dimensions to achieve the same accuracy as that of the previous study [37] (Fig. 2a).

In *Drosophila pseudoobscura*, after feature extraction with iLearn, PyFeat, and Pse-in-one, a total of 1784 dimensional features with a classification accuracy of more than 90% were combined. After using MRMD2.0 to reduce the

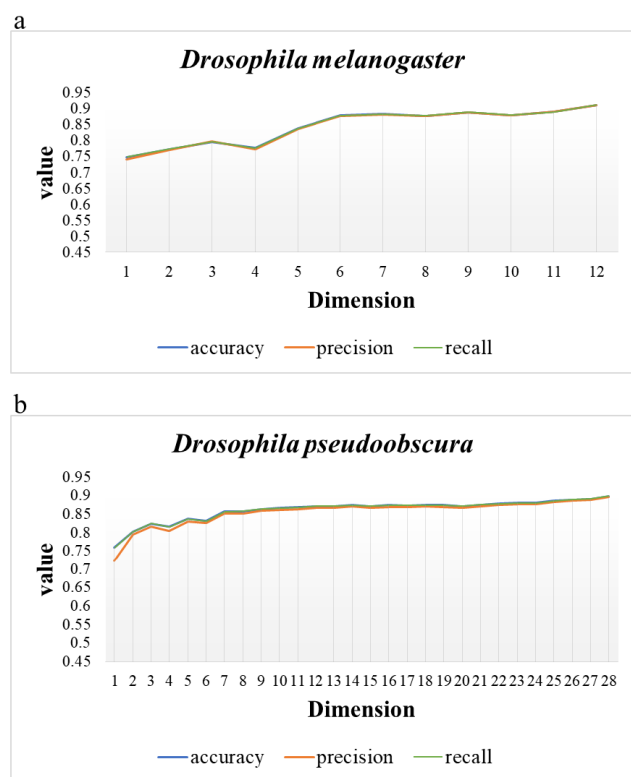
**TABLE 2.** Comparison of the classification results with the previous study.

a <i>Drosophila melanogaster</i>		Parameters of performance		
Methods		Accuracy	Precision	Recall
This study	PyFeat	95.7%	95.8%	95.7%
	iLearn	83.7%-91.9%	84.0%-92.1%	83.7%-91.9%
	Pse-in-One	88.8%-92.2%	88.8%-92.2%	88.8%-92.2%
Jiang et al. (2016)		91.0%	91.0%	91.0%

b <i>Drosophila pseudoobscura</i>		Parameters of performance		
Methods		Accuracy	Precision	Recall
This study	PyFeat	93.6%	93.6%	93.6%
	iLearn	85.3%-92.0%	86.0%-92.1%	85.3%-92.0%
	Pse-in-One	90.7%-92.3%	90.6%-92.2%	90.7%-92.3%
Jiang et al. (2016)		92.0%	92.0%	92.0%

Note: Compared with the previous study, it can be seen that the feature extraction methods we adopted are better than the previous study, and the optimal feature extraction method in this study is PyFeat.

**FIGURE 2.** Lowest number of dimensional features that distinguish positive and negative cases well for the different species.

dimension, the classification accuracy was approximately 90%, which was slightly lower than that of the previous study [37]. The lowest number of dimensional features that could distinguish positive and negative cases was found to be a total of 28 dimensions (Fig. 2b).

We further used t-SNE to visualize them in a 2D feature space and the positive and negative cases were easily separated by two distinct background areas of SVM, as shown in Figures 3a-3b. As can be seen in the figures, the points

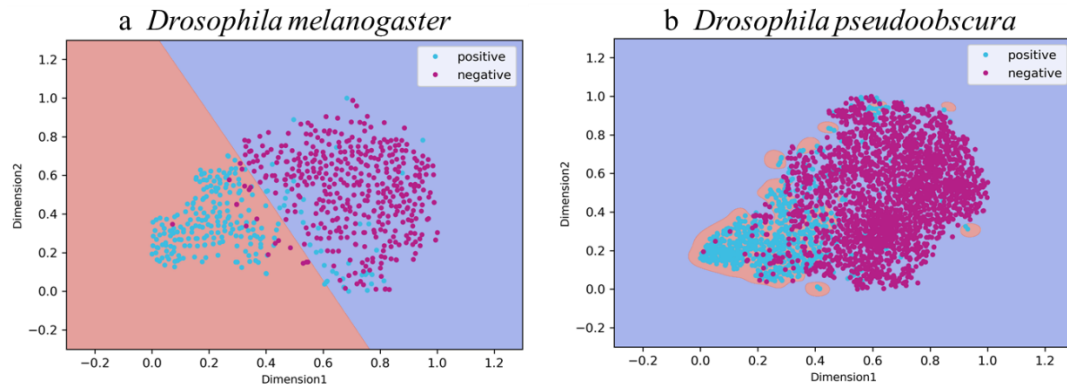
of the positive cases were allocated in the red area, and the points of the negative cases were allocated in the blue area. In this method, the positive cases and negative cases were clearly separated, indicating that the two features were very discriminative.

### C. DETERMINATION OF THE MOST DISCRIMINATIVE FEATURES

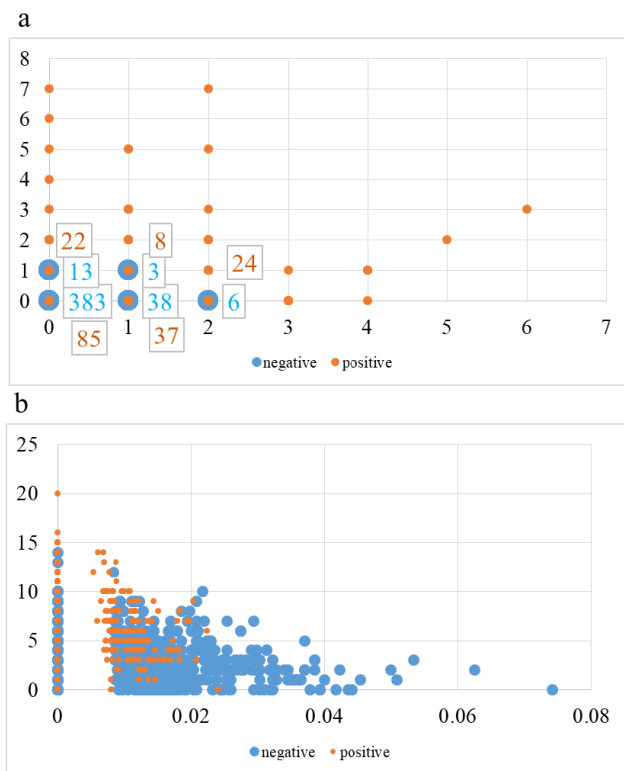
The MRMD2.0 dimensionality reduction produced a list ranked by classification accuracy. In the list, the features with higher ranks were more discriminative. For *Drosophila melanogaster*, the most discriminative features for miRNA recognition were the numbers of G\_GUG and C\_AGU in the miRNA whole sequence. The feature extraction method integrates the 13 small methods from PyFeat, corresponding to monoTriKGap, and the accuracy of the first two dimensional features was 77.2%. The scatter plot using the two-dimensional features is shown in Fig. 4a. The algorithm for the feature extraction method is detailed in the Supplementary Material.

In Fig. 4a, the negative cases are overwhelmed by the positive cases, so we magnified the negative cases. The data label shows the number of feature vectors for the positive and negative cases that coincided. There are 443 negative feature vectors and 238 positive feature vectors. Except for the overlapped points, the remaining points are the feature vectors of the positive cases. As can be seen, when the value of the feature vector is greater than 2, it is likely a pre-miRNA sequence.

For *Drosophila pseudoobscura*, the most discriminative features for miRNA recognition were the 4-tuple nucleotide composition and the occurrences of a 4-length neighbouring nucleic acids that differ by at most  $m$  mismatches ( $m < 4$ ). The feature extraction methods are PseKNC from iLearn and Mismatch from Pse-in-One and the accuracy of the first two dimensional features was 80.2%. The scatter plot using the



**FIGURE 3.** Dimensionless scatter plots using t-SNE. Note: Note that the positive cases are represented as green points, and the negative cases are represented as red points.



**FIGURE 4.** Scatter plots of the best two-dimensional features to distinguish miRNA and non-miRNA for the two kinds of *Drosophila*. Note: Note that the orange points represent positive cases, and the blue points represent negative cases.

two-dimensional features is shown in Fig. 4b. The algorithms for the two feature extraction methods are detailed in the Supplementary Material.

In Fig. 4b, when we magnify the points of the negative cases and shrink the points of the positive cases, we can see that when the value of the feature vector is less than 0.02, the subject sequence is likely a pre-miRNA sequence. Although there are some overlapping points, the positive cases and negative cases are easier to separate, indicating that these two features are discriminative for distinguishing of miRNAs from non-miRNAs.

#### IV. DISCUSSION

Previous studies have shown that miRNAs are much more abundant and functional than previously thought and the identification and prediction of miRNA has important biological significance in many bioinformatics studies [90]–[99]. There are many ways to identify miRNAs, but many of these methods have disadvantages. Our goal in this study was to find the lowest number of dimensional features and the most discriminative features that could distinguish miRNAs from non-miRNAs. Using machine learning and computational prediction to identify miRNAs can reduce the cost and achieve fast, accurate and good results. In this study, we developed an improved method which uses PyFeat, iLearn and Pse-in-one for feature extraction and the random forest as the classifier. N-fold cross-validation test was used to train the classification model together with t-SNE and MRMD2.0 for dimension reduction. Compared with MRMD2.0, t-SNE is more comprehensive. It is better for determining whether the positive cases and negative cases are split and makes it easier to determine whether the extracted features were discriminative or not. However, t-SNE cannot produce a profile with a feature list like MRMD2.0. Therefore, the two methods could be used together to achieve a better effect. Compared with the previous study [37], we improved the classification accuracy, precision and recall and we were able to find the lowest number of dimensional features and the most discriminative features that could distinguish miRNAs from non-miRNAs. With this method, it is easy to distinguish the positive cases and negative cases and determine whether the extracted features were discriminative for identification of miRNAs. We were also able to identify the discriminative features for identification of pre-miRNAs in two *Drosophila* species. In addition, our feature selection method can be applied to identify lncRNAs [100] and predict their target genes [101] and functions [102], [103].

#### ACKNOWLEDGMENT

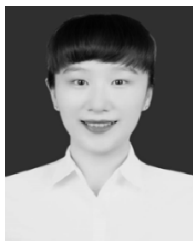
The authors would like to thank Dr. Q. Zou for his insightful advices in the study and manuscript.

## REFERENCES

- [1] M. Tsuzuki and Y. Watanabe, "Profiling new small RNA sequences," *Methods Mol. Biol.*, vol. 1456, pp. 177–188, Oct. 2017.
- [2] A. E. Murmann, E. T. Bartom, M. J. Schipma, J. Vilker, S. Chen, and M. E. Peter, "6mer seed toxicity in viral microRNAs," *iScience*, vol. 23, Dec. 2020, Art. no. 100737.
- [3] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, and K.-C. Chou, "Identification of real microRNA precursors with a pseudo structure status composition approach," *PLoS ONE*, vol. 10, no. 3, 2015, Art. no. e0121501.
- [4] A. Laganà, "Computational prediction of microRNA targets," *Adv. Exp. Med. Biol.*, vol. 887, pp. 231–252, 2015.
- [5] Y. Xu, M. Guo, X. Liu, C. Wang, and Y. Liu, "Inferring the soybean (*Glycine max*) microRNA functional network based on target gene network," *Bioinformatics*, vol. 30, no. 1, pp. 94–103, Jan. 2014.
- [6] K. K. W. To, W. Fong, C. W. S. Tong, M. Wu, W. Yan, and W. C. S. Cho, "Advances in the discovery of microRNA-based anticancer therapeutics: Latest tools and developments," *Expert Opinion Drug Discovery*, vol. 15, no. 1, pp. 63–83, Jan. 2020.
- [7] R. T. Marquez and A. P. McCaffrey, "Advances in microRNAs: Implications for gene therapists," *Hum. Gene Therapy*, vol. 19, no. 1, pp. 27–38, 2008.
- [8] A. Kopkova, J. Sana, T. Machackova, M. Vecera, L. Radova, K. Trachtova, V. Vybihal, M. Smrcka, T. Kazda, O. Slaby, and P. Fadrus, "Cerebrospinal fluid microRNA signatures as diagnostic biomarkers in brain tumors," *Cancers*, vol. 11, no. 10, p. 1546, 2019.
- [9] Y. Zhang, C. Kou, S. Wang, and Y. Zhang, "Genome-wide differential-based analysis of the relationship between DNA methylation and gene expression in cancer," *Current Bioinf.*, vol. 14, no. 8, pp. 783–792, Dec. 2019.
- [10] L. H. Zhijin, W. Zhao, Y. Xu, Z. Zhang, M. Wang, and X. Zhou, "Integrative analysis of DNA methylation and gene expression profiles identifies MIR4435-2HG as an oncogenic lncRNA for glioma progression," *Gene*, vol. 715, Oct. 2019, Art. no. 144012, doi: 10.1016/j.gene.2019.144012.
- [11] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "MiR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, pp. D98–D104, Jan. 2009.
- [12] L. Yu, J. Zhao, and L. Gao, "Predicting potential drugs for breast cancer based on miRNA and tissue specificity," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 971–982, 2018.
- [13] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 283–291, Jan./Feb. 2019.
- [14] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods," *Briefings Bioinform.*, Oct. 2019, doi: 10.1093/bib/bbz080.
- [15] X. Zeng, W. Lin, M. Guo, and Q. Zou, "A comprehensive overview and evaluation of circular RNA detection tools," *PLOS Comput. Biol.*, vol. 13, no. 6, 2017, Art. no. e1005420.
- [16] C. Jeyaram, M. Philip, R. C. Perumal, J. Benny, J. M. Jayakumari, and M. S. Ramasamy, "A computational approach to identify novel potential precursor miRNAs and their targets from hepatocellular carcinoma cells," *Current Bioinf.*, vol. 14, no. 1, pp. 24–32, Dec. 2018.
- [17] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu, and Y. Wang, "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Syst. Biol.*, vol. 4, no. S1, May 2010.
- [18] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *Int. J. Data Mining Bioinf.*, vol. 8, no. 3, pp. 282–293, 2013.
- [19] Q. Jiang, S. Jin, Y. Jiang, M. Liao, R. Feng, L. Zhang, G. Liu, and J. Hao, "Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells," *Mol. Neurobiol.*, vol. 54, no. 1, pp. 594–600, Jan. 2017.
- [20] G. Liu, Y. Zhao, S. Jin, Y. Hu, T. Wang, R. Tian, Z. Han, D. Xu, and Q. Jiang, "Circulating vitamin E levels and Alzheimer's disease: A mendelian randomization study," *Neurobiol. Aging*, vol. 72, pp. 189.e1–189.e9, Dec. 2018.
- [21] G. Liu, Y. Zhang, L. Wang, J. Xu, X. Chen, Y. Bao, Y. Hu, S. Jin, R. Tian, W. Bai, W. Zhou, T. Wang, Z. Han, J. Zong, and Q. Jiang, "Alzheimer's disease rs11767557 variant regulates EPHA1 gene expression specifically in human whole blood," *J. Alzheimer's Disease*, vol. 61, no. 3, pp. 1077–1088, Jan. 2018.
- [22] G. Wang, Y. Wang, W. Feng, X. Wang, J. Y. Yang, Y. Zhao, Y. Wang, and Y. Liu, "Transcription factor and microRNA regulation in androgen-dependent and-independent prostate cancer cells," *BMC Genomics*, vol. 9, no. 2, p. S22, 2008.
- [23] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon  $\gamma$ -stimulated HeLa cells," *PLoS ONE*, vol. 5, no. 7, 2010, Art. no. e11794.
- [24] A. Charrier, E. Vergne, C. Joffrion, A. Richer, N. Dousset, and E. Chevreau, "An artificial miRNA as a new tool to silence and explore gene functions in apple," *Transgenic Res.*, vol. 28, nos. 5–6, pp. 611–626, Dec. 2019.
- [25] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artif. Intell. Med.*, vol. 77, pp. 53–63, Mar. 2017.
- [26] M. Faiza, K. Tanveer, S. Fathi, Y. Wang, and K. Raza, "Comprehensive overview and assessment of microRNA target prediction tools in homo sapiens and drosophila melanogaster," *Current Bioinf.*, vol. 14, no. 5, pp. 432–445, Jun. 2019.
- [27] G. Liu, S. Jin, Y. Hu, and Q. Jiang, "Disease status affects the association between rs4813620 and the expression of Alzheimer's disease susceptibility gene TRIB3," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 45, pp. E10519–E10520, Nov. 2018.
- [28] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification in arabidopsis using multiple histone markers," *BioMed Res. Int.*, vol. 2015, Sep. 2015, Art. no. 61402.
- [29] A. D. Jayaprakash, O. Jabado, B. D. Brown, and R. Sachidanandam, "Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing," *Nucleic Acids Res.*, vol. 39, no. 21, p. e141, Nov. 2011.
- [30] X. Zhu, H.-D. Li, L. Guo, F.-X. Wu, and J. Wang, "Analysis of single-cell RNA-seq data by clustering approaches," *Current Bioinf.*, vol. 14, no. 4, pp. 314–322, Apr. 2019.
- [31] A. Sboner, X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein, "The real cost of sequencing: Higher than you think!" *Genome Biol.*, vol. 12, no. 8, p. 125, 2011.
- [32] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "DeepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, Dec. 2019, doi: 10.1093/bioinformatics/btz418.
- [33] D. Duan, K.-X. Zheng, Y. Shen, R. Cao, L. Jiang, Z. Lu, X. Yan, and J. Li, "Label-free high-throughput microRNA expression profiling from total RNA," *Nucleic Acids Res.*, vol. 39, no. 22, p. e154, Dec. 2011.
- [34] R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green, "A framework for improving microRNA prediction in non-human genomes," *Nucleic Acids Res.*, vol. 43, no. 20, Jul. 2015, Art. no. gkv698.
- [35] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinf.*, vol. 6, no. 1, p. 310, 2005.
- [36] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 192–201, Jan. 2014.
- [37] L. Jiang, J. Zhang, P. Xuan, and Q. Zou, "BP neural network could help improve pre-miRNA identification in various species," *BioMed Res. Int.*, vol. 2016, Aug. 2016, Art. no. 9565689.
- [38] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Brief Bioinform.*, pp. 1–11, Apr. 2019.
- [39] R. Muhammod, S. Ahmed, D. M. Farid, S. Shatabda, A. Sharma, and A. Dehngangi, "PyFeat: A Python-based effective feature generation tool for DNA, RNA and protein sequences," *Bioinformatics*, vol. 35, no. 19, pp. 3831–3833, Oct. 2019.
- [40] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.
- [41] Y. Yao, X. Li, B. Liao, L. Huang, P. He, F. Wang, J. Yang, H. Sun, Y. Zhao, and J. Yang, "Predicting influenza antigenicity from hemagglutinin sequence data based on a joint random forest method," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 1545.

- [42] L. Wei, J. Tang, and Q. Zou, "SkipCPP-pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC Genomics*, vol. 18, no. S7, p. 742, Oct. 2017.
- [43] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinform.*, vol. 17, no. 1, p. 398, Dec. 2016.
- [44] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 14, no. 3, pp. 687–695, May 2017.
- [45] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.
- [46] B. Liu, S. Chen, K. Yan, and F. Weng, "IRO-PsekGCC: Identify DNA replication origins based on pseudo k-Tuple GC composition," *Frontiers Genet.*, vol. 10, p. 842, Sep. 2019.
- [47] Y. Chu, A. C. Kaushik, X. Wang, W. Wang, Y. Zhang, X. Shan, B. R. Salahub, Y. Xiong, and D. Q. Wei, "DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Brief Bioinform.*, pp. 1–12, Dec. 2019.
- [48] G. Liu, Y. Hu, Z. Han, S. Jin, and Q. Jiang, "Genetic variant rs17185536 regulates SIM1 gene expression in human brain hypothalamus," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 9, pp. 3347–3348, Feb. 2019.
- [49] G. Wang, X. Luo, J. Wang, J. Wan, S. Xia, H. Zhu, J. Qian, and Y. Wang, "MeDReaders: A database for transcription factors that bind to methylated DNA," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D146–D151, Jan. 2018.
- [50] I. M. Scott, W. Lin, M. Liakata, J. E. Wood, C. P. Vermeer, D. Allaway, J. L. Ward, J. Draper, M. H. Beale, D. I. Corol, J. M. Baker, and R. D. King, "Merits of random forests emerge in evaluation of chemometric classifiers by external validation," *Analytica Chim. Acta*, vol. 801, pp. 22–33, Nov. 2013.
- [51] L. Lebanov, L. Tedone, A. Ghiasvand, and B. Paull, "Random forests machine learning applied to gas chromatography—Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils," *Talanta*, vol. 208, Feb. 2020, Art. no. 120471.
- [52] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers Genet.*, vol. 9, p. 745, Jan. 2019.
- [53] B. Liu, X. Gao, and H. Zhang, "BioSeq-analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Res.*, vol. 47, no. 20, p. e127, 2019.
- [54] B. Liu, C. C. Li, and K. Yan, "DeepSVM-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings Bioinform.* Oct. 2019, doi: 10.1093/bib/bbz098.
- [55] T. D. Capellini, G. Vaccari, E. Ferretti, S. Fantini, M. He, M. Pellegrini, L. Quintana, G. Di Giacomo, J. Sharpe, L. Selleri, and V. Zappavigna, "Scapula development is governed by genetic interactions of Pbx1 with its family members and with Emx2 via their cooperative control of Alx1," *Development*, vol. 137, no. 15, pp. 2559–2569, Aug. 2010.
- [56] T. Fang, Z. Zhang, R. Sun, L. Zhu, J. He, B. Huang, Y. Xiong, and X. Zhu, "RNAm5CPred: Prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 739–747, Dec. 2019.
- [57] X. Shan, X. Wang, C.-D. Li, Y. Chu, Y. Zhang, Y. Xiong, and D.-Q. Wei, "Prediction of CYP450 enzyme–substrate selectivity based on the network-based label space division method," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4577–4586, Nov. 2019.
- [58] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, May 2019.
- [59] H. Lin, Z.-Y. Liang, H. Tang, and W. Chen, "Identifying sigma70 promoters with novel pseudo nucleotide composition," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 4, pp. 1316–1321, Jul. 2019.
- [60] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *saccharomyces cerevisiae*," *Briefings Funct. Genomics*, vol. 18, no. 6, pp. 367–376, Oct. 2019.
- [61] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," *BMC Bioinform.*, vol. 19, no. 1, p. 306, Dec. 2018.
- [62] C. Jia, Y. Zuo, and Q. Zou, "O-GlcNAcPred-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique," *Bioinformatics*, vol. 34, no. 12, pp. 2029–2036, Jun. 2018.
- [63] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwok, K.-C. Chou, J. Song, and C. Jia, "MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, Sep. 2019.
- [64] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, and H. Lin, "iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, May 2019.
- [65] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in *saccharomyces cerevisiae* using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, Jun. 2019.
- [66] H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings Bioinform.*, Jun. 2019.
- [67] L. Wei, P. Xing, G. Shi, Z. Ji, and Q. Zou, "Fast prediction of protein methylation sites using a sequence-based feature selection technique," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 4, pp. 1264–1273, Jul./Aug. 2019.
- [68] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *J. Proteome Res.*, vol. 18, no. 7, pp. 2931–2939, Jul. 2019.
- [69] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Trans. Nanobiosci.*, vol. 16, no. 4, pp. 240–247, Jun. 2017.
- [70] P. Feng, H. Ding, H. Lin, and W. Chen, "AOD: The antioxidant protein database," *Sci. Rep.*, vol. 7, no. 1, pp. 1–4, Dec. 2017.
- [71] B. Liu, "BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinform.*, vol. 20, no. 4, pp. 1280–1294, Jul. 2019.
- [72] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "K-Skip-n-Gram-RF: A random forest based method for Alzheimer's disease protein identification," *Frontiers Genet.*, vol. 10, Feb. 2019.
- [73] K. Qu, F. Gao, F. Guo, and Q. Zou, "Taxonomy dimension reduction for colorectal cancer prediction," *Comput. Biol. Chem.*, vol. 83, Dec. 2019, Art. no. 107160.
- [74] W. Li, J. E. Cerise, Y. Yang, and H. Han, "Application of t-SNE to human genetic data," *J. Bioinform. Comput. Biol.*, vol. 15, no. 04, Aug. 2017, Art. no. 1750017.
- [75] H. Yang, W. Yang, F.-Y. Dao, H. Lv, H. Ding, W. Chen, and H. Lin, "A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*," *Briefings Bioinform.*, Oct. 2019, Art. no. bbz123.
- [76] R. C. Li, T. Garg, T. Cun, L. Shieh, G. Krishnan, D. Fang, and J. H. Chen, "Impact of problem-based charting on the utilization and accuracy of the electronic problem list," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 5, pp. 548–554, May 2018.
- [77] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.
- [78] L. Wei, S. Wan, J. Guo, and K. K. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.
- [79] B. Liu and K. Li, "iPromoter-2L2. 0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 80–87, Dec. 2019.
- [80] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers Microbiology*, vol. 9, p. 2571, Oct. 2018.
- [81] Q. Xu, Y. Xiong, H. Dai, K. M. Kumari, Q. Xu, H.-Y. Ou, and D.-Q. Wei, "PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm," *J. Theor. Biol.*, vol. 417, pp. 1–7, Mar. 2017.
- [82] X. Lu, X. Qian, X. Li, Q. Miao, and S. Peng, "DMCM: A data-adaptive mutation clustering method to identify cancer-related mutation clusters," *Bioinformatics*, vol. 35, no. 3, pp. 389–397, Feb. 2019.

- [83] X. Lu, X. Li, P. Liu, X. Qian, Q. Miao, and S. Peng, "The integrative method based on the module-network for identifying driver genes in cancer subtypes," *Molecules*, vol. 23, no. 2, p. 183, 2018.
- [84] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism," *Bioinformatics*, vol. 36, no. 4, pp. 1037–1043, Feb. 2019, doi: [10.1093/bioinformatics/btz694](https://doi.org/10.1093/bioinformatics/btz694).
- [85] Y. Wang, G. Yu, J. Wang, G. Fu, M. Guo, and C. Domeniconi, "Weighted matrix factorization on multi-relational data for LncRNA-disease association prediction," *Methods*, vol. 173, pp. 32–43, Feb. 2020.
- [86] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, p. 3140, 2018.
- [87] L. Xu, G. Liang, S. Shi, and C. Liao, "SeqSVM: A sequence-based support vector machine method for identifying antioxidant proteins," *Int. J. Mol. Sci.*, vol. 19, no. 6, p. 1773, 2018.
- [88] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, "Is there any sequence feature in the RNA pseudouridine modification prediction problem?" *Mol. Therapy-Nucleic Acids*, vol. 19, pp. 293–303, Mar. 2020.
- [89] H. Yang, H. Lv, H. Ding, W. Chen, and H. Lin, "IRNA-2OM: A sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens," *J. Comput. Biol.*, vol. 25, no. 11, pp. 1266–1277, Nov. 2018.
- [90] F. D. Mairinger, R. Werner, E. Flom, J. Schmeller, S. Borchert, M. Wessolly, J. Wohlschlaeger, T. Hager, T. Mairinger, J. Kollmeier, D. C. Christoph, K. W. Schmid, and R. F. H. Walter, "miRNA regulation is important for DNA damage repair and recognition in malignant pleural mesothelioma," *Virchows Archiv*, vol. 470, no. 6, pp. 627–637, Jun. 2017.
- [91] S. Yang, R. Fan, Z. Shi, K. Ji, J. Zhang, H. Wang, M. Herriid, Q. Zhang, J. Yao, G. W. Smith, C. Dong, "Identification of a novel microRNA important for melanogenesis in alpaca (*Vicugna pacos*)," *J. Animal Sci.*, vol. 93, no. 4, pp. 1622–1631, 2015.
- [92] J. Du, Y. Wu, Y. Zhang, L. Wu, X. Wang, and S. Tao, "Large-scale information entropy analysis of important sites in mature and precursor miRNA sequences," *Sci. China C, Life Sci.*, vol. 52, no. 8, pp. 771–779, Aug. 2009.
- [93] M. Koerner, K. Chhatbar, S. Webb, J. Cholewa-Waclaw, J. Selfridge, D. De Sousa, B. Skarnes, B. Rosen, M. Thomas, J. Bottomley, R. Ramirez-Solis, C. Lelliott, D. Adams, and A. Bird, "An orphan CpG island drives expression of a let-7 miRNA precursor with an important role in mouse development," *Epigenomes*, vol. 3, no. 1, p. 7, 2019.
- [94] X. Fu, W. Zhu, L. Cai, B. Liao, L. Peng, Y. Chen, and J. Yang, "Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures," *Frontiers Genet.*, vol. 10, pp. 119:1–119:12, Feb. 2019.
- [95] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, Jul. 2018.
- [96] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul. 2017.
- [97] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings Bioinf.*, vol. 17, no. 2, pp. 193–203, 2016.
- [98] L. Wang, Z. Xuan, S. Zhou, L. Kuang, and T. Pei, "A novel model for predicting LncRNA-disease associations based on the LncRNA-MiRNA-Disease interactive network," *Current Bioinf.*, vol. 14, no. 3, pp. 269–278, Mar. 2019.
- [99] J. S. Mattick, "The functional genomics of noncoding RNA," *Science*, vol. 309, no. 5740, pp. 1527–1528, Sep. 2005.
- [100] A. R. Gawronski, M. Uhl, Y. Zhang, Y.-Y. Lin, Y. S. Niknafs, V. R. Ramnarine, R. Malik, F. Feng, A. M. Chinnaiyan, C. C. Collins, S. C. Sahinalp, and R. Backofen, "MechRNA: Prediction of lncRNA mechanisms from RNA-RNA and RNA-protein interactions," *Bioinformatics*, vol. 34, no. 18, pp. 3101–3110, Sep. 2018.
- [101] L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, "LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D140–D144, 2019.
- [102] Q. Jiang, R. Ma, J. Wang, X. Wu, S. Jin, J. Peng, R. Tan, T. Zhang, Y. Li, and Y. Wang, "LncRNA2Function: A comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data," *BMC Genomics*, vol. 16, no. 3, p. S2, 2015.
- [103] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: A comprehensive Web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, Jun. 2018.



**TIERYING YU** was born in Yantai, Shandong, China, in 1997. She is currently pursuing the B.S. degree from Shandong Agricultural University, Tai'an, China. She was recommended to the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, for a straight Ph.D. degree. She is now studying bioinformatics with the University of Electronic Science and Technology of China for the period, from October 2019 to June 2020.

Her current research interests include scallop stock improvement, big data, and biocomputing.



**MIN CHEN** was born in Liaocheng, Shandong, China, in 1991. She received the B.S., M.S., and Ph.D. degrees from Shandong Agricultural University, Tai'an, China, in 2018.

Since 2018, she has been with the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences. She is the author of eight articles. Her current research interests include scallop stock improvement, chloroplast development, and chlorophyll synthesis in peaches.



**CHUNDE WANG** was born in Qingdao, Shandong, China, in 1967. He received the B.S. degree in marine biology from the Ocean University of China, Qingdao, China, in 1989, the M.S. degree in malacology from the Institute of Oceanology, Chinese Academy of Sciences, Qingdao, in 1992, and the Ph.D. degree from Dalhousie University, Halifax, NS, Canada, in 2001.

He was a Research Assistant with the Institute of Oceanology, Chinese Academy of Sciences, from 1992 to 1995. From 2000 to 2006, he worked as a Senior R&D Consultant in scallop breeding and aquaculture with Pec-Nord Inc., Quebec, QC, Canada. Since 2007, he has been with Qingdao Agricultural University (QAU), where he is currently a Professor. He is also a Researcher with the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences. He is the author of more than 40 articles. He holds over 28 patents and three officially approved new scallop strains. His research interests include scallop stock improvement, seed production and adult grow-out techniques, QTL mapping, scallop genome-wide association studies, and genome-selection breeding techniques.

Dr. Wang is a Senior Member of the Chinese Society of Malacology and a member of the National Shellfish Association of USA.

• • •