Molecular Therapy
Nucleic Acids
Original Article

Selecting Essential MicroRNAs Using a Novel Voting Method

Xiaoqing Ru,[1,2,5] Peigang Cao,[3,5] Lihong Li,[2] and Quan Zou[1,4]

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; [2]School of Information and Electrical Engineering, Hebei University of Engineering, Handan, China; [3]Department of Cardiology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China; [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

Among the large number of known microRNAs (miRNAs), some miRNAs play negligible roles in cell regulation. Therefore, selecting essential miRNAs is an important initial step for a deeper understanding of miRNAs and their functions. In this study, we generated 60 classification models by combining 12 representative feature extraction methods and 5 commonly used classification algorithms. The optimal model for essential miRNA classification that we obtained is based on the Mismatch feature extraction method combined with the random forest algorithm. The F-Measure, area under the curve, and accuracy values of this model were 93.2%, 96.7%, and 93.0%, respectively. We also found that the distribution of the positive and negative examples of the first few features greatly influenced the classification results. The feature extraction methods performed best when the differences between the positive and negative examples were obvious, and this led to better classification of essential miRNAs. Because each classifier's predictions for the same sample may be different, we employed a novel voting method to improve the accuracy of the classification of essential miRNAs. The performance results showed that the best classification results were obtained when five classification models were used in the voting. The five classification models were constructed based on the Mismatch, pseudo-distance structure status pair composition, Subsequence, Kmer, and Triplet feature extraction methods. The voting result was 95.3%. Our results suggest that the voting method can be an important tool for selecting essential miRNAs.

INTRODUCTION

MicroRNAs (miRNAs) are short noncoding RNAs that are found widely in eukaryotes.[1] Their breadth and diversity indicate that they have a very wide variety of biological functions. They are involved in many important biological processes in cells, including regulating the expression of genes that encode proteins involved in biological development,[2–4] cell proliferation,[5] differentiation,[6] and apoptosis.[7] miRNAs are associated with cancer[8–10] and other diseases.[11–15] Drugs that target genes have been developed based on miRNA gene silencing and have been applied to some previously incurable diseases that threaten human health.[16–20] miRNAs also play important roles in cell adaptation to abnormal environments, such as freezing, dehydration, and hypoxia.[21–23] Because of the many biological functions of miRNAs, a lot of attention has been given to miRNA-related problems in bioinformatics.[24–30]

Accurate identification of miRNA sequences is one such problem that has achieved good results. For example, in 2013, Wei et al.[31] constructed a classifier to identify miRNAs using a high-quality negative set and reported a classification accuracy rate of 93%. In 2015, Peace et al.[1] proposed a framework for improving miRNA prediction in non-human genomes using sequence conservation and phylogenetic distance information. Their framework uses accuracy, sensitivity, and specificity parameters to obtain species-specific predictions. In 2016, Jiang et al.[5] used a backpropagation neural network algorithm to identify miRNAs in *Arabidopsis*. In their model, the precision and recall rates were 95% and 96%, respectively; however, these results do not make much sense for the in-depth study of miRNAs. The reasons for this failure were likely because of the recent dramatic increase in known miRNAs (e.g., miRBase [Release 22.1: October 2018] contains 38,589 miRNA sequences from 271 species[32]) and the proposal that some miRNAs or miRNA families have negligible effects in cell development.[33] Therefore, to efficiently study the biological mechanisms of miRNAs, it is necessary to detect essential miRNAs from among the many other miRNAs.

Two important factors that influence miRNA prediction results are the feature extraction method and classification algorithm selected. A good feature extraction method will fully express the sequence information. The existing methods for RNA feature extraction can be divided into four categories: those based on ribonucleic acid composition, autocorrelation, pseudo ribonucleic acid composition, or predicted structure composition.[34,35] Methods based on RNA sequence composition include basic kmer (Kmer), Mismatch, and Subsequence.[31,36–39] Kmer[31] represents RNA sequences as the frequency of occurrence of k adjacent bases and is the simplest of the three methods. Methods based on autocorrelation include dinucleotide-based auto-covariance (DAC),[40] dinucleotide-based cross-covariance

Received 28 May 2019; accepted 8 July 2019;
https://doi.org/10.1016/j.omtn.2019.07.019.

[5]These authors contributed equally to this work.

Correspondence: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China.
E-mail: zouquan@nclab.net

Table 1. Performances of the Three Feature Extraction Methods with Different k Values

	Kmer			Mismatch			Subsequence		
K Values	Sn	Sp	ACC	Sn	Sp	ACC	Sn	Sp	ACC
2	67.0	80.6	73.9	96.4	89.7	93.0	91.7	88.6	90.1
3	76.4	85.2	80.9	94.1	93.1	93.6	96.4	87.5	90.7
4	83.5	90.9	87.2	89.4	92.0	90.7	90.5	88.6	89.5

ACC, accuracy; Sn, sensitivity; Sp, specificity.

(DCC),[41] dinucleotide-based auto-cross-covariance (DACC; a combination of DAC and DCC),[42] Moran autocorrelation (MAC),[43] Geary autocorrelation (GAC),[44] and normalized Moreau-Broto autocorrelation (NMBAC).[45] Methods based on the pseudo-RNA composition[46,47] include general parallel correlation pseudo-dinucleotide composition (PC-PseDNC-General) and its variant general series correlation pseudo-dinucleotide composition (SC-PseDNC-General). The equations used in PC-PseDNC-General and SC-PseDNC-General differ in that they calculate the correlation factors that reflect the sequence or the order correlations, respectively, among all of the consecutive dinucleotides along an RNA sequence.[42] Methods based on the predicted structure composition include local structure-sequence triplet elements (Triplet),[48–50] pseudo-structure status composition (PseSSC),[26] and pseudo-distance structure status pair composition (PseDPC).[13]

The most commonly used classification algorithms are random forest and support vector machine. Random forest[51–57] can be considered an integrated algorithm that reduces the one-sidedness and inaccuracy of a single decision tree by combining multiple different decision trees. Support vector machine[13,48,58–66] maximizes the classification of positive and negative examples by constructing a hyperplane. Other machine learning algorithms also have been used for classification and recognition, such as neural networks,[5,67,68] Naive Bayes,[69,70] evolutionary algorithms,[71] and ensemble learning.[72–76]

The aims of this study were: (1) to construct a classification model by combining 12 different feature extraction algorithms and 5 classification algorithms to find the most suitable model for essential miRNA classification; (2) to explore the distribution of positive and negative examples under different feature extraction methods, and to determine the influence of distribution differences between positive and negative examples on classification results; and (3) to further improve the classification accuracy of essential miRNAs through a novel voting method. The performance of the optimal classification model shows the validity of our conclusions and methods.

RESULTS AND DISCUSSION

Determine the Parameters for Kmer, Mismatch, and Subsequence

For these three feature extraction algorithms, the parameter k, which has 4^k-dimensional features, has to be set. For k = 1, the extracted features do not represent the complete sequence information. For k > 5, the extracted features will have more than 1,024 dimensions. When the dimensions are very high, the computational time can be very long, and over-fitting phenomenon and dimensionality disaster may occur. To avoid these problems, we set k = 2, 3, and 4. The results for each of the methods on the pre-miRNA dataset are shown in Table 1. Each performance value was taken from the best classification model under each method.

In the Kmer-based model, the performance was best for k = 4. In the models that used the Mismatch and Subsequence feature extraction methods, all three k values produced similar results that were better than the classification results obtained with the Kmer-based model (Table 1). Smaller k values will require a shorter computational time, so k = 2 was selected as the best value for the Mismatch- and Subsequence-based models.

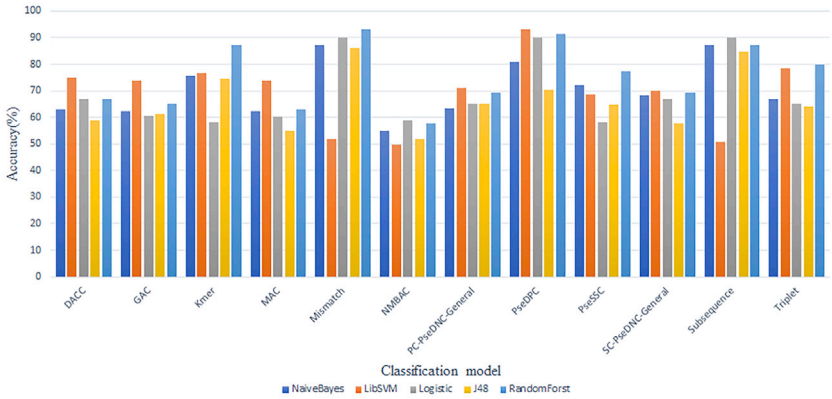

Figure 1. Accuracy of All the Classification Models

Table 2. Performances of the Three Best Classification Models

Methods	Dimensions	F-Measure (%)	AUC (%)	ACC (%)
Mismatch	16	93.2	96.7	93.0
PseDPC	515	92.6	93.0	93.0
Subsequence	16	90.2	95.1	90.1

ACC, accuracy; AUC, area under the curve.

Selection of the Best Classification Model for Predicting Essential miRNAs

A total of 12 feature extraction methods were used in this study. The models based on Kmer, Mismatch, and Subsequence involve parameter setting, as described in Determine the Parameters for Kmer, Mismatch, and Subsequence. We combined the 12 feature extraction methods with 5 commonly used classification algorithms to obtain 60 classification models. The accuracy of these models on the pre-miRNA dataset is shown in Figure 1.

The accuracy of each of the classification models varied depending on the feature extraction method that was used (Figure 1). Three classification models had accuracies >90%, namely, Mismatch + random forest, PseDPC + support vector machine, and Subsequence + Logistic. Detailed performance information is shown in Table 2.

The Mismatch + random forest model, which has low dimensionality and good performance, was considered the optimal model for predicting essential miRNAs in the dataset (Table 2).

Representation of Important Features of the Classification Models

We explored the distribution of positive and negative examples under the different feature extraction methods. As shown in Figure 1, the accuracy was <70% for all of the classification models when combined with the NMBAC and SC-PseDNC-General feature extraction methods, indicating these methods were not suitable for predicting essential miRNAs. From among the remaining 50 classification models, those with the best classification performance under each feature extraction method were selected. Among the 10 selected models, those based on the Mismatch, Triplet, and PC-PseDNC-General methods all showed better performances when combined with the random forest algorithm. On the basis of the ANOVA of these three feature extraction methods, we took out the first four-dimensional features that had the greatest influence on the classification results. The obtained distribution of positive and negative examples is shown in Figure 2. Clearly, the difference between the positive and negative examples is larger with Mismatch than with the other two methods. This indicates that a feature extraction method that produces a large difference in the distribution of positive and negative samples contributes to a better final classification result.

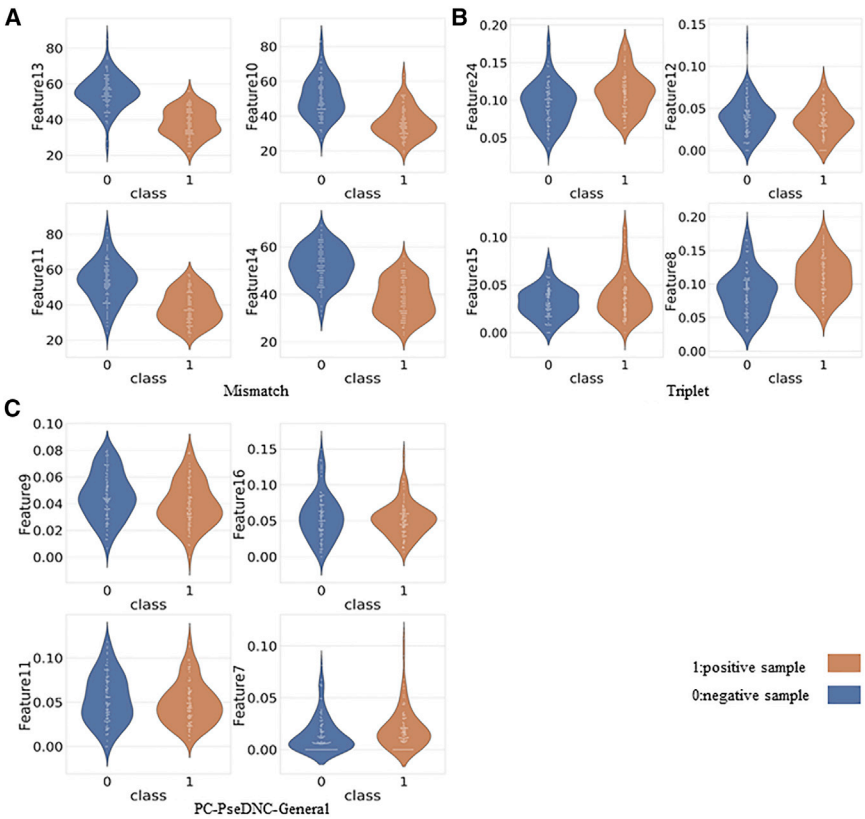

Figure 2. Distribution of Positive and Negative Examples of Important Features Obtained Using Different Extraction Methods

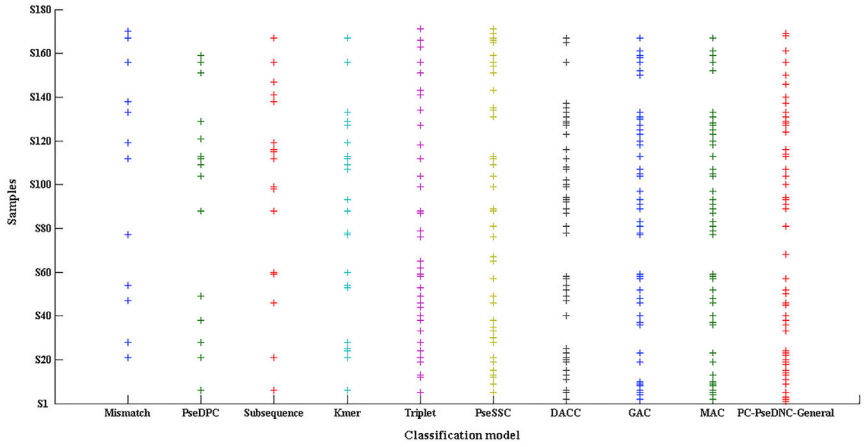

Figure 3. Distribution of Mispredicted Samples for Each Classification Model

Optimal Voting Results

To achieve higher prediction accuracy, we used a novel voting method to predict all samples based on the results from the 10 selected classification models described in Representation of Important Features of the Classification Models. The predictions of these 10 models for all samples are shown in Figure 3.

We obtained four types of voting results from the 10 classification models, and the best results for each type are shown in Table 3. Each voting process eliminates two classification models, and the eliminated models have strong correlations. For example, from the distribution of the erroneously predicted samples shown in Figure 3, the classification models based on the GAC and MAC feature extraction methods had strong correlations, and the correct or mispredicted samples were almost the same, so these methods were eliminated in the second type of voting. Excluding the most relevant classification models was beneficial to the final voting result, as shown in Table 4.

As shown in Table 3, the results obtained by voting on the classification model based on the Mismatch, PseDPC, Subsequence, Kmer, and Triplet feature extraction methods were the best (accuracy rate of 95.3%). The accuracy of voting was higher than the accuracy of the case alone.

The predictions of the classification model based on the Triplet method were worse than those of the model based on Kmer (Figure 3), but after participating in the voting with the model based on Mismatch and PseDPC, the number of samples that were mispredicted with Triplet was less than the number with Kmer. This is because the model based on Kmer had a stronger correlation with the other two classification models. We chose to vote with the classification model based on Mismatch and PseDPC because these two feature extraction methods performed best in all of the classification models, and the correlation between them was very low.

The results shown in Tables 3 and 4 fully demonstrate that the novel voting method proposed in this study can achieve excellent results for the selection of essential miRNAs.

Table 3. Best Results for Four Types of Voting

Category	Model Included	No. of Samples that Were Mispredicted	Voting Results (%)
1	Mismatch + PseDPC + Subsequence,Kmer + Triplet + PseSSC, DACC,GAC + MAC	14	91.9
2	Mismatch + PseDPC + Subsequence,Kmer + Triplet + PseSSC + DACC	9	94.7
3	Mismatch + PseDPC + Subsequence,Kmer + Triplet	8	95.3
4	Mismatch + PseDPC + Subsequence	9	94.7

Conclusions

The aim of this study was to select essential miRNAs from a large number of miRNA sequences, thus making the study of the biological mechanisms of miRNAs more efficient. We used known mouse miRNAs as the dataset. We used different feature extraction methods to represent these data, then combined the extracted features with different classification algorithms to construct classification models. The final classification result was determined by a novel voting method, which gave a final voting result of 95.3%. This result showed that this method was effective in identifying the essential miRNAs in the dataset. In future work, we will focus on detecting new essential miRNAs, analyzing their function, and exploring the relationship between new miRNAs and diseases.[27,77]

MATERIALS AND METHODS

The general pipeline used in this study is shown in Figure 4.

Acquisition of Datasets

Acquisition of essential pre-miRNA sequences: miRNA genes produce primary miRNA (pri-miRNA) sequences that are 300–1,000 nt long. The pri-miRNAs are processed to precursor miRNA (pre-miRNA) sequences that are 60–70 nt long. Mature miRNAs, which are 20–24 nt long, are formed from pre-miRNAs by the action of enzymes.[26,48,78] The hairpin structure of pre-miRNAs is an important feature that is widely used to identify miRNAs.[48] In this study, we used pre-miRNA sequences from miRBase (http://www.mirbase.org/). We collected a total of 91 pre-miRNA sequences that are essential in mice from Bartel's 2018 review of metazoan miRNAs.[79] The 91 pre-miRNA sequences were from several families.

Table 4. Verifying the Impact of Model Relevance on Voting

Reserved Classification Model	No. of Samples that Were Mispredicted	Voting Results (%)
Mismatch + PseDPC + Subsequence	9	94.7
Mismatch + PseDPC + Kmer	14	91.9
Mismatch + PseDPC + Triplet	9	94.7

Acquisition of pre-miRNA sequences of unknown importance: We downloaded all the mice pre-miRNA sequences (a total of 1,234) from miRBase. All the sequences that belonged to the same families as the 91 essential miRNAs were excluded. The remaining pre-miRNAs (1,090) were considered to be non-essential pre-miRNAs.

To shorten the computational time and to ensure better performance results, we removed redundant sequences[80] and obtained a final dataset that contained 85 essential and 88 non-essential pre-miRNA sequences.

Feature Extraction Methods

Many methods have been used for extracting features of RNA sequences. In this study, we used a number of feature extraction methods on the selected pre-miRNA sequences. We expressed a pre-miRNA sequence as $R = r_1, r_2, r_3, r_4, r_5, \ldots, r_L$, where $r_i \in \{A, U, G, C\}$,[33] and L is the sequence length. These features can be generated easily using the repRNA web server[81] and the BioSeq-Analysis platform.[82]

Kmer and Mismatch

Kmer is a simple feature extraction method in which k indicates the number of bases in a subsequence. For a given k value, there will be 4^k seed sequences.[31] For example, for k = 2, there are 16 subsequences, AA, AC, AU, AG, UA, UU, …, CC. Mismatch calculates the number of occurrences of subsequences containing k adjacent bases and uses a parameter m (maximum allowed error match = $0 \leq m < k$).[83] For k = 2, m is 0, 1. For example, for the subsequence AC, the A– and –C subsequences are all regarded as AC.

Subsequence

Subsequence also counts the number of occurrences of a subsequence, but, in particular, it takes into account the length of the sequence that is eligible and the influence factor ∂.[84] This method allows for spacing between the bases in a given subsequence. For example, there are four ways in which a search for AAC in a sequence is eligible: AAC, AAXXXC, AXXXAC, and AXXAC (where X can be U or G). For this example, the number of occurrences of AAC can be expressed as: $1 + \partial 6 + \partial 6 + \partial 5$. For AAC, where there is no gap between the bases, the length is considered to be 1.

Triplet

Triplet extracts features based on secondary structure information in the sequence. Two states are considered for each nucleotide: matched (represented by "(" or ")" for A paired with U and G paired with C) and unmatched (represented by "."). The possible structural forms for any three adjacent nucleotides are: "(((," "((.," "(.(," ".((," "(..," ".(. ," "..(," and "…." For the four bases (A, U, G, C) in a pre-miRNA sequence, there will be 8×4 different triplets. Triplet counts the number of occurrences of these 32 triplets in a sequence.[48]

PseDPC

Pre-miRNA sequences can be represented by 10 secondary structure states: A, C, G, U, A–U, U–A, G–C, C–G, G–U, and U–G. Within a given range of distance thresholds, the PseDPC algorithm counts the frequency of occurrence of the two-two combination between the structural states. When the distance d is equal to 0, only the frequencies of each of the 10 states are counted. When d is ≥ 1, there will be 100 combinations of the 10 states, so the frequency of occurrence of the 100 combinations is counted. Therefore, 10 + 100d dimensional features can be extracted. Then the features of the free energy between the two structural states of a combination are extracted, giving φ features, where φ is the highest counted rank of the structural correlation along an RNA chain.[13]

Voting Process

The results produced by each classifier were integrated using a voting process as follows. First, each classification algorithm will correctly or incorrectly predict each of the samples (a sequence is considered a sample). Correctly predicted samples are not marked, the samples that were mispredicted are marked with "+," and the number of classifiers is represented by n. Second, when n is a singular number ≥ 3, the number of times each sample is erroneously predicted under n classifiers is counted and recorded as m, where $m \geq n + 1/2$; the sample is considered to be a sample that was mispredicted. Third, when n is a double number ≥ 4, the number of times each sample is mispredicted under the n classifiers is counted, and the sample

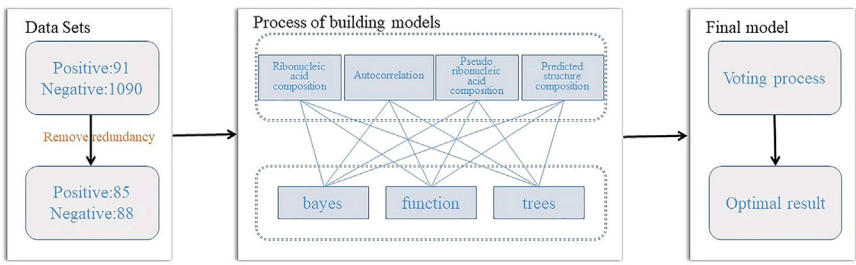

Figure 4. Flowchart of the Processes Used in This Study

corresponding to m = n/2 is selected. Suppose that there are z samples that meet the requirements, and the z samples are predicted differently by the different classifiers, the classifier with the highest number of mispredicted samples is eliminated. Then, n is singular and step 2 can be repeated. Fourth, when n satisfies the condition of being a singular number, steps 2 and 3 can count the number of samples considered to be mispredicted in various voting processes, from which the final voting result can be derived.

AUTHOR CONTRIBUTIONS

X.R. implemented the experiments and drafted the manuscript. P.C., L.L., and Q.Z. initiated the idea, conceived the whole process, and finalized the paper. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The work was supported by the National Key R&D Program of China (grant 2018YFC0910405) and the Natural Science Foundation of China (grant 61771331).

REFERENCES

1. Peace, R.J., Biggar, K.K., Storey, K.B., and Green, J.R. (2015). A framework for improving microRNA prediction in non-human genomes. Nucleic Acids Res. *43*, e138.
2. La Torre, A., Georgi, S., and Reh, T.A. (2013). Conserved microRNA pathway regulates developmental timing of retinal neurogenesis. Proc. Natl. Acad. Sci. USA *110*, E2362–E2370.
3. Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. Sci. Rep. *6*, 34820.
4. Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. BMC Bioinformatics *19* (*Suppl 5*), 116.
5. Jiang, L., Zhang, J., Xuan, P., and Zou, Q. (2016). BP Neural Network Could Help Improve Pre-miRNA Identification in Various Species. BioMed Res. Int. *2016*, 9565689.
6. Le, M.T., Xie, H., Zhou, B., Chia, P.H., Rizk, P., Um, M., Udolph, G., Yang, H., Lim, B., and Lodish, H.F. (2009). MicroRNA-125b promotes neuronal differentiation in human cells by repressing multiple targets. Mol. Cell. Biol. *29*, 5290–5305.
7. Körner, C., Keklikoglou, I., Bender, C., Wörner, A., Münstermann, E., and Wiemann, S. (2013). MicroRNA-31 sensitizes human breast cells to apoptosis by direct targeting of protein kinase C epsilon (PKCepsilon). J. Biol. Chem. *288*, 8750–8761.
8. Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2019). Discovering Cancer Subtypes via an Accurate Fusion Strategy on Multiple Profile Data. Front. Genet. *10*, 20.
9. Yu, L., Zhao, J., and Gao, L. (2018). Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity. Int. J. Biol. Sci. *14*, 971–982.
10. Pavithra, D., Sabitha, K., and Rajkumar, T. (2018). Identification of small molecule inhibitors for differentially expressed miRNAs in gastric cancer. Comput. Biol. Chem. *77*, 442–454.
11. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37*, D98–D104.
12. Cheng, L., and Hu, Y. (2018). Human Disease System Biology. Curr. Gene Ther. *18*, 255–256.
13. Liu, B., Fang, L., Liu, F., Wang, X., and Chou, K.C. (2016). iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J. Biomol. Struct. Dyn. *34*, 223–235.
14. Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM rs3851179 variant confers susceptibility to Alzheimer's disease in Chinese population. Mol. Neurobiol *54*, 3131–3136.
15. Hu, Y., Zhao, T., Zang, T., Zhang, Y., and Cheng, L. (2019). Identification of Alzheimer's Disease-Related Genes Based on Data Integration Method. Front. Genet. *9*, 703.
16. Kelly, P.S., Gallagher, C., Clynes, M., and Barron, N. (2015). Conserved microRNA function as a basis for Chinese hamster ovary cell engineering. Biotechnol. Lett. *37*, 787–798.
17. Jiang, Q., Jin, S., Jiang, Y., Liao, M., Feng, R., Zhang, L., Liu, G., and Hao, J. (2017). Alzheimer's Disease Variants with the Genome-Wide Significance are Significantly Enriched in Immune Pathways and Active in Immune Cells. Mol. Neurobiol. *54*, 594–600.
18. Liu, G., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., Han, Z., Xu, D., and Jiang, Q. (2018). Circulating vitamin E levels and Alzheimer's disease: a Mendelian randomization study. Neurobiol. Aging *72*, 189.e1–189.e9.
19. Liu, G., Zhang, Y., Wang, L., Xu, J., Chen, X., Bao, Y., Hu, Y., Jin, S., Tian, R., Bai, W., et al. (2018). Alzheimer's Disease rs11767557 Variant Regulates EPHA1 Gene Expression Specifically in Human Whole Blood. J. Alzheimers Dis. *61*, 1077–1088.
20. Liu, G., Wang, T., Tian, R., Hu, Y., Han, Z., Wang, P., Zhou, W., Ren, P., Zong, J., Jin, S., and Jiang, Q. (2018). Alzheimer's Disease Risk Variant rs2373115 Regulates GAB2 and NARS2 Expression in Human Brain Tissues. J. Mol. Neurosci. *66*, 37–43.
21. Biggar, K.K., Kornfeld, S.F., Maistrovski, Y., and Storey, K.B. (2012). MicroRNA regulation in extreme environments: differential expression of microRNAs in the intertidal snail Littorina littorea during extended periods of freezing and anoxia. Genomics Proteomics Bioinformatics *10*, 302–309.
22. Biggar, K.K., and Storey, K.B. (2012). Evidence for cell cycle suppression and microRNA regulation of cyclin D1 during anoxia exposure in turtles. Cell Cycle *11*, 1705–1713.
23. Wu, C.W., Biggar, K.K., and Storey, K.B. (2013). Dehydration mediated microRNA response in the African clawed frog Xenopus laevis. Gene *529*, 269–275.
24. Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2018). FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association. BMC Genomics *19* (*Suppl 10*), 911.
25. Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics *34*, 1953–1956.
26. Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.C. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE *10*, e0121501.
27. Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. Int. J. Data Min. Bioinform. *8*, 282–293.
28. Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., Qian, J., and Wang, Y. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. Nucleic Acids Res. *46* (D1), D146–D151.
29. Wang, G., Wang, F., Huang, Q., Li, Y., Liu, Y., and Wang, Y. (2015). Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. BioMed Res. Int. *2015*, 757530.
30. Gong, W., Huang, Y., Xie, J., Wang, G., Yu, D., and Sun, X. (2017). Genome-wide identification and characterization of conserved and novel microRNAs in grass carp (Ctenopharyngodon idella) by deep sequencing. Comput. Biol. Chem. *68*, 92–100.
31. Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. IEEE/ACM Trans. Comput. Biol. Bioinform *11*, 192–201.
32. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47* (D1), D155–D162.
33. Alvarez-Saavedra, E., and Horvitz, H.R. (2010). Many families of C. elegans microRNAs are not essential for development or viability. Curr. Biol. *20*, 367–373.
34. Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., Tan, R., Zhang, T., Li, Y., and Wang, Y. (2015). LncRNA2Function: a comprehensive resource for functional

investigation of human lncRNAs based on RNA-seq data. BMC Genomics *16* (*Suppl 3*), S2.

35. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47* (D1), D140–D144.
36. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., and Noble, W.S. (2004). Mismatch string kernels for discriminative protein classification. Bioinformatics *20*, 467–476.
37. El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008). Predicting flexible length linear B-cell epitopes. Comput. Syst. Bioinformatics Conf. *7*, 121–132.
38. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. *43*, W65–W71.
39. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.
40. Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. *36*, 3025–3030.
41. Dong, Q., Zhou, S., and Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics *25*, 2655–2662.
42. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.
43. Horne, D.S. (1988). Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers *27*, 451–477.
44. Sokal, R.R., and Thomson, B.A. (2006). Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am. J. Phys. Anthropol. *129*, 121–131.
45. Feng, Z.P., and Zhang, C.T. (2000). Prediction of membrane protein types based on the hydrophobic index of amino acids. J. Protein Chem. *19*, 269–275.
46. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics3024762535, 1469–1477.
47. Dao, F.Y., Lv, H., Wang, F., Feng, C.Q., Ding, H., Chen, W., and Lin, H. (2019). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics *35*, 2075–2083.
48. Xue, C., Li, F., He, T., Liu, G.P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics *6*, 310.
49. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl. Based Syst. *163*, 787–793.
50. Tan, J.X., Li, S.H., Zhang, Z.M., Chen, C.X., Chen, W., Tang, H., and Lin, H. (2019). Identification of hormone binding proteins based on machine learning methods. Math. Biosci. Eng. *16*, 2466–2480.
51. Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., Yang, J., Sun, H., Zhao, Y., and Yang, J. (2017). Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. Sci. Rep. *7*, 1545.
52. Cutler, A., Cutler, D., and Stevens, J. (2011). Random Forests. Machine Learning *45*, 157–176.
53. Ding, Y., Tang, J., and Guo, F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinformatics *17*, 398.
54. Liu, B., Yang, F., Huang, D.S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.
55. Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., and Gao, L. (2017). Prediction of Novel Drugs for Hepatocellular Carcinoma Based on Multi-Source Random Walk. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 966–977.
56. Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., and Hu, Y. (2018). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genomics *19* (*Suppl 1*), 919.
57. Cheng, L., Shi, H., Wang, Z., Hu, Y., Yang, H., Zhou, C., Sun, J., and Zhou, M. (2016). IntNetLncSim: an integrative network analysis method to infer human lncRNA functional similarity. Oncotarget *7*, 47864–47874.
58. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. IEEE Intell. Syst. *13*, 18–28.
59. Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. Inf. Sci. *418–419*, 546–560.
60. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: A Sequence-Based Predictor for Identifying 2′-O-Methylation Sites in Homo sapiens. J. Comput. Biol *25*, 1266–1277.
61. Yang, W., Zhu, X.J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. Curr. Bioinform. *14*, 234–240.
62. Liu, Y., Wang, X., and Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief. Bioinform. *20*, 330–346.
63. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796–2800.
64. Sun, Y., Xiong, Y., Xu, Q., and Wei, D. (2014). A hadoop-based method to predict potential effective drug combination. BioMed Res. Int. *2014*, 196858.
65. He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. BMC Bioinformatics *19*, 306.
66. Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA Promoter Identification in Arabidopsis Using Multiple Histone Markers. BioMed Res. Int. *2015*, 861402.
67. Song, T., Rodríguez-Patón, A., Zheng, P., and Zeng, X. (2018). Spiking Neural P Systems with Colored Spikes. IEEE Trans. Cogn. Dev. Syst. *10*, 1106–1115.
68. Cabarle, F.G.C., Adorna, H.N., Jiang, M., and Zeng, X. (2017). Spiking Neural P Systems With Scheduled Synapses. IEEE Trans. Nanobioscience *16*, 792–801.
69. Feng, P.M., Ding, H., Chen, W., and Lin, H. (2013). Naïve Bayes classifier with feature selection to identify phage virion proteins. Comput. Math. Methods Med. *2013*, 530696.
70. Feng, P.M., Lin, H., and Chen, W. (2013). Identification of antioxidants from sequence information using naïve Bayes. Comput. Math. Methods Med. *2013*, 567529.
71. Xu, H., Zeng, W., Zhang, D., and Zeng, X. (2019). MOEA/HD: A Multiobjective Evolutionary Algorithm Based on Hierarchical Decomposition. IEEE Trans. Cybern. *49*, 517–526.
72. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.
73. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.
74. Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. Mol. Ther. Nucleic Acids *12*, 635–644.
75. You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics *34*, 2465–2473.
76. Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D.Q. (2018). PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method. Front. Microbiol. *9*, 2571.
77. Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y., and Wang, Y. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst. Biol. *4* (*Suppl 1*), S2.
78. Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. *35*, W339–W344.

79. Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20–51.

80. Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. Brief. Bioinform. *10*, 1106–1115.

81. Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K.-C. (2016). repRNA: a web server for generating various feature vectors of RNA sequences. Mol. Genet. Genomics *291*, 473–481.

82. Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief. Bioinform., *2017*, bbx165.

83. Luo, L., Li, D., Zhang, W., Tu, S., Zhu, X., and Tian, G. (2016). Accurate Prediction of Transposon-Derived piRNAs by Integrating Various Sequential and Physicochemical Features. PLoS ONE *11*, e0153268.

84. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002). Text Classification using String Kernels. J. Mach. Learn. Res. *2*, 419–444.