The Author(s) *Journal of Biomedical Semantics* 2017, **8**(Suppl 1):30
DOI 10.1186/s13326-017-0143-z

Journal of Biomedical Semantics

RESEARCH Open Access

A novel method to identify pre-microRNA in various species knowledge base on various species

Tianyi Zhao[1], Ningyi Zhang[1], Ying Zhang[2], Jun Ren[3], Peigang Xu[1], Zhiyan Liu[1], Liang Cheng[4*] and Yang Hu[3*]

From Biological Ontologies and Knowledge bases workshop on IEEE BIBM 2016
Shenzhen, China. 16 December 2016

Abstract

Background: More than 1/3 of human genes are regulated by microRNAs. The identification of microRNA (miRNA) is the precondition of discovering the regulatory mechanism of miRNA and developing the cure for genetic diseases. The traditional identification method is biological experiment, but it has the defects of long period, high cost, and missing the miRNAs that but also many other algorithms only exist in a specific period or low expression level. Therefore, to overcome these defects, machine learning method is applied to identify miRNAs.

Results: In this study, for identifying real and pseudo miRNAs and classifying different species, we extracted 98 dimensional features based on the primary and secondary structure, then we proposed the BP-Adaboost method to figure out the overfitting phenomenon of BP neural network by constructing multiple BP neural network classifiers and distributed weights to these classifiers. The novel method we proposed, from the 4 evaluation terms, have achieved greatly improvement on the effect of identifying true pre-RNA compared to other methods. And from the respect of identifying species of pre-RNA, the novel method achieved more accuracy than other algorithms.

Conclusions: The BP-Adaboost method has achieved more than 98% accuracy in identifying real and pseudo miRNAs. It is much higher than not only BP but also many other algorithms. In the second experiment, restricted by the data, the algorithm could not get high accuracy in identifying 7 species, but also better than other algorithms.

Keywords: Pre-miRNA identification, BP neural network, Adaboost

Background

MicroRNAs(miRNAs) are a class of small single-strand and non-coding RNA molecules of approximately 22 nucleotides in length, miRNAs play important roles in many biological process including affecting stability, metabolism, signal translation, disease development and translation of mRNAs [1]. Meanwhile, miRNAs are also very important in the treatment of diseases, such as: cancer [2], X chromosomal defects [3], DiGeorge disease [4], etc. As the development of science and technology, people pay more and more attention to miRNA research, amount of novel miRNAs are discovered, the number and functional features are far beyond our imagination [5, 6]. The main challenge of studying miRNAs is how to find miRNAs and the action sites, at present the main methods for identifying miRNAs are cDNA clone and sequencing and computational prediction, the expression of cDNA sequencing method is low and costs amount of time and funding [7–12]. Therefore, computational method are more prevalent, several algorithms have been proposed to detect pre-miRNAs, the main challenge is to discriminate the real pre-miRNAs from the pseudo ones and identify novel miRNAs.

* Correspondence: liangcheng@hrbmu.edu.cn; huyang@hit.edu.cn
[4]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150001, China
[3]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China
Full list of author information is available at the end of the article

Recently, for miRNAs identification, machine learning techniques have been widely used. Sequence composition and structural conformation features are applied to train the learning system, then the classifiers employ multiple features to obtain the final prediction. Xue et al. discovered the significant difference of local contiguous subsequence between real and pseudo miRNAs. Therefore, they applied the three-character group local structure-sequence features to describe the samples, and based on SVM they proposed the triplet-SVM to identify novel miRNAs and miRNAs from specific species [13]. Zhao et al. employed parallel triplet local structure-sequence feature, however they chose the first nucleotide of the contiguous triplet group as the local structure-sequence feature, and add two MFE related features and two nucleotide pairing features, then apply the SVM classifier PMirP [14]. Jiang et al. added MFE and *P*-value as the features based on the feature set in [13], and proposed the classifier Mipred based on RF method [15], Limin Jiang. et [16] applied BP neural network to identify real and pseudo pre-miRNAs, and proved the superiority of BP neural network by comparing with triplet-SVM、RF methods.

Neural network [17] and other classifiers of data driving tend to occur overfitting phenomenon. BP neural network is a widely used classification algorithm, it has strong self-learning ability and is particularly suitable for solving internal mechanism problems. However, the algorithm tends to be overfitting and the output is unstable. The boosting algorithm [18] integrates multiple weak classifiers to obtain a strong classifier and avoid overfitting phenomenon. Freund [19] promoted the boosting algorithm to Adaboost(adaptive boosting) so that the new algorithm can be more suitable for practical applications.

Therefore, in this study, we proposed BP-Adaboost algorithm to establish multiple BP neural network classifiers and distribute the weights of classifiers through Adaboost framework. Eventually a strong classifier with high accuracy is obtained.

Methods

Feature extraction

N-gram frequency

In the recent years, for pre-miRNA identifying, studies have shown that the local primary sequence is crucial to the pre-miRNA sequence [20]. Therefore, the n-gram frequency is the most commonly used feature in the primary sequence feature selecting [21, 22]. However, there is still no exact criteria for choosing the value of n. Thus, n is often chosen by comparing the effect of n-gram frequency with different n-values. In this study, we chose n as 3. Thus, for a certain sequence, there are 64(43) combinations in a triple-nucleotide group, then we computed the frequency occurrence of these 64 combinations in the sequence.

Energy characteristics features

Some studies showed that the minimum free energy (MFE) indicates the stability of a secondary structure. The real pre-miRNA sequences have a lower minimum free energy than that of the randomly generated pre-miRNA sequences. Therefore, the minimum free energy of a pre-miRNA sequence is also considered as a feature in distinguishing the pre-miRNA sequences. RNAfold is used to compute the MFE value of a secondary structure.

Structural-diversity based features

The base-pair of nucleotide in the sequences is also a remarkable characteristic in distinguishing real and pseudo pre-miRNAs. The traditional nucleotide pairing are A-U pairing and C-G pairing, but in pre-miRNA sequences there are also other forms of nucleotide pairing, such as the G-U pairing. Therefore, in this study, the G-U pairing is also included as one of the features.

Triple structure sequence

To highly specify the primary sequence features, the secondary structure is also a significant feature. The software RNAfold is employed to calculate the potential secondary structure. In the predicted secondary structure, there are two states for each nucleotide of the sequence, matched or non-matched, indicated by brackets '(' or ')', and dots '.'. In this study, the two brackets are not distinguished, which means every ')' is replaced by '('. For any three nucleotides groups, there are 8 (2×3) possible characters combinations, including '(((', '((.', '(.(', '(..', '.((', '.(.', '..(' and '…'. Considering the first nucleotide of the three characters group, then there are 32 (4 × 8) different combinations, which are denoted as 'A(((','U(..', etc. For a given sequence, the 32 dimensional feature vector is sufficient information for miRNA identification. Then the calculated 32D feature is employed to train the classifier.

Using the feature extraction methods which we mentioned above, we can extract 98D features from any pre-miRNA sequence in total.

First we integrate the obtained pre-sequence into primary sequence and secondary structure sequence. For the primary sequence, we choose the n-gram parameter(n = 3) and extracted 64 dimensional feature. In addition to calculating the potential structure, the software RNAfold is applied to predict the second structure sequence. In this structure, we extracted 32 dimensional features according to the triple structure sequence. We also extract the energy characterization MFE as a feature of pre-miRNA sequence. The possible nucleotide pairing G-U is included as the last feature. Therefore, altogether we extracted 98 features.

The Flow chart of Feature extraction is illustrated in Fig. 1.

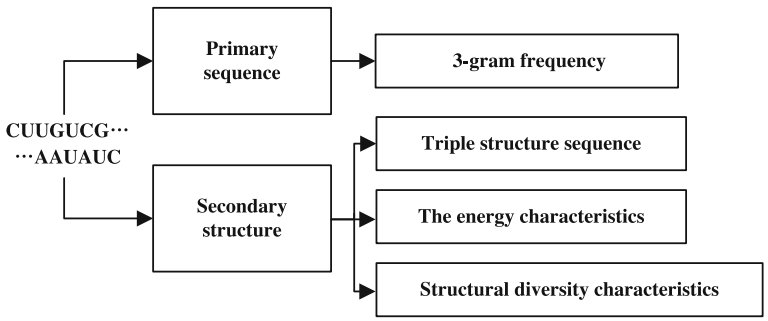

Fig. 1 Flow chart of Feature extraction

Methods and framework

BP-Adaboost

Due to BP neural network tends to be caught in overfitting phenomenon and unstable output, in this study, we proposed a new method BP-Adaboost based on BP neural network. We employed BP neural network as a weak classifier to establish multiple classification model by training repeatedly. Finally, a strong classifier is obtained after adjusting the weights through Adaboost.

The framework is shown in Fig. 2.

First, we establish N BP network classifiers by the extracted features and their corresponding labels. While training and establishing classifiers, each classifier will get a corresponding weight. In the end, we obtained a strong classifier by combining these N weight-distributed classifiers.

The construction of BP neural network classifier

To accomplish the construction of classifier, we also need to set the various parameters besides the obtained features and the corresponding labels.

First, we need to choose the number of nodes in the hidden layer, since there is no specific criterion at present, we choose the number through the empirical formula as followed.

$$M = \sqrt{N + L} + a \tag{1}$$

The number of nodes(M) equals to a constant($a \in [1, 10]$) plus the square root of the number of feature dimensions(N) plus the output(L). In this study, N = 98, L = 1. From the formula above, in this study, we choose M = 12.

After setting the number of nodes (12) and hidden layers(3) of the BP neural network, the structure of BP network is 98–12-1. Then set the paramaters (Epochs, The learning rate, Error bounds) and functions(Performance function, Transfer function of hidden layer nodes, Transfer function of output nodes,The training function) of BP network. The parameters and functions of BP network are listed in Table 1.

Method process

For a given set of multiple classification training data $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, the input data $x_N \in X \quad \mathrm{R}^n$, with an arbitrary integer label y_N. First, initialize the weight distribution, the initial weight of each sample is 1/N. Then train the sample to get the first classifier, and reduce the weight of the correct classification samples while raising the weight of improper classification samples. By the statistics of the weights

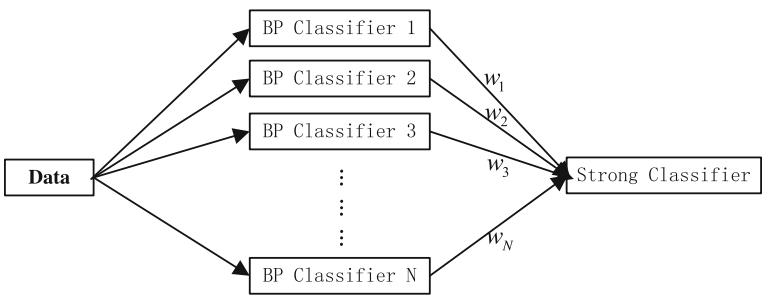

Fig. 2 Frame of BP-Adaboost

Table 1 Parameters and functions of BP neural network

Setting items	The value set
Epochs	50
The learning rate	0.1
Performance function	MSE
Error bounds	0.01
Transfer function of hidden layer nodes	Tansig
Transfer function of output nodes	Purelin
The training function	Trainlm

of improper classification samples, the weights of corresponding classifiers are obtained. Repeat the process above, we can get multiple classifiers and the corresponding weights, then the ultimate strong classifier are obtained. The method process is as followed,

Algorithm BP-Adaboost

Input: $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$

Output: Strong classifier $G(x)$

Step 1. Initialize the weight of training data set

$$w_i = \frac{1}{N}, i = 1, 2, 3, \cdots, N$$

Step 2. Obtain M weak classifiers

for m=1:M

1) Train samples and obtain BP weak classifier: $G_m(x)$

2) Calculate the error rate of $G_m(x)$

$$err_m = \sum_{i=1}^{N} w_{mi} \, || \, (G_m(x_i) \neq y_i)$$

3) Get the weight of the weak classifier

$$a_m = \frac{1}{2} \log \frac{1 - err_m}{err_m}$$

4) Update the weight of sample data

$$w_{mi} = \frac{w_{mi}}{Z_m} \exp(-a_m y_i G_m(x_i)), i = 1, 2, \cdots, N$$

5) Calculate Z (normalization factor)

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-a_m y_i G_m(x_i))$$

end

Step 3. Obtain strong classifier

$$G(x) = \arg\max \sum_{m: G_m(x) = y} a_m$$

Results

Data description

The dataset of pre-miRNAs was downloaded from http://bioinf.sce.carleton.ca/SMIRP [23]. There are 7 species samples in the data set and each species has both a positive sample set and a negative sample set.

These 7 species are *Anolis carolinensis*、*Arabidopsis lyrata*、*Arabidopsis thaliana*、*Drosophila melanogaster*、*Drosophila pseudoobscura*、Epstein barrvirus、Xenopus tropicalis. As each species has a negative sample set, altogether we obtain 8 classes of pre-miRNAs, one of them is pseudo pre-miRNAs. The total number of the gene sequences of the whole data set is 12,846, among them 9264 sequences are pseudo pre-miRNAs, and the rest 3582 of them are true pre-miRNAs.

In this article, we distinguish the real pre-miRNAs from the pseudo ones before classifying these 7 species. *V*-fold cross-validation with moderate computational complexity is widely used for model selection. Usually, a value of V between 5 and 10 is selected based on experience. In this study, V = 10. First, the 12,846 sequences are randomly divided into 10 groups, and choose 9 of them to be training samples. The last one is tested as the testing set for a total of 10 training times. The final statistical results are averaged.

Evaluation criteria

The four kinds of prediction results are true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Many evaluation indicators can be used for the classification results. First, the accuracy rate (ACC) is the proportion of the correct classification. Precision and recall are common used evaluation criteria in pattern recognition, precision represents the proportion of true positive samples of the classified positive samples, and recall represents the proportion of correctly classified positive samples of the whole positive samples, specificity represents the proportion of the correctly classified negative samples of the whole negative samples, the computational formula is as follows,

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$specificity = \frac{TN}{TN + FP} \tag{5}$$

The authentic classification of pre-miRNAs

In this study, the label of pseudo pre-miRNAs is 0, and the label of real pre-miRNAs is 1(for all species).

Figure 3 shows the curves of the four error metrics for 10 experiments, blue dot-solid line is the ACC curve, purple dotted line is precision curve, red solid line is

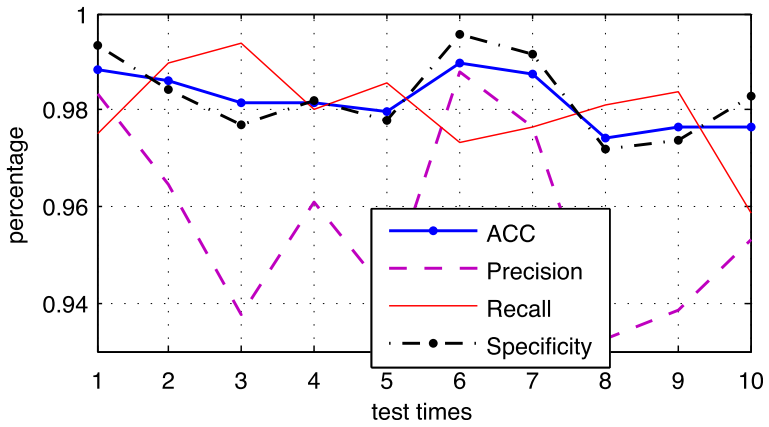

Fig. 3 Results of 4 error standards of ten experiments

recall curve, and black dash-dotted line is specificity curve. It is observed that the fluctuation of the precision curve is relatively large, and the rest three curves are more stable.

The error statistics of the average results of 10 experiments are shown in the Table 2. The table shows that BP-Adaboost algorithm is superior to other 4 algorithms in these 4 accuracy assessment, and Naïve Bayes is the worst. The accuracy of BP-Adaboost algorithm reaches 98.22%, and this represents the superiority and effectiveness of this novel method we proposed in distinguishing real and pseudo mi-RNAs. The table also shows the accuracy of BP neural network is only second to BP-Adaboost algorithm, and that's the reason why we choose the BP neural network combined with Adaboost algorithm. Due to the randomness of BP network, by weighting multiple classifiers to obtain the final results effectively improves the classification accuracy and stability.

Species classification of pre-miRNAs

The Fig. 4 shows the classification results from the statistic of each classified real pre-miRNA sequence.

Table 2 Comparison of the BP-Adaboost with alternative models

Algorithm	ACC	Precision	Recall	Specificity
BP-Adaboost	0.9822	0.9576	0.9797	0.9830
BP	0.9541	0.9429	0.9736	0.9800
Random Forest	0.9336	0.9270	0.9744	0.9772
Naïve Bayes	0.7026	0.4831	0.9721	0.5987
SVM	0.8811	1	0.5729	1

There are 7 species in the graph, red bar shows the true number of the species, blue bar shows the correctly classified number. From the graph we can tell the accuracy of the classification of some species are not ideal, the reason is the number of pseudo pre-miRNA sequences is large. In the cross validation, real pre-miRNA sequences of the samples are not enough, so the number of real pre-miRNA sequences is smaller after the classification. Therefore, the samples of some kind of species are more likely not enough.

The classification accuracy of each species is shown as Table 3.

It can be seen from the table that the accuracy of BP-Adaboost algorithm is superior to other algorithms, although the accuracy of the rest 4 algorithms is higher in some specific species, the accuracy of the algorithm in this study is the highest in total.

Discussions

In this paper, we proposed a new method to identify the pre-miRNAs. We use the Adabost algorithm to generate ten BP classifiers to finish the identification. It can provide a new thinking of solving the problem of miRNAs identification. We use several original algorithms to compare with our method, and found that our method can achieve better performance than them. Although the method can fully play the generalization of BP, and make the whole method hardly over-fit, it still has the problems such as: its performance is not ideal in solving multi-classification problem with imbalanced samples, its training time is longer than normal BP algorithm. In the respect of experiment, the method should be tested in many other data sets to verify the effectiveness of BP-Ababoost.

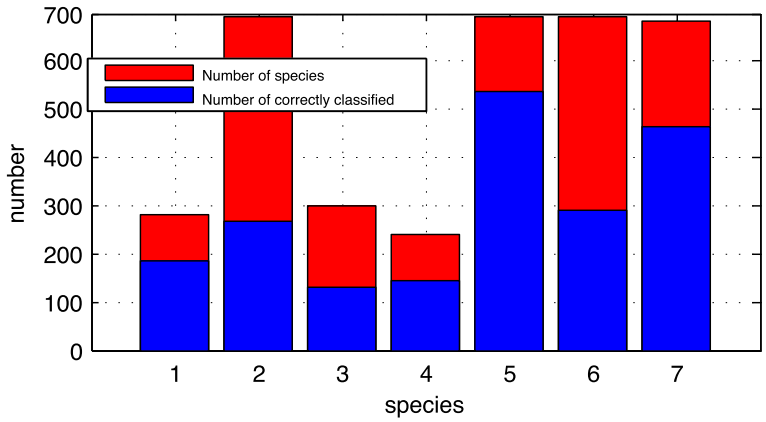

Fig. 4 Results of 7 species classify by BP-Adaboost

Conclusions

The identification of miRNAs is significant for human to study its function and understand its network regulation mechanism, discovering more novel miRNAs can also promote the prediction of miRNA target genes and the development of new drugs. In this study, we proposed a method combined BP neural network with Adaboost algorithm, it can effectively overcome the defects of unstable output and overfitting phenomenon, our method obtained a strong classifier by integrating multiple week classifiers (BP neural network classifiers) and distributing the weights to them. The data set of traditional classification of real and pseudo pre-miRNA sequences combined the real and pseudo sequences of one species together, in this study, we combined the real and pseudo sequences of 7 different species together, which increased the diversity and difficulty of the classification. In the end, we obtained a high accuracy identification result of real and pseudo pre-miRNAs. Beyond that, in this study we also classified 7 different species from pre-miRNAs which is the part that few people are paying attention to. Due to that, the sample data is not enough, though the accuracy of our classifier is higher than other methods, but the overall classification result is still need to be proved. However, the method we proposed is still able to provide guidance for the miRNA identification.

Table 3 Accuracy's comparison of the BP-Adaboost with alternative models in 7 species

Species	BP-Adaboost	BP	RF	Naïve Bayes	SVM
Anolis carolinensis	0.66	0.10	0.78	0.69	0.14
Arabidopsis lyrata	0.39	0.25	0.53	0.21	0
Arabidopsis thaliana	0.45	0.23	0.67	0.54	0
Drosophila melanogaster	0.61	0.20	0.51	0.75	0.21
Drosophila pseudoobscura	0.79	0.35	0.31	0.41	0.14
Epstein barrvirus	0.42	0.26	0.24	0.06	0.10
Xenopus tropicalis	0.68	0.43	0.45	0.07	0
Total	0.57	0.29	0.51	0.30	0.22

Abbreviations

ACC: Accuracy rate; BP-Adaboost: Back Propagation neural network fused with Adaboost algorithm; FN: False Negative; FP: False Positive; miRNA: microRNA; MSE: Mean Square Error; RF: Random Forest algorithm; SVM: Support Vector Machine algorithm; TN: True Negative; TP: True Positive

Acknowledgments

Yang Hu, Zhiyan Liu and Liang Cheng are the corresponding author. Tianyi Zhao, Ningyi Zhang and Ying Zhang are the co-first author.

Funding

This work was supported by the National Natural Science Foundation of China (No: 61,571,152 and 61,502,125, $2000), the National High-tech R&D Program of China (863 Program, $2500) [Nos: 2014AA021505, 2015AA020101, 2015AA020108], the National Science and Technology Major Project [Nos: 2013ZX03005012 and 2016YFC1202302, $1000], Heilongjiang Postdoctoral Fund (Grant No. LBH-Z15179, $800), and China Postdoctoral Science Foundation (Grant No. 2016 M590291, $500).

Availability of data and materials

The dataset of pre-miRNAs was downloaded from http://bioinf.sce.carleton.ca/SMIRP

About this supplement

This article has been published as part of Journal of Biomedical Semantics Volume 8 Supplement 1, 2017: Selected articles from the Biological Ontologies and Knowledge bases workshop. The full contents of the supplement are available online at https://jbiomedsem.biomedcentral.com/articles/supplements/volume-8-supplement-1.

Authors' contributions

TZ and NZ implemented the first version of the BP-Adaboost. JR, PX, ZL, LC updated the algorithm. YZ and YH wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China. [2]Department of Pharmacy, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China. [3]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China. [4]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150001, China.

Published: 20 September 2017

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.
2. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A. 2004;101(9):2999–3004.
3. Jin P, Zarnescu DC, Ceman S, Nakamoto M, Mowrey J, Jongens TA, Nelson DL, Moses K, Warren ST. Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. Nat Neurosci. 2004;7(2):113–7.
4. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432(7014):235–40.
5. Mattick JS. The functional genomics of noncoding RNA. Science. 2005; 309(5740):1527–8.
6. Michalak P. RNA world - the dark matter of evolutionary genomics. J Evol Biol. 2006;19(6):1768–74.
7. E. Berezikov, E. Cuppen, and R. H. Plasterk, "Approaches to microRNA discovery," Nat Genet, vol. 38 Suppl, pp. S2–S7, 2006.
8. Bentwich I. Prediction and validation of microRNAs and their targets. FEBS Lett. 2005;579(26):5904–10.
9. Peng J, Wang T, Hu J, Wang Y, Chen J. Constructing networks of organelle functional modules in Arabidopsis. Curr Genomics. 2016;17(5):427–38.
10. Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, Cheng L, Wang Y, Chen J. Predicting disease-related genes using integrated biomedical networks. BMC Genomics. 2017;18(1):1043.
11. Peng J, Li H, Liu Y, Juan L, Jiang Q, Wang Y, Chen J. InteGO2: a web tool for measuring and visualizing gene semantic similarities using Gene Ontology. BMC Genomics. 2016;17(Suppl 5):530.
12. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. Bioinformatics. 2016;32(8):1185–94.
13. Xue C, Li F, He T, Liu GP, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005;6:310.
14. Zhao D, Wang Y, Luo D, Shi X, Wang L, Xu D, Yu J, Liang Y. PMirP: a pre-microRNA prediction method based on structure-sequence hybrid features. Artif Intell Med. Jun, 2010;49(2):127–32.
15. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007;35(Web Server issue):W339–44.
16. Jiang L, Zhang J, Xuan P, Zou Q. BP neural network could help improve pre-miRNA identification in various species. Biomed Res Int. 2016;2016: 9565689.
17. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods. 2000;43(1):3–31.
18. Kearns M, Valiant L. Cryptographic limitations on learning Boolean formulae and finite automata. J ACM (JACM). 1994;41(1):67–95.
19. Freund, Yoav, Schapire, Robert E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer & System Sciences. 1997;55(1):119–39.
20. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics. 2004;20(17):2911–7.
21. Liu H, Wong L. Data mining tools for biological sequences. J Bioinforma Comput Biol. 2003;1(1):139–67.
22. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015;43(W1):W65–71.
23. Peace RJ, Biggar KK, Storey KB, Green JR. A framework for improving microRNA prediction in non-human genomes. Nucleic Acids Res. 2015; 43(20):gkv698.