Computers in Biology and Medicine 45 (2014) 157–160

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

miRClassify: An advanced web server for miRNA family classification and annotation

Quan Zou [a,*], Yaozong Mao [a], Lingling Hu [a], Yunfeng Wu [a], Zhiliang Ji [b,c,**]

[a] School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China
[b] State Key Laboratory of Stress Cell Biology, School of Life Sciences, Xiamen University, Xiamen 361005, China
[c] The Key Laboratory for Chemical Biology of Fujian Province, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 31 July 2013
Accepted 11 December 2013

Keywords:
Microrna family
Classification
Machine learning

ABSTRACT

MicroRNA (miRNA) family is a group of miRNAs that derive from the common ancestor. Normally, members from the same miRNA family have similar physiological functions; however, they are not always conserved in primary sequence or secondary structure. Proper family prediction from primary sequence will be helpful for accurate identification and further functional annotation of novel miRNA. Therefore, we introduced a novel machine learning-based web server, the miRClassify, which can rapidly identify miRNA from the primary sequence and classify it into a miRNA family regardless of similarity in sequence and structure. Additionally, the medical implication of the miRNA family is also provided when it is available in PubMed. The web server is accessible at the link http://datamining.xmu.edu.cn/software/MIR/home.html.

1. Introduction

MicroRNA is a kind of short RNA of about 22 nucleotides that is acknowledged to be deeply involved in the posttranscriptional regulation [18,8]. Since the first miRNA let-4 was identified in 2004, more than 21,000 matured miRNAs from 193 species have been identified and validated up to date via molecular technologies like northern blot, in situ hybridization and newly emerged next generation sequencing [12]. Besides, efficient computational tools like PlantMiRNAPred [14] have been developed to predict novel miRNAs from primary RNA sequences based on conservation in core sequence or hairpin structure against known miRNAs. Even so, it is believed that many miRNAs have not been properly identified yet due to the distant homology in sequence. Moreover, majority of the identified miRNAs are still under-studied that their target genes and impact on biological processes remain unclear.

Similar to the proteins family definition, it is acknowledged that miRNAs deriving from the identical ancestor in the phylogenetic tree can be grouped into a family. The family members intend to execute similar biological functions. For example, many miR-17-92 family members are recognized as oncogenic genes that promoted the development of lung cancers, particularly the small lung cancer [13]. This inspires us an alternative way for better identification and annotation of miRNAs. It is noticed that most of the current bioinformatics tools mainly focus on the identification of miRNAs [4], their targets [11] and distinguishing from pseudo-hairpins [17,6,14,10]. Pseudo means the sequence which can be fold as a hairpin but not be a pre-miRNA. Few attentions have been paid for miRNA family identification. Currently, there are two major sources for miRNA family classification. One is the miRBase [1] which classify miRNAs into family based on sequence similarity in the seed regions with manual adjustment. The other one is the Rfam database [2] which has almost the same family classification but with smaller dataset. According to the latest miRBase, there are 1733 miRNA families, which share the similar biological and medicine function. However, the related biological and medicine knowledge was introduced in different references and there is few such dataset collecting these related knowledge. When a novel miRNA gene is found with RT-PCR, it is not easy to guess the family information and the biological function. So it is necessary to develop a supervised method for predicting the family information for a novel miRNA sequence, and collecting the related biological and medicine references from PubMed database, which is the motivation of our web server.

The homology-based approaches are good in classification of those highly similar sequences by aligning against known miRNAs; however, they are inadequate for distantly related sequences [15]. For instance, rno-let-7b-3p (*Rattus norvegicus*) and sme-let-7d (*Schmidtea mediterranea*) are two members of let-7 family, which share no significant similarity in the entire sequence. Besides, Ding's group introduced a supervised support vector machine (SVM) model, which can effectively identify the miRNAs regardless of sequence homology [3]. However, the over-fitting problem due

* Corresponding author. Tel.: +86 13656009020.
** Corresponding author at: State Key Laboratory of Stress Cell Biology, School of Life Sciences, Xiamen University, Xiamen 361005, China. Tel./fax: +86 592 2182897.
E-mail addresses: zouquan@xmu.edu.cn (Q. Zou), appo@xmu.edu.cn (Z. Ji).

0010-4825/$ - see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.12.007

to the significant imbalance of the existing training datasets has not been properly solved. Therefore, we proposed a novel hierarchical model of random forest in this study, which can predict miRNA based on the primary sequence and properly assign it into a definite miRNA family in a cascade manner. A user-friendly interactive interface was also constructed for remote access.

2. Methods

2.1. Datasets and features

The miRNA family information for model construction is derived from the miRBase version 19.0, which includes 21,264 miRNAs in 1543 miRNA families [1]. Of these families, 69% are composed of less than five members, accounting for about 17% of the total miRNAs. Comparatively, about 8% of miRNA families contain 21–100 members per family, and 2% of families are very large families, containing more than 100 members per family. The large families (about 10% of total families) together cover about 65% of miRNAs.

For a given RNA sequence, *k*-grams are first extracted as features [3]. *k*-grams stand for *k* consecutive nucleotides. The higher the *k* is, the more features there are. For instance, $k=3$ is referred to as a "trigram", which has $4^3=64$ features. To choose an optimal *k* for the model, a feature comparison study was undertaken by adopting different *k* ranging from 3 to 6. As well, the 32-dimensional features often used in previous miRNA classification were included [17]. The prediction accuracies of the RF models constructed upon these features were evaluated by conventional 10-fold cross validation, and the results are given in Table 1. Generally, the 6-g features have better performance than other feature sets. Although it is a bit time-consuming than 3–5-g feature sets, it is still efficient than the 32-dimensional features. Therefore, considering the balance of execution speed and the prediction accuracy, *k* is set as 6 in this study.

2.2. Model construction

To deal with the significant unbalance in the dataset and the robustness problem that it may induce, we developed a hierarchical random forest (RF) classifier model to categorize the novel miRNA sequences into families. Random forest is a classifying technique of ensemble learning. It is called as "forest" because that it consists of several decision trees. The two major ideas of random forest are bagging and random feature selection. In bagging, classifiers are trained on a bootstrap training data and the prediction is voted by these classifiers. Random forest selects some of the features randomly to split at each node when constructing the decision trees. Each tree in the forest is constructed to the largest extent possible without any pruning. Training samples are selected randomly and replaced from the original training set and cross-validation is used for assessing the performance of each tree [6]. In our work, 100 decision trees were employed to construct the random forest. And the probability of prediction is the rates of the decision trees which are major in the voting and predicting the final result.

As illustrated in Fig. 1, there are two core components in the models. One is the pre-miRNA predictor. It will distinguish the miRNAs from the pseudo-hairpins. The other is a three-layer classifier for miRNA family classification. Each of these three layers contains an independent RF classifier. The first layer is composed of 20 classes, corresponding to the top 19 largest miRNA families from miRBase and one integrated from the remaining families. The second layer is composed of 100 classes, corresponding to 99 largest miRNA families, excluding the 19 miRNA families in the first layer, and one integrated from the remaining families. In the third layer, all remaining miRNA families are incorporated for model construction. We have counted all the families members. There are 22 families whose members are more than 100. And there are 123 families whose members are more than 20. So we choose 20 nodes in the first layer, since the former 20 families are more large than others. We choose 100 nodes in the second layer, since they all consist more than 20 members. For a query sequence, a pre-miRNA prediction is normally demonstrated before further miRNA family classification. The putative pre-miRNA sequence will then go through the three-layer classifier one by one until it is properly assigned into a miRNA family. In this way, the unbalanced dataset problem is, to some extent, solved and the method becomes more robust.

Table 1
Model performance on random forest by adopting different feature sets, including *N*-gram and Xue's 32-dimensional features.

Features	Overall accuracy		
	First layer	Second layer	Third layer
$N=3$	90.84	77.10	60.21
$N=4$	93.19	82.64	67.23
$N=5$	93.21	84.72	68.69
$N=6$	95.14	85.56	69.59
Xue's	90.20	75.05	51.98

2.3. Evaluation of the model

The model was evaluated by conventional 10-fold cross validation for each of three-layer classifier. The overall accuracies of 95.71%, 91.43% and 72.74% were achieved for the first layer RF classifier to the third one respectively. The performance of the model was also compared with several different classifiers adopting 6-g features, including the random tree (RT), the decision tree (DT), the support vector machine (SVM) and the nearest neighbor (NN). Boosting, which is a popular ensemble learning, is also compared. The results are given in Table 2. It seems that the RF model shows comparatively best performance than other algorithms in this multi-class classification. Boosting can improve the performance; however, it is very time consuming. Since RF can work as well as Boosting, we employ RF as our classifier in the web server.

3. Description of the server

A user-friendly web server was constructed on a Linux workstation with Apache-Tomcat-7.0.27, and developed for the biomedical application. Users can input the putative novel pre-miRNA sequences in FASTA format. The program first predicts whether the sequence is a pre-miRNA molecule. If it is, the web server predicts the family information via three hierarchical classifiers. The sizes of the training models are 12 M, 80 M, and 1.28 G for these classifiers. Running a program with more than 1 G of model data is difficult for common computers. Thus, we employed a super workstation with 144 G memory for our web server, and users can retrieve their results within a few seconds. When searching for the "detail" in "Disease relationship", the related literature that was mined from PubMed is listed for further biomedical research [16]. The miRNA family name is searched in the titles and abstracts from PubMed. If there is a disease name appearing with the miRNA family name, we extract the relationship from the abstract or the title. The miRNA family–disease relationship dataset is the fusion of Jiang's [7] and Lu's work [9].

4. Conclusion

Although miRNA mature part can be easily obtained in the next generation sequencing (NGS) experiments and miRBase families

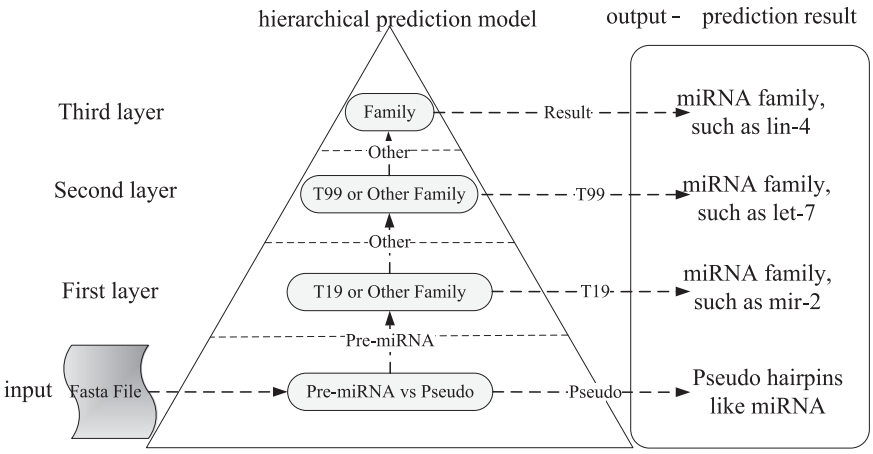

Fig. 1. Model of the hierarchical random forest classifier.

Table 2
Performance comparison of different classifiers with 6-gram features.

Classifiers*	Acc of first layer	Acc of second layer	Acc of third layer
Random tree	90.92	75.06	47.04
Random forest	95.14	85.56	69.59
Decision tree	95.11	59.20	61.84
Support vector machine (SVM)	85.88	61.41	55.59
Nearest neighbor	93.85	89.63	69.74
Boost+random forest	95.36	87.02	72.10
Boost+SVM	84.54	60.14	26.63

are explained by multiple alignments, which are steered by mature sequence segments instead of precursors, miRClassify is still biologically meaningful since novel miRNAs are usually found by RT-PCR or northern blot experiments [5]. Researchers can find the expression of precursor sequence, but it is not easy to make sure the mature part. So it is helpful and useful to predict the family and function from the precursor sequences.

So far as we know, the miRClassify is the only on-line server for miRNA family classification. It can not only predict novel miRNA from the primary sequence regardless of sequence or structure homology, but also provide information of family classification and medical implications which has not been done in previous studies. The server will facilitate better understanding of miRNAs and their impact on nowadays medicine.

Our server can deal with the pre-miRNA candidates so far. It cannot extract the miRNA or miRNA clusters from long DNA segment. Moreover, computationally finding the mature part from the precursor is still an open problem. These will be the future work for improving miRClassify.

Conflict of interest statement

There is no conflict of interest.

Acknowledgments

The work was supported by the Natural Science Foundation of China (Nos. 61370010, 31271405, 81101115), the Natural Science Foundation of Fujian Province of China (Nos. 2011J01371, 2011J05158) and the 2013 Program for New Century Excellent Talents in Fujian Province University.

References

[1] Ana Kozomara, et al., miRBase: integrating microRNA annotation and deep-sequencing data, Nucl. Acids Res. 39 (2011) D152–D157.
[2] W.S. Burge, et al., Rfam 11.0: 10 years of RNA families, Nucl. Acids Res. 41 (2013) D226–D232.
[3] Jiandong Ding, et al., miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM, BMC Bioinf. 12 (2011) 216.
[4] Lingling Hu, et al., Benchmark comparison of ab initio microRNA identification methods and software, Genet. Mol. Res. 11 (2012) 4525–4538.
[5] Yong Huang, et al., Computational identification and characteristics of novel microRNAs from the silkworm (*Bombyx mori* L.), Mol. Biol. Rep. 37 (2010) 3171–3176.
[6] Peng Jiang, et al., MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combine features, Nucl. Acids Res. 35 (2007) W339–W344.
[7] Qinghua Jiang, et al., miR2Disease: a manually curated database for microRNA deregulation in human disease, Nucl. Acids Res. 37 (2009) D98–D104.
[8] Bing Liu, et al., Identifying miRNAs targets and functions, Briefings Bioinf. (2012), http://dx.doi.org/10.1093/bib/bbs075.
[9] Ming Lu, et al., An analysis of human microRNA and disease associations, PLoS One 3 (2008) e3420.
[10] Kwang Loong Stanley Ng, Santosh K. Mishra, De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures, Bioinformatics 23 (2007) 1321–1330.
[11] Leyi Wei, et al., Computational analysis of miRNA target identification, Curr. Bioinf. 7 (2012) 512–525.
[12] Vernell Williamson, et al., Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation, Briefings Bioinf. 14 (2012) 36–45.
[13] Changchun Xiao, et al., Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes, Nat. Immunol. 9 (2008) 405–414.
[14] Ping Xuan, et al., PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs, Bioinformatics 27 (2011) 1368–1376.
[15] Ping Xuan, et al., MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs, PLoS One 6 (2011) e27422.
[16] Ping Xuan, et al., Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors, PLoS One 8 (2013) e70204.
[17] Chenghai Xue, et al., Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinf. 6 (2005) 310.
[18] Dong Yue, et al., Understanding microRNA regulation: a computational perspective, IEEE Signal Process. Mag. 29 (2012) 77–88.

Quan Zou born in 1982. He received the Bachelor degree in Computer Science and Economics at the same time in 2004, and received Ph.D. Degree in 2009 from the School of Computer Science and Technology in Harbin Institute of Technology. He is an Assistant Professor in the Department of Computer Science in Xiamen University now. He is a Member of IEEE, ACM, and CCF.

Yaozong Mao received the Bachelor degree in 2012 from Tianjin University of Commerce, China. At present he is a graduate student in Department of Computer Science at Xiamen University. His research interests include bioinformatics and CUDA computing.

Lingling Hu received the Bachelor degree in 2011 from Henan University, China. At present she is a graduate student in Department of Computer Science at Xiamen University. Her research interests include microRNA identification and classification.

Yunfeng Wu received the B.E. and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China, in 2003 and 2008, respectively. He worked as a Post-Doctoral Fellow at the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada. He is currently an Associate Professor at the School of Information Science and Technology, Xiamen University, Xiamen, Fujian Province, China. His research interests include biomedical signal processing, pattern recognition, and neural engineering. He is a Senior Member of IEEE, and Member of BMES, INNS, and CCF.

Zhiliang Ji received the B.E. degree from Tsinghua University in China, and M.S. and Ph.D. degrees from National University of Singapore, Singapore, in 2000 and 2004, respectively. Now he is a Professor in Xiamen University. His research interests include system biology and bioinformatics.