Computational Biology and Chemistry xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Predicting microRNA biological functions based on genes discriminant analysis

Tao Ding[a,b], Junhua Xu[a], Mengmeng Sun[a], Shanshan Zhu[a], Jie Gao[a,*]

[a] School of Science, Jiangnan University, Wuxi, China
[b] School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne, UK

ARTICLE INFO

Article history:
Received 7 September 2017
Accepted 25 September 2017
Available online xxx

Keywords:
microRNA
Family classification
Genes discriminant analysis
10-fold cross-validation

ABSTRACT

Although thousands of microRNAs (miRNAs) have been identified in recent experimental efforts, it remains a challenge to explore their specific biological functions through molecular biological experiments. Since those members from same family share same or similar biological functions, classifying new miRNAs into their corresponding families will be helpful for their further functional analysis. In this study, we initially built a vector space by characterizing the features from miRNA sequences and structures according to their miRBase family organizations. Then we further assigned miRNAs into its specific miRNA families by developing a novel genes discriminant analysis (*GDA*) approach in this study. As can be seen from the results of new families from *GDA*, in each of these new families, there was a high degree of similarity among all members of nucleotide sequences. At the same time, we employed 10-fold cross-validation machine learning to achieve the accuracy rates of 68.68%, 80.74%, and 83.65% respectively for the original miRNA families with no less than two, three, and four members. The encouraging results suggested that the proposed *GDA* could not only provide a support in identifying new miRNAs' families, but also contributing to predicting their biological functions.

1. Introduction

MicroRNAs (miRNAs) are one kind of endogenous non-coding RNAs (ncRNAs) with the length of 20–25 nucleotides. They can bind to the 3′ untranslated regions (UTRs) and suppress the expression of their target messenger RNAs (mRNAs) at post-transcriptional level through sequence-specific base pairing. Accumulated studies have shown that miRNAs play critical roles in significant signalling pathways, biological processes, and pathophysiologies (Bartel, 2004; Ambros, 2001, 2004; Meister and Tuschi, 2004; Backes et al., 2017). Up to now, several biological experiments for identifying the biological functions of miRNAs have made advances by direct cloning (Liu et al., 2009), forward and reverse genetics (Gurumurthy et al., 2016), and so on. However, these methods would be a high-cost when applied to large-scale miRNAs. In fact, a striking feature of these miRNAs is that their loci are usually clustered in the genome (Subramanian et al., 2005,). More in-depth studies, including miRNA co-expression and primary transcript identification, suggest that the majority of miRNA clusters are transcribed as a single unit and share common biological functions (Baskerville and Bartel, 2005; Rhee et al., 2013). Meanwhile, miRNA's families can be viewed as a special case of gene families or genes sets (Li et al., 2015), and miRNA families are highly conserved in nucleotide sequences and secondary structures among closely related species over evolutionary time (Cuperus et al., 2011; Kaczkowski et al., 2008; Kamanu et al., 2013; Hertel and Stadler, 2015). Therefore, given an existing miRNA family system (the miRBase (Griffiths-Jones et al., 2007) family organization), we proposed a bioinformatics method to accurately and efficiently identify miRNAs' families for the sake of their biological function prediction. We reviewed literature in the following sections for the convenience of descriptions.

By clustering miRNAs, Sanger et al. (Griffiths-Jones et al., 2003) constructed an RNA family database: Rfam, which was a collection of multiple sequence alignments and covariance models representing non-coding RNA families. Ding et al. (Ding et al., 2011) employed *N-grams* to extract features from precursor miRNA sequences, then trained a multiclass *SVM* classifier to classify new miRNAs. Based on the advanced machine learning-based web server, Zou et al. (Zou et al., 2014) proposed a novel hierarchical model of random forests in the study (miRClassify). This model could predict miRNAs according to its primary miRNA sequence

* Corresponding author.
E-mail addresses: 1213834817@qq.com (T. Ding), 1447067414@qq.com (J. Xu), sunmeng926@126.com (M. Sun), 371980576@qq.com (S. Zhu), gaojie@jiangnan.edu.cn (J. Gao).

https://doi.org/10.1016/j.compbiolchem.2017.09.008
1476-9271/

and properly assigned it into a specific miRNA family in a cascade manner.

Most of traditional miRNA family research approaches (Griffiths-Jones et al., 2003; Ding et al., 2011; Zou et al., 2014) started with constructing feature vectors by primary or precursor miRNAs, and then used different algorithms to establish the miRNA family system. However, with the enormous miRNAs across all species, these methods would require a large amount of computing resources and time. For instance, although miRFam had a decent performance on classifying genes, the dimensions of feature vector that the model constructed had reached to 340. It would be time-consuming when applied to large-scale validation miRNAs.

This paper proposed a novel effective classification approach named Genes Discriminant Analysis (*GDA*) to assign new miRNAs to their corresponding families. The basic idea of *GDA* is similar to Mahalanobis distance discriminant analysis in multivariate statistical analysis (Pemajayantha et al., 2003). So, the proposed *GDA* algorithm does not have very complicated or expensive computing components due to using an existing miRNA family system. As such, *GDA* would have a good time complexity that allowed it to run faster. Furthermore, we took into consideration the highly conserved sequences and structures of miRNAs, and constructed a 116-dimensional vector to improve the model performance. In those, the proposed model would contribute to accurately identifying miRNAs' family and annotating new miRNAs' biological functions.

2. Methods

MiRNA genes produce three major RNA products: primary miRNAs (pri-miRNAs), precursor miRNAs (pre-miRNAs), and mature miRNAs. The pri-miRNA, pre-miRNA, and mature miRNA each contain identical seed sequences and all have the potential to interact with target mRNAs (Kim, 2005). Besides, the miRNA families are highly conserved in nucleotide sequences and secondary structures (Kaczkowski et al., 2008; Lee et al., 2007). Taken together, we initially built a vector space by characterizing features of miRNAs from their mature sequences and secondary structures. And then new miRNAs would be classified properly and rapidly by the novel approach *GDA* in this study.

2.1. Data sources

MiRBase is the data source used in this study. It is the central online repository of miRNA nomenclature, sequence data, annotation, and target prediction (Griffiths-Jones et al., 2007). From version 5.0, miRBase begins to classify miRNAs into different families. In the past ten years, the number of miRNA genes' registers in miRBase increased rapidly with the development of experimental (Lim et al., 2003; Grad et al., 2003) and computational miRNAs detecting methods (Ng and Mishra, 2007; Vander Burgt et al., 2009). We explored the miRBase from version 8.2 (July 2006) to version 21 (June 2014) and illustrated the changes of human miRNAs and gene families during the last several years in Table 1.

TargetScan, an algorithm for predicting miRNA targets, refers to bases 2–8 of mature miRNA segment (numbered from the 5′ end of miRNA) as the "miRNA seed regions" in 2004 (Lewis et al., 2003). The study also found that the conserved 5′ region of miRNA is extremely important for target identification and keeps highly conserved in the evolutionary process. Because of that, we integrated the 1700 mature miRNAs and 579 miRNA families of Homo sapiens in miRBase as experimental sets. In this paper, 860 unclassified miRNAs would be assigned into these 579 families through *GDA*.

Table 1
miRNA family numbers from miRBase version 8.2 to version 21.

	miRNAs	miRNA Families	Unclassified miRNAs
Version: 8.2	462	342	114
Version: 21	1767	579	860
Growth Multiple	4.1	1.7	7.5

2.2. Feature vectors

For the given mature miRNA sequences, they only contain four bases *A*,*U*, *C*, and *G*. We introduced a method like *N-Grams* (Liu et al., 2014) (*N-Grams* also known as *k-mers* or *l-tuples*, stand for *N* consecutive nucleotides) to transform miRNA sequences into numeric vectors. We selected $N = 1, 2, 3$ to refer it as "unigram-base", "bigram-bases", and "trigram-bases", respectively. By converting a mature miRNA sequence to a set of "*N-Grams*", it can be embedded into an 84-dimensional (4 + 16 + 64) vectors space. All these elements of feature vectors are calculated as follows,

$$x_i' = \frac{n_i}{n} \times 100\%\ \ i = o,\ op,\ opq \text{ and } o,\ p,\ q \in \{A,\ C,\ G,\ U\} \quad (1)$$

where x'_i is the occurrence frequency rate of the content of unigram-base, bigram-bases, and trigram-bases. n_i is the frequency of each bases group. n is the total frequencies of all bases group. The sum of all dimensional values in the feature vector x' is 1.

$$x' = (x_A,\ \cdots,\ x_U\ ,\ x_{AA},\ \cdots,\ x_{UU},\ x_{AAA},\ \cdots,\ x_{UUU}) \quad (2)$$

For the given precursor miRNA sequences, secondary structures of all candidate miRNAs are predicted by RNAfold using minimum free energy (MFE) rule (Hofacker, 2003). Eventually, we got the characteristic sequences by the popular representation-bracket notation (Bose et al., 2015), where dots "." are placed under the unpaired bases, and brackets "(" are placed under the paired bases.

For any 3 adjacent nucleotides, there are 8 possible structural compositions: "(((", "((.", "(.", "(.(", ".((", ".(.", ".(", and " . . . ". Considering the middle nucleotide, there are 32 (4×8) possible structure-sequence combinations, and they are denoted as "A(((", "C((.", etc. The number of each triplet element appeared is counted for each precursor sequences to produce the 32-dimensional feature vector. The sum of all dimensional values in the feature vector x″ is 1.

$$x'' = (x_{A\ldots}, x_{A..(}, \cdots, x_{U((.}, x_{U(((}) \quad (3)$$

Therefore, any miRNA can be converted into a 116-dimensional (84 + 32) vectors according to its mature sequence and secondary structure.

$$x = (x', x'') \quad (4)$$

For example, Fig. 1 illustrated the characteristic sequence of hsa-mir-1973 was as follows.

2.3. Genes discriminant analysis

For each new miRNA $x = (x_1, x_2, \cdots, x_{116})^T$, it would be assigned into one of 579 miRNA families. We constructed classification functions for each family. It was defined as:

$$d^2(x, \pi_i) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i),\ \ i = 1,\ 2,\ \ \ldots,\ 579 \quad (5)$$

In this equation, μ_i was the population mean of miRNA family π_i, and Σ_i^{-1} was the inverse covariance matrix in 116*116 of the family π_i. Supposed that a sample x_i was obtained from family π_i, where

Please cite this article in press as: T. Ding, et al., Predicting microRNA biological functions based on genes discriminant analysis, Comput. Biol. Chem. (2017), https://doi.org/10.1016/j.compbiolchem.2017.09.008