Genomics 98 (2011) 73–78

Contents lists available at ScienceDirect

Genomics

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

Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM

Yanqiu Wang [a,1], Xiaowen Chen [a,1], Wei Jiang [a,1], Li Li [b], Wei Li [a], Lei Yang [a], Mingzhi Liao [a], Baofeng Lian [a], Yingli Lv [a], Shiyuan Wang [a], Shuyuan Wang [a], Xia Li [a,*]

[a] College of Bioinformatics Science and Technology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province, Harbin Medical University, Harbin 150081, PR China
[b] Key Laboratory of Arrhythmias, Ministry of Education, School of Medicine, Tongji University, Shanghai 200092, PR China

ARTICLE INFO

Article history:
Received 19 October 2010
Accepted 29 April 2011
Available online 7 May 2011

Keywords:
Human microRNA precursors
Classification
Feature selection
Support vector machine
Genetic algorithm

ABSTRACT

MicroRNAs (miRNAs) are non-coding RNAs that play important roles in post-transcriptional regulation. Identification of miRNAs is crucial to understanding their biological mechanism. Recently, machine-learning approaches have been employed to predict miRNA precursors (pre-miRNAs). However, features used are divergent and consequently induce different performance. Thus, feature selection is critical for pre-miRNA prediction. We generated an optimized feature subset including 13 features using a hybrid of genetic algorithm and support vector machine (GA–SVM). Based on SVM, the classification performance of the optimized feature subset is much higher than that of the two feature sets used in microPred and miPred by five-fold cross-validation. Finally, we constructed the classifier miR-SF to predict the most recently identified human pre-miRNAs in miRBase (version 16). Compared with microPred and miPred, miR-SF achieved much higher classification performance. Accuracies were 93.97%, 86.21% and 64.66% for miR-SF, microPred and miPred, respectively. Thus, miR-SF is effective for identifying pre-miRNAs.

1. Introduction

MicroRNAs (miRNAs) are a family of ~22nt endogenous non-coding RNAs involving in post-transcriptional regulation [1]. Mature miRNAs are usually cleaved from ~90nt miRNA precursors (pre-miRNAs), characterized by a stem-loop or stem-loop-like structure that can be used as a characteristic for identifying novel miRNAs. The first miRNA (lin-4) is discovered in 1993 [1,2]; currently 15,172 miRNAs in 142 species are in the latest version of miRBase (version 16), including 1048 human miRNAs [3]. Recent estimates suggest that there is a large number of undiscovered miRNAs in many species [4,5].

Currently, computational prediction and experimental approaches have been used to discover novel miRNAs. A cDNA cloning technique is frequently used in experimental approaches. Although the cDNA cloning is direct and reliable, capturing miRNAs with low-expression levels or miRNAs that are expressed in a time-specific or tissue-specific manner is difficult. In recent years, computational prediction has been used to identify potential pre-miRNAs, since it is not affected by time or tissue specificity of miRNA expression. In particular, machine learning approaches including support vector machine (SVM) [6–9], random forest (RF) [10], hidden Markov model (HMM) [11–13] and naive Bayes classifier (NBC) [14] have been used. Sequence compositions and RNA folding measures of secondary structure have been used as inputting features in these approaches. The divergent features used in these approaches result in different outcomes. Hence, selecting effective feature subset is very important for identifying new pre-miRNAs.

In miRabela [15], 40 distinctive sequence and structural features from the hairpins are employed to identify pre-miRNAs in the genomic regions near the known mammalian miRNAs. This method predicted about 50 to 100 novel pre-miRNAs for several species; about 30% of potential pre-miRNAs predicted were experimentally validated. miPred, which is an SVM-based classifier, was constructed using 29 “global and intrinsic” features from hairpin folding characteristics, and is used to predict human pre-miRNAs. Its accuracy in test set reaches 93.50% for human [6]. A recent microPred approach used 21 features related to sequence composition and thermodynamic stability to distinguish human pre-miRNAs from pseudo hairpins. MicroPred based on SVM achieves high classification results for both sensitivity (SE) (90.02%) and specificity (SP) (97.28%) [8].

Although these approaches achieve satisfactory performance in several species, they have limitations. MiRabela and miPred consider the importance of the stem-loop structure characteristics for identifying real pre-miRNAs, but do not filter out features in the feature set that may lead to poor classification performance. For microPred, only fine features from the initial feature set remain after a

* Corresponding author at: College of Bioinformatics Science and Technology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province, Harbin Medical University, Harbin 150081, PR China. Fax: +86 451 8661 5922.
E-mail address: lixia@hrbmu.edu.cn (X. Li).
[1] Equal contribution.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.04.011

filter feature selection. However, the *filter* model requires no feedback from the classifier and estimates the classification performance by indirect assessments such as distance measures, which reflect how well the classes separated from each other. Although feature selection in microPred improves classification accuracy, feature selection using the *filter* model does not provide much higher classification accuracy. By contrast, the *wrapper* feature selection methods are classifier dependent, in which the "goodness" of the selected feature subset is evaluated directly by classification accuracy. Based on previous studies, the classification accuracy of the *wrapper* model is higher than the *filter* model in feature selection [16–19].

In this study, to find a more effective feature subset for classification, we firstly extract as many characteristics of pre-miRNAs as possible from the literatures [6,8,15]. Then, a *wrapper* feature selection that evaluated if the features were useful or not, a hybrid of genetic algorithm and support vector machine (GA–SVM), was employed to identify an optimized feature subset. Based on SVM, the classification performance of the optimized feature subset is much higher than that of the initial features without feature selection and the two feature sets used in microPred and miPred by using five-fold cross-validation. The SVM-based classifier miR-SF is constructed for predicting the most recently identified human pre-miRNAs in miRBase (version 16) by using the optimized feature subset. Compared with the two SVM-based microPred and miPred, miR-SF (93.97%) achieves much higher classification accuracy than microPred (86.21%) and miPred (64.66%).

2. Materials and methods

2.1. Data set

The sequences of human miRNA precursors were downloaded from release 15 of miRBase registry database, which included 940 human pre-miRNA entries. The secondary structure of all sequences used in this study was predicted using the RNAfold procedure in the Vienna RNA package version 1.8.1 [20]. After removing the pre-miRNAs with multiple loops in the stem-loop structure, 906 pre-miRNAs were obtained, which composed the positive samples. 657 of those were contained in release 12 of miRBase and were used to generate the optimized feature subset, while the remaining 249 pre-miRNAs were used to evaluate the optimized feature subset.

As negative samples, 8494 pseudo hairpins were extracted from protein coding regions according to the RefSeq and UCSC refGene annotations. These were also used in microPred [8] and miPred [6]. To distinguish miRNAs from other small RNAs, 754 other small RNAs used in microPred [8,21] were obtained, of which 129 small RNAs without multiple loops were also taken as negative samples.

2.2. Feature set

Constructing an initial feature set was very important for identifying the optimized feature subset. In order to extract an effective classification feature subset, we selected as many features of pre-miRNAs as possible based on the literatures. We obtained 185 features for the original feature set, characterized by sequence composition and RNA folding measures of secondary structure. Sequence features mainly included the frequency of single nucleotides, dinucleotides and trinucleotides in the pre-miRNA sequences, while secondary structure features included adjusted Shannon entropy, distance between internal loops and the frequency of the minimum free energy structure, etc. All features are described in detail in the supplementary material.

All feature values were extracted by analyzing pre-miRNA primary and secondary structures. Fig. 1 shows the primary and secondary structure of the hsa-let-7e precursor and the locations of some terms in the secondary structure. The secondary structure of all pre-miRNA sequences was predicted using RNAfold under default parameters. After extracting all feature values, we found that "the number of symmetrical loops with each part exactly containing seven bases" and "the number of asymmetrical loops with the longest part exactly containing seven bases" were zero for all sequences used. These two features were removed, and the remaining 183 features formed the initial feature set.

2.3. Identification of an optimized feature subset

GA–SVM was proposed for selecting an optimal feature gene subset for disease classification in our previous study [22]. This is a hybrid of genetic algorithm (GA) and support vector machine (SVM) that fully utilizes the unique merits of the two data-mining approaches. Here, for each sample set, GA–SVM was employed to extract an optimal feature subset for distinguishing real pre-miRNAs from pseudo hairpins and other small RNAs. GA was used to extract the optimal feature subset based on the process of nature selection, in which the fitness value of an individual (feature subset) is given based on the classification accuracy of the SVM classifiers. First, we randomly generated *N* fixed-length individuals (the set of the feature subsets) as an initial population, encoded as *N* binary strings for easier operation. Based on the corresponding feature value submatrix for each individual, SVM classifiers were constructed, and the average classification accuracy based on five-fold cross-validation was denoted as the fitness value of this individual. Individuals with higher fitness values had a greater chance to be selected to generate new feature subsets than those with low values by crossover and mutation. Better individuals were retained by survival of the fittest, and finally

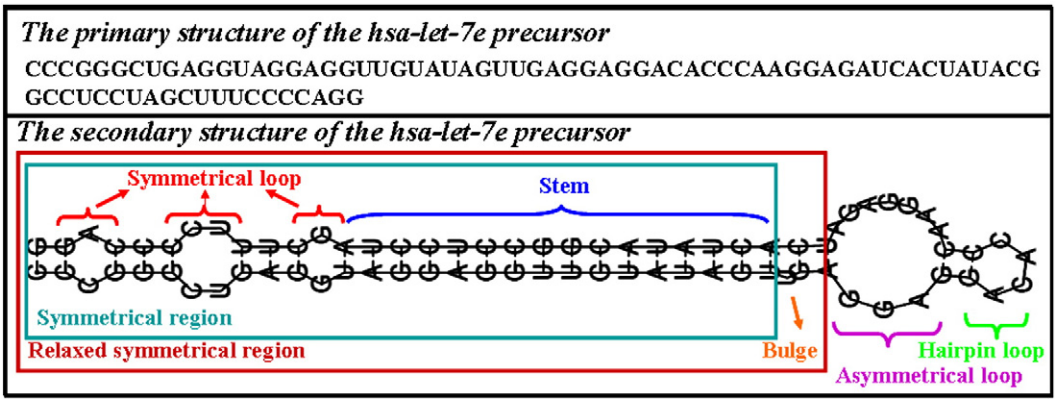

Fig. 1. The primary and secondary structure of the hsa-let-7e precursor and the locations of some terms in the secondary structure. The upper part shows the primary structure of hsa-let-7e and the lower one indicates the secondary structure and the relevant terms with different colors.

the optimal individual was obtained as an optimal feature subset when the stop condition was met. In this study, for a certain sample set, GA–SVM with default parameters was employed to extract an optimal feature subset for distinguishing two classes, using the following steps:

(1) *Encoding*. We constructed the feature value matrix M^K and the initial feature set with n features for a certain sample set. M^K consisted of the feature values of the n ($n=183$, when $K=0$) features of all 1314 samples in this sample set. Next, N fixed-length binary strings (individuals) were randomly generated to form the initial population D_0^K, which contained N individuals (we set $N=40$). Each binary string represented a feature subset, and the value of each position in the binary string was encoded as either 1 or 0; with 1 representing presence of the particular feature in the subset, and 0 representing absence. To avoid the loss of important features, in the initial population, about half of the features in each individual were retained, so the number of features in the feature subsets was large in early generations. For example, M^0 had 183 features, while the number of features in each individual of D_0^0 is approximately 92.

(2) *Evaluating feature subsets*. We employed SVM to construct the classifiers for each individual (feature subset) in the initial population. For each individual, the feature value submatrix $SM^K{}_j$ ($j=1,2,\ldots,N$) including only the features in this individual was obtained from the original matrix M^K. Based on the submatrix $SM^K{}_j$, the adaptability of the individual j was evaluated by the average accuracy of the classifiers based on five-fold cross validation, which was denoted as $eval_j = eval(SM_j^K)$ for the individual j. Where the fitness value was $eval_j = \left(\sum_{d=1}^{25} ACC_d\right)/25$, d was the number of sample sets generated by one five-fold cross validation, ACC_d was the classification accuracy of the test samples T_d in the dth sample set set_d (detailed description was shown in five-fold cross validation in methods) using the SVM algorithm. Discriminant function in SVM was given by

$$\widehat{y} = f(x) = sgn\left\{\sum_{i=1}^{L} a_i y_i K(x_i \cdot x) - b\right\} \quad (1)$$

where x was the test sample vector, x_i was the training sample vector, L was the number of test samples, y_i was label of samples, a_i was the Lagrange multiplier related with x_i. For SVM, $a_i \neq 0$, sgn{} was symbol function, and $K(x_i \cdot x)$ was a kernel function (linear kernel was used in this study). The proportion of samples correctly classified in the test samples was obtained as the classification accuracy (ACC).

(3) *Producing new population by crossover and mutation*. Since the initial set of feature subsets (initial population) is not, in general, the whole subsets of the feature set, but is randomly generated, the new set of the feature subsets (new generation population) was generated by crossover and mutation to obtain the optimal feature subset. Based on the fitness value of each individual in the initial population D_0^K, the highest $N/2$ individuals were obtained to form the set of the feature subsets denoted D_{01}^K, which did not undergo crossover and mutation and directly entered the next generation. Using the choice probability $P_j = (eval_j)/\sum_{j=1}^{N} eval_j$, we randomly selected individuals in the initial population D_0^K to generate another $N/2$ new individuals denoted D_{02}^K by crossover and mutation. Here the higher the fitness value of an individual j was, the more the selected probability was for that individual. For the initial population D_0^K, based on crossover probability 0.6, we extracted two individuals at random to perform a single crossover four times, generating eight individuals. Two individuals with the highest fitness values of the eight individuals were extracted as the members of D_{02}^K. Mutation was employed to change the values of some positions (adding and deleting features) in randomly selected individuals in D_0^K based on a mutation probability of 0.05. Four individuals were selected in each batch, and the individual with the highest fitness value was added to D_{02}^K. Finally, the new population D_1^K was formed by combining D_{01}^K with D_{02}^K.

(4) *Extracting an optimal feature subset*. The best individual (feature subset) with the highest $eval_j$ in the current generation D_1^K was obtained by calculating the fitness value of all individuals. When the classification accuracy difference of the two best individuals in neighboring two generations was less than 0.001, or the maximum generation reached 100, iteration was stopped, and the last best individual was extracted as the best feature subset in the current set of feature subsets in which each feature number was approximately $n/2$. To further select the optimal feature subset, the initial feature set was replaced with the above extracted best feature subset, the initial feature value matrix M^K was replaced with the corresponding submatrix SM^K (in which the column of the feature value matrix was all features of the new feature subset) of this new feature subset, and the above steps were repeated until the last two feature subsets extracted were the same. This feature subset was extracted as the optimal feature subset of this sample set.

2.4. Measures of evaluating the optimized feature subsets

Based on the optimized feature subset, we constructed the SVM classifier for evaluating the classification power of this feature subset. Five-fold cross validation and the five indices including sensitivity (SE), specificity (SP), accuracy (ACC), F-measure (Fm) and Matthews correlation coefficient (MCC) were used to evaluate the classification performance of the optimized feature subset. The formulas of these indices were:

$$\mathrm{SE} = TP/(TP + FN) \times 100\% \quad (2)$$

$$SP = TN/(TN + FP) \times 100\% \quad (3)$$

$$\mathrm{ACC} = (TP + TN)/(TN + FP + TP + FN) \times 100\% \quad (4)$$

$$Fm = 2TP/(2TP + FP + FN) \times 100\% \quad (5)$$

$$\mathrm{MCC} = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)} \quad (6)$$

where TP was the number of real pre-miRNAs correctly identified, FN was the number of real pre-miRNAs missed; TN was the number of pseudo hairpins correctly identified, and FP was the number of pseudo hairpins incorrectly classified.

2.5. Five-fold cross validation

For the optimized feature subset, we used five-fold cross-validation to evaluate its classification performance. First, positive and negative samples in a sample set were randomly divided into five non-overlapping parts of roughly equal size, denoted as P_i ($i=1, 2, \ldots, 5$) for positive samples (real pre-miRNAs) and N_i ($i=1, 2, \ldots, 5$) for negative samples. A combination of P_i and N_i was used as the test set, and the rest of the sample set was used as the training set. Thus, all combinations produced 25 pairs of training and test sets, $set_d = \{L_d, T_d\}$ ($d=1, 2, \ldots, 25$). Here, L_d was training

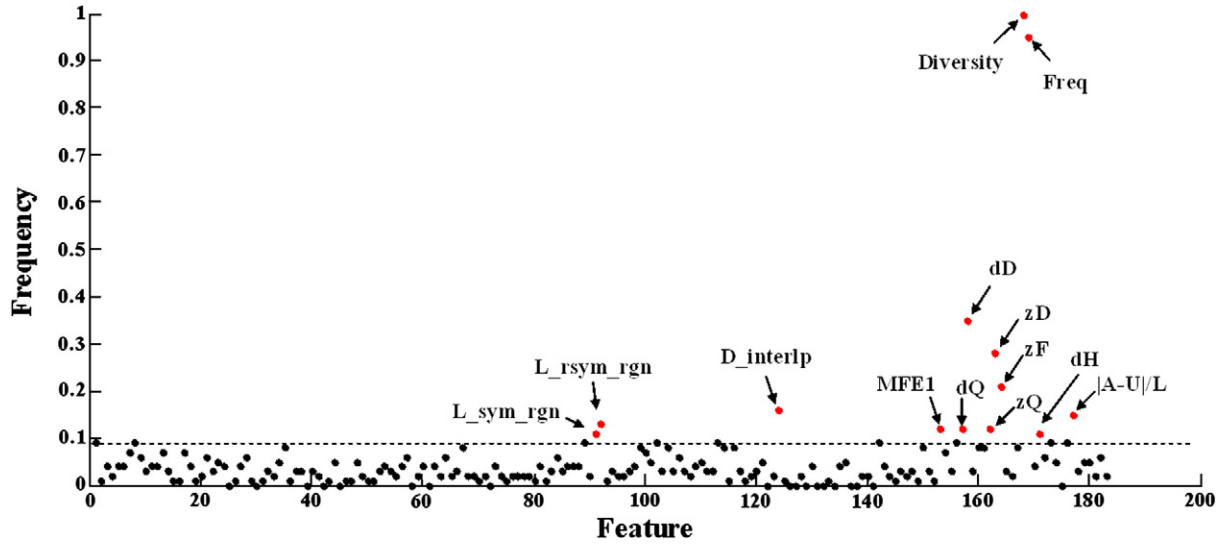

Fig. 2. Frequency of 183 features in training sets. The *x*-axis is 183 features, and the *y*-axis is the frequency of features on training sets. The highest feature frequency on 1000 random datasets of 0.09 is denoted with a dotted line. The frequencies of 13 features on training sets are denoted with red solid circles, and the frequencies of other features are denoted with black solid circles. Meanwhile, 13 feature names are given by this figure.

set, while T_d was test set. Then, a set of 25 classifiers was constructed based on 25 different training sets L_d, and the values of the five classification measures for the corresponding test set T_d were calculated. The average values of each of five measures were considered as the final output.

3. Results

Using the GA–SVM algorithm we identified an optimized feature subset for distinguishing real pre-miRNAs from non-pre-miRNAs. Then, the classification performance of the optimized feature subset was compared with the initial features and the two previous feature sets by using five-fold cross-validation based on SVM. Finally, the miR-SF classifier was constructed by using this optimized feature subset. The classification performance of miR-SF was compared with two previous approaches (microPred and miPred) by using the identification power of the most recently identified miRNAs in miRBase (version 16).

3.1. Optimized feature subset extracted

We extracted 657 pre-miRNAs (positive samples) from release 12 of miRBase after removing the pre-miRNAs with multiple loops in the stem-loop structure. The number of the positive samples was significantly less than the negative samples, so we established an unbiased classifier by selecting 657 negative samples, including 129 other small RNAs and randomly selected 528 negative samples from 8494 extracted protein-coding sequences. This was repeated 100 times to form 100 different negative sample sets. Thus, 100 different training sets were obtained by combining 657 positive samples with one of the 100 random negative sample sets, 100 times. Each of the 100 training sets was composed of 657 real pre-miRNAs and 657 non-pre-miRNAs. These training sets were employed to extract an optimized feature subset.

The GA–SVM algorithm was used to select an optimal feature subset for each training set. However, each optimal feature subset was extracted based on only one training set. In order to improve the classification power, we integrated the 100 optimal feature subsets. First, the frequency of each feature in all 100 optimal feature subsets was calculated, with a maximum value of 1 and minimum value of 0. Feature frequencies are shown in Fig. 2, marked with solid circles. The higher the feature frequency was, the more important that feature was for pre-miRNA identification. Next, we employed the permutation technique to determine the threshold of feature frequency by disturbing category labels 10 times randomly for each of the 100 training sets. Based on the 1000 random sets, the random feature frequencies were calculated, with a highest frequency of 0.09 (dotted line in Fig. 2). Here, 0.09 was set as the threshold of feature frequency. Finally, 13 features were statistic significance, which had the feature frequencies larger than 0.09. The 13 features composed the final optimized feature subset. Feature name, detailed description and frequency of each feature are listed in Table 1. The 13 features related to RNA secondary structure folding measures are denoted with red solid circles in Fig. 2. The result showed that structure features could be more effective than sequence composition for classification of pre-miRNAs, which was consistent with previous approaches such as miPred, microPred and G^{2}DE's [13]. Some features were found in

Table 1
The optimized feature subset containing 13 features.

Feature	Frequency	Description
Diversity	1	The structural diversity
Freq	0.95	The frequency of the MFE structure
dD	0.35	Adjusted base pair distance
zD	0.28	Normalized variants of dD
zF	0.21	Normalized variants of dF, where dF is compactness of the tree-graph representation of the sequence
D_interlp	0.16	Average distance between internal loops
|A-U|/L	0.15	The ratio of |A-U| to length of sequence, where |A-U| is the number of (A-U) base pairs in secondary structure
L_rsym_rgn	0.13	Length of the longest relaxed symmetry region, where the relaxed symmetry region is composed of consecutive stems, symmetrical loops and asymmetrical loops, and the maximally allowed asymmetrical base number in asymmetrical loops is 4.
MFEI1	0.12	The ratio of MFE to %G + C content, where MFE is minimum free energy
dQ	0.12	Adjusted Shannon entropy
zQ	0.12	Normalized variants of dQ
L_sym_rgn	0.11	Length of the longest symmetry region, where the symmetry region is composed of consecutive stems and symmetrical loops, but not bulges, asymmetrical loops and hairpin loops
dH	0.11	Structure enthalpy

Table 2
Performance comparison of four feature sets using five-fold cross validation.

Features	SE	SP	ACC	Fm	MCC
13 optimized features	**100**%	**97.98**%	**98.99**%	**99.01**%	**98.02**%
21 microPred features	99.98%	97.06%	98.52%	98.55%	97.10%
29 miPred features	87.10%	92.63%	89.87%	89.55%	80.02%
All 183 initial features	97.93%	97.13%	97.53%	97.55%	95.11%

The best performance in each index is highlighted with bold front.

many approaches [6,8,13,15] and might be crucial for the prediction of pre-miRNAs, such as MFEI1, dQ, zQ, dD and zD.

3.2. Performance evaluation using five-fold cross validation

In order to evaluate the classification performance of the identified optimized feature subset, it was compared with the initial feature set used here and the feature subsets used in microPred and miPred. Firstly, the remaining 249 pre-miRNAs in release 15 of miRBase and the re-extracted 249 negative samples were combined as the reference sample set. Next, by using the reference sample set, five-fold cross validation based on SVM was performed on the four feature sets (the identified optimized feature subset, the feature subset identified in microPred, the feature subset used in miPred and the initial feature set used here). Finally, the classification performance was evaluated by five indices: SE, SP, ACC, Fm and MCC. As a result, our optimized feature subset achieved the higher SE (100%), SP (97.98%), ACC (98.99%), Fm (99.01%) and MCC (98.02%) than that of the other two feature subsets and the initial total features (Table 2). Thus, the identified optimized feature subset was effective for distinguishing real pre-miRNAs from non-pre-miRNAs.

3.3. Accuracy evaluation using the most recently confirmed pre-miRNAs in miRBase

Based on release 15 of miRBase, the optimized feature subset was used to distinguish the pre-miRNAs from non-pre-miRNAs, and achieved better performance than the initial features and the two other feature subsets (in microPred and miPred). To evaluate the power of identification for new miRNA sequences, 119 newly discovered human pre-miRNAs in release 16 of miRBase were extracted, of which 116 pre-miRNAs had no multiple loops. These pre-miRNAs were predicted using miR-SF, microPred and miPred. As a result, miR-SF achieved the highest prediction accuracy, at 93.97% (109/116), while microPred was 86.21% (100/116) and miPred was 64.66% (75/116). These results indicated that miR-SF is a powerful approach for predicting novel pre-miRNAs.

4. Conclusion

Identification of miRNAs is the first step in understanding their biological characteristics. In recent years, many approaches have been proposed to predict pre-miRNAs, using different feature sets and yielding different performances. However, these did not consider the effect of feature selection on classification. Selection of the feature subset is important for distinguishing pre-miRNAs from non-pre-miRNAs. In this study, the GA–SVM algorithm proposed in our previous work was applied to identify the optimized feature subset including 13 features. Based on this subset, miR-SF was constructed for predicting novel pre-miRNAs.

Firstly, using the reference sample set, the classification performance of the optimized feature subset was compared with the initial 183 features without feature selection and the two existing feature subsets used in previous methods. Based on SVM, we evaluated the performances of the four feature sets by using five-fold cross-validation, which showed that the identified optimized feature set achieved the highest classification power. This demonstrated that feature selection is crucial for identifying pre-miRNAs. Secondly, we constructed the SVM classifier (miR-SF) based on the optimized feature subset. In the prediction of the most recently identified pre-miRNAs in release 16 of miRBase, the accuracy of miR-SF was much higher than that of the two other methods (microPred and miPred). These results demonstrated that miR-SF was effective for predicting novel pre-miRNAs.

Moreover, very significantly, all 13 features extracted in this study were RNA folding measures of secondary structure. This showed that structural features might be more effective than sequence composition for identifying pre-miRNAs. The frequency of the two features “the structural diversity” and “the frequency of MFE structure” were significantly higher than other features. These were important for predicting RNA secondary structure [20,23]. Furthermore, stem-loop secondary structure was important for predicting the pre-miRNAs [7,24]. Thus, the two features we proposed might be promising characteristics for distinguishing the pre-miRNAs from non-pre-miRNAs.

Finally, the classifier based on the optimized feature subset should have the ability to predict novel miRNAs in human genome. However, ~11 million hairpins are identified by scanning the entire human genome [25]. For the huge number of candidate hairpins, enormous false positives can be produced in genome-wide prediction of pre-miRNAs, even if the specificity is 97.98%. How to reduce the false positive rate would be the next problem that needs to be thought deeply. And with more and more miRNAs identified, the false-positive predictions should also be reconsidered [4,5]. Furthermore, although the classification performance of miR-SF was satisfactory, the initial feature set extracted from the known literature did not contain all possible features describing pre-miRNAs. Thus, some effective features might be omitted. With further research on pre-miRNA characteristics, we might find more effective features using GA–SVM, and miR-SF might identify many more potential pre-miRNAs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 30871394, 30900837 and 61073136], the National High Tech Development Project of China, the 863 Program [grant number 2007AA02Z329], the National Basic Research Program of China, the 973 Program [grant number 2008CB517302], Scientific Research Fund of Heilongjiang Provincial Health Department [grant number 2009-253].

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ygeno.2011.04.011.

References

[1] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297.
[2] R.C. Lee, R.L. Feinbaum, V. Ambros, The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75 (1993) 843–854.
[3] S. Griffiths-Jones, H.K. Saini, S. van Dongen, A.J. Enright, miRBase: tools for microRNA genomics, Nucleic Acids Res. 36 (2008) D154–D158.
[4] K.C. Miranda, T. Huynh, Y. Tay, Y.S. Ang, W.L. Tam, A.M. Thomson, B. Lim, I. Rigoutsos, A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes, Cell 126 (2006) 1203–1217.
[5] A. Oulas, A. Boutla, K. Gkirtzou, M. Reczko, K. Kalantidis, P. Poirazi, Prediction of novel microRNA genes in cancer-associated genomic regions—a combined computational and experimental approach, Nucleic Acids Res. 37 (2009) 3276–3287.
[6] K.L. Ng, S.K. Mishra, De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures, Bioinformatics 23 (2007) 1321–1330.
[7] C. Xue, F. Li, T. He, G.P. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinformatics 6 (2005) 310.

[8] R. Batuwita, V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction, Bioinformatics 25 (2009) 989–995.
[9] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001.
[10] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, Z. Lu, MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features, Nucleic Acids Res. 35 (2007) W339–W344.
[11] J.W. Nam, K.R. Shin, J. Han, Y. Lee, V.N. Kim, B.T. Zhang, Human microRNA prediction through a probabilistic co-learning model of sequence and structure, Nucleic Acids Res. 33 (2005) 3570–3581.
[12] S. Kadri, V. Hinman, P.V. Benos, HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models, BMC Bioinformatics 10 (Suppl 1) (2009) S35.
[13] C.H. Hsieh, D.T. Chang, C.H. Hsueh, C.Y. Wu, Y.J. Oyang, Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm, BMC Bioinformatics 11 (Suppl 1) (2010) S52.
[14] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L.C. Showe, M.K. Showe, Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier, Bioinformatics 22 (2006) 1325–1334.
[15] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen, M. Zavolan, Identification of clustered microRNAs using an ab initio prediction method, BMC Bioinformatics 6 (2005) 267.
[16] D.D. Finlay, C.D. Nugent, P.J. McCullagh, N.D. Black, Mining for diagnostic information in body surface potential maps: a comparison of feature selection techniques, Biomed. Eng. Online 4 (2005) 51.
[17] Y. Peng, Z. Wu, J. Jiang, A novel feature selection approach for biomedical data classification, J. Biomed. Inform. 43 (2010) 15–23.
[18] K.Z. Mao, Feature subset selection for support vector machines through discriminative function pruning analysis, IEEE Trans. Syst. Man Cybern. B Cybern. 34 (2004) 60–67.
[19] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artif. Intell. Med. 31 (2004) 91–103.
[20] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (2003) 3429–3431.
[21] B.H. Zhang, X.P. Pan, S.B. Cox, G.P. Cobb, T.A. Anderson, Evidence that miRNAs are different from other RNAs, Cell. Mol. Life Sci. 63 (2006) 246–254.
[22] L. Li, W. Jiang, X. Li, K.L. Moser, Z. Guo, L. Du, Q. Wang, E.J. Topol, S. Rao, A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (2005) 16–23.
[23] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers. 29 (1990) 1105–1119.
[24] D. Zhao, Y. Wang, D. Luo, X. Shi, L. Wang, D. Xu, J. Yu, Y. Liang, PMirP: a pre-microRNA prediction method based on structure-sequence hybrid features, Artif. Intell. Med. 49 (2010) 127–132.
[25] I. Bentwich, A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, Z. Bentwich, Identification of hundreds of conserved and nonconserved human microRNAs, Nat. Genet. 37 (2005) 766–770.