GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Genetic algorithm-based efficient feature selection for classification of pre-miRNAs

P. Xuan[1], M.Z. Guo[1], J. Wang[2], C.Y. Wang[1], X.Y. Liu[1] and Y. Liu[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, Heilongjiang, P.R. China
[2]School of Computer and Information Science, Southwest University,
Chongqing, P.R. China

Corresponding author: M.Z. Guo
E-mail: maozuguo@hit.edu.cn

Genet. Mol. Res. 10 (2): 588-603 (2011)
Received November 7, 2010
Accepted January 8, 2011
Published April 12, 2011
DOI 10.4238/vol10-2gmr969

ABSTRACT. In order to classify the real/pseudo human precursor microRNA (pre-miRNAs) hairpins with *ab initio* methods, numerous features are extracted from the primary sequence and second structure of pre-miRNAs. However, they include some redundant and useless features. It is essential to select the most representative feature subset; this contributes to improving the classification accuracy. We propose a novel feature selection method based on a genetic algorithm, according to the characteristics of human pre-miRNAs. The information gain of a feature, the feature conservation relative to stem parts of pre-miRNA, and the redundancy among features are all considered. Feature conservation was introduced for the first time. Experimental results were validated by cross-validation using datasets composed of human real/pseudo pre-miRNAs. Compared with *microPred*, our classifier *miPredGA*, achieved more reliable sensitivity and specificity. The accuracy was improved nearly 12%. The feature selection algorithm is useful for constructing more efficient classifiers for identification of real human pre-miRNAs from pseudo hairpins.

Key words: Feature selection; Genetic algorithm; Pre-miRNA; Information gain; Conservation

INTRODUCTION

MicroRNAs (miRNA) are non-coding RNAs about 21~23 nucleotides (nt) in length, which can play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Bartel, 2004; Chatterjee and Grosshans, 2009). It has been shown that miRNAs usually participate in a set of important life processes, including growth processes, hematopoiesis, organ formation, apoptosis, and cell proliferation. Furthermore, they are closely related to many kinds of human diseases, including cancer (Bushati and Cohen, 2007). Due to the difficulty of systematically detecting miRNAs from a genome using existing experimental techniques, computational methods play important roles in the identification of new miRNAs.

Precursor miRNAs (pre-miRNAs) of 60~70 nt have stem-loop hairpin structures, which are an important characteristic feature used in the computational identification of miRNAs. Recently, the *ab initio* method based on machine learning was presented and applied to distinguish real pre-miRNAs from candidate hairpin sequences. Through learning from known miRNAs and pre-miRNAs, the features of primary sequence and second structure are extracted. These features are used to construct the classifiers, such as support vector machines (SVM) (Sewer et al., 2005; Xue et al., 2005; Ng and Mishra, 2007; Batuwita and Palade, 2009), probabilistic co-learning model (Nam et al., 2005), naive Bayes (Yousef et al., 2006, 2008), random forest (Jiang et al., 2007), and kernel density estimation (Chang et al., 2008). These classifiers could then classify a candidate sequence as a real pre-miRNA or a pseudo pre-miRNA. However, there are quite a lot extracted features. Not all these features are beneficial, because some features provide little information gain or because some features are redundant relative to other features. Therefore, it is necessary to select the most representative feature subset, which contributes to the improvement of the classification performance.

Triplet-SVM (Xue et al., 2005) classifies human real pre-miRNAs and pseudo pre-miRNAs with 32 structure-sequence features, which considers the structure compositions of 3 adjacent nucleotides and the middle nucleotide among the 3, such as “U(((“ and “A((.”. Xue et al. (2005) used all 32 features to train an SVM classifier. *MiPred* (Jiang et al., 2007) is the extension of *Triplet-SVM*, which uses 2 additional features such as the minimum of free energy (*MFE*) and the randomization test (P value), totaling 34 features. In addition, *MiPred* estimates and ranks the relative importance of a feature. It has been shown that P value and *MFE* are more important than the other 32 structure-sequence features. *miPred* (Ng and Mishra, 2007) is another SVM-based method for classifying human pre-miRNAs from genome pseudo hairpins based on a group of 29 global and intrinsic folding features. These 29 features are evaluated with the F-scores *F1* and *F2* on the class-conditional distributions to measure their discriminative power. A subset of 26 features is selected and 3 strongly correlated features are excluded. *microPred* (Batuwita and Palade, 2009) collects the 29 features from *miPred* and presents 19 new features, totaling 48 features. The following feature selection methods based on filtering are applied for searching the feature space: Divergence, Transformed divergence and Jeffries-Matusita distance. However, the methods described above do not consider feature conservation and redundancy among the features. Therefore, we present a novel method to select a feature subset from the original feature set of pre-miRNA.

MATERIAL AND METHODS

Features of pre-miRNA

The current research indicates that pre-miRNAs have many features about both the primary sequence and secondary structure. These features could be used to classify the real pre-miRNA and pseudo hairpin sequences with an *ab initio* method.

miPred extracted 29 global and intrinsic folding features from human real and pseudo pre-miRNAs. These features are: 1) seventeen base composition variables, including 16 dinucleotide frequencies, that is, $XY\%$ where $X,Y \in \{A,C,G,U\}$, and $(G+C)\%$ content; 2) six folding measures, adjusted base pairing propensity, dP (Schultes et al., 1999), adjusted MFE of folding denoted as dG (Seffens and Digby, 1999; Freyhult et al., 2005), adjusted base pair distance dD (Moulton et al., 2000; Freyhult et al., 2005), adjusted Shannon entropy dQ (Freyhult et al., 2005), MFE index 1 ($MFEI_1$) (Zhang et al., 2006), and MFE index 2 ($MFEI_2$); 3) one topological descriptor, which is the degree of compactness dF (Fera et al., 2004; Gan et al., 2004), and 4) five normalized variants of dP, dG, dQ, dD, and dF: zP, zG, zQ, zD, and zF.

In addition to the above 29 features, *microPred* extracted 19 new features, totaling 48 features. These features are: 1) two *MFE*-related features, MFE index 3 ($MFEI_3$) and MFE index 4 ($MFEI_4$); 2) four RNAfold-related features, normalized ensemble free energy, *NEFE*, frequency of the MFE structure *Freq*, structural diversity denoted as *Diversity*, and a combined feature *Diff*; 3) six thermodynamic features, structure entropy dS and dS/L, structure enthalpy dH and dH/L, melting energy of the structure T_m and T_m/L, where L is the length of pre-miRNA sequence, and 4) seven-base pair-related features: $|A\text{-}U|/L$, $|G\text{-}C|/L$, $|G\text{-}U|/L$, average base pairs per stem *Avg_BP_Stem*, $(A\text{-}U)\%/n\_stems$, $(G\text{-}C)\%/n\_stems$, $(G\text{-}U)\%/n\_stems$, where *n_stems* is the number of stems in the secondary structure.

It has been shown that the above 48 features could efficiently represent the characteristic in primary sequence and secondary structure of pre-miRNA (Batuwita and Palade, 2009). Therefore, we selected a representative feature subset from these 48 features and avoided the redundant features, which is helpful for improving classification performance.

Influencing factors of feature selection

Feature selection aims to select a group of more representative features, which could conserve most information of the original data and distinguish each sample in the dataset. Our feature selection method considers some effective influencing factors, including information gain, feature conservation, and feature redundancy. Here, feature conservation is introduced in this study for the first time.

Information gain

Since all the features of pre-miRNAs are discrete, the feature discrimination is measured by information gain based on Shannon entropy. Suppose a feature of pre-miRNA is x, and the entropy of x is denoted as $H(x)$. When the value of feature y is known, the conditional entropy is $H(x|y)$.

The information gain of features x and y is $IG(x, y)$ (Quinlan, 1993) as shown in

Equation 1. Classification of real or pseudo pre-miRNAs is a two-class problem. *IG*(*c*, *x*) is the information gain of feature *x* relative to classification feature *c*, and *IG* (*c*, *x*) = *H*(*c*) - *H*(*c*| *x*). *IG*(*c*, *x*) are used to measure feature discrimination for the training dataset composed of real or pseudo pre-miRNAs. The features with greater information gain should be selected first.

$$IG\ (x, \mathrm{y}) = H\ (\mathrm{x}) - H\ (x|\mathrm{y}) \qquad \text{(Equation 1)}$$

However, some features have very small information gain. The features would not improve classification performance and would even have a negative effect on the classifier. Thus, they are useless features and should not be selected.

Feature conservation

Pre-miRNAs are typically 60~70 nt, and contain an ~22-bp double-stranded stem and an ~10-nt terminal loop. Recently, computational phylogenetic shadowing showed that the stems of pre-miRNAs are highly conserved in whole genome alignments, whereas most terminal loop sequences are only loosely conserved (Berezikov et al., 2005). Therefore, the conservation degree of a feature is measured through observing the consistent degree of nucleotide sequence in stems. If a feature can reflect the conservation of stems well, a candidate hairpin sequence with a value of the feature similar to one of real pre-miRNAs is more likely a real pre-miRNA. This kind of features should be selected first.

Feature redundancy

The similarity between feature *x* and *y* is calculated with Equation 2. Thus, *Sim*(*x*, *y*) ranges from 0 to 1. *Sim*(*x*, *y*) = 0 means that these two features *x* and *y* are completely irrelevant. *Sim(x*, *y*) = 1 means that *x* and *y* are completely relevant.

$$Sim(x, y) = 2\left[\frac{IG(x, y)}{H(x) + H(y)}\right] \qquad \text{(Equation 2)}$$

When *Sim*(*x*, *y*) is greater than a threshold, the features *x* and *y* are redundant. Selecting both features simultaneously is not useful for improving classification performance. At this time, the feature with greater information gain and feature conservation should be selected and the other one should be filtered out.

Feature conservation extraction

In order to extract the feature conservation, we truncate the conserved stems from pre-miRNA hairpins and count how many same k-mer sequences there are between two stems from two pre-miRNAs. k-mer restricts special k nucleotides to being adjacent. 2-mer and 3-mer are not strict enough for the combination of adjacent nucleotides. However, 5-mer is too strict. Therefore, 4-mer is selected to measure the similarity between two stems.

Figure 1 shows the procedure of extracting feature conservation based on a clustering algorithm. 1) Given the primary sequences of pre-miRNAs, such as the primary sequence of hsa-mir-192, the secondary structures of the pre-miRNAs are predicted by RNAfold (Hofacker et al., 1994). The central loop and the unpaired part between the 5' and 3' arm are then cut off to obtain the conserved stem. 2) Both arms in the stem of pre-miRNAs are scanned with a sliding window whose length is 4 nt and the step length is 1 nt. The frequencies of each 4-mer in the 5' and 3' arm are counted. 3) In the initial stage, each known real pre-miRNA is to be as a single cluster. 4) Two clusters are iteratively merged with the most similar stems into one cluster until the value of similarity between any two clusters is less than a threshold. The threshold is determined by our experiment. When the threshold is set at 12, most of the pre-miRNAs with similar stems could be gathered into a cluster. 5) After the process of clustering, the feature differences are calculated. The feature difference is used to measure the average variation of a feature among all the clusters. Finally, the conservation of each feature is calculated.

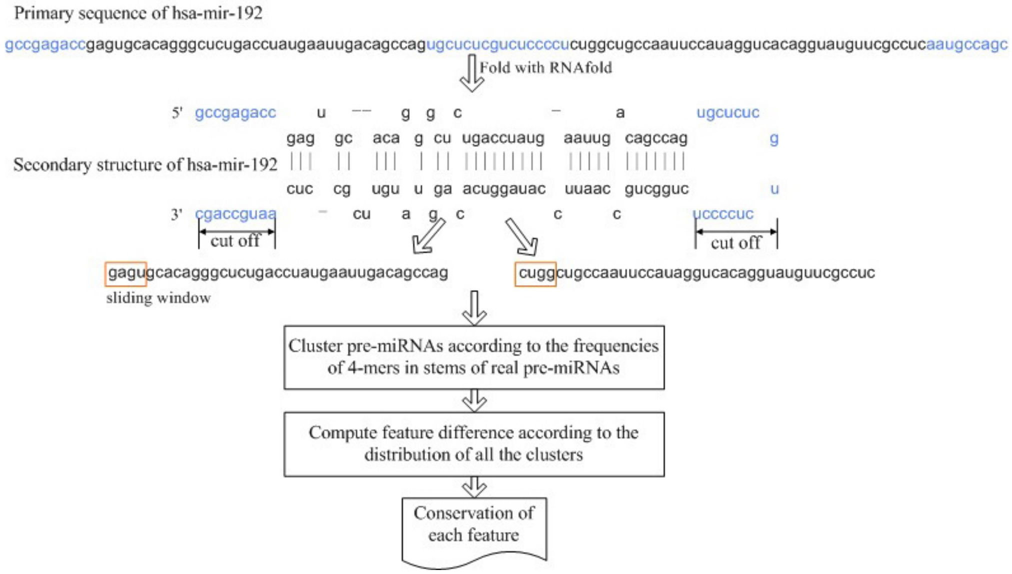
Primary sequence of hsa-mir-192
gccgagaccgagugcacagggcucugaccuaugaauugacagccagugcucucgucuccccucuggcugccaauuccauaggucacagguauguucgccucaaugccagc
Fold with RNAfold
5' gccgagacc u –– g g c – a ugcucuc
gag gc aca g cu ugaccuaug aauug cagccag g
Secondary structure of hsa-mir-192
cuc cg ugu u ga acuggauac uuaac gucgguc u
3' cgaccguaa – cu a g c c c ucccuc
cut off
cut off
gagugcacagggcucugaccuaugaauugacagccag
cuggcugccaauuccauaggucacagguauguucgccuc
sliding window
Cluster pre-miRNAs according to the frequencies of 4-mers in stems of real pre-miRNAs
Compute feature difference according to the distribution of all the clusters
Conservation of each feature

Figure 1. Procedure of extracting feature conservation based on clustering.

Suppose x is a feature and the pre-miRNAs have been gathered into M clusters. N_i is the number of pre-miRNAs in the i^{th} cluster, v_{ij} is the 48-dimensional feature vector of the j^{th} pre-miRNA in the i^{th} cluster, and $v_{ij}[k]$ is the k^{th} dimensional feature value of the j^{th} pre-miRNA. The vector set of the i^{th} cluster is $V_i = \{v_{i1},v_{i2},...v_{iNi}\}$. The mean value of the k^{th} feature in the i^{th} cluster is Avg_{ik}, which is shown as Equation 3. The root-mean-square value of the k^{th} feature is $DAvg_{ik}$ in Equation 4. The average difference value of the k^{th} feature in M clusters is described with Equation 5. As shown in Equation 6, $Con(x_k)$ represents the conservation degree of feature x_k.

$$Avg_{ik} = \frac{\sum_{j=1}^{N_i} v_{ij}[k]}{N_i} \qquad \text{(Equation 3)}$$

$$DAvg_{ik} = \sqrt{\frac{\sum_{j=1}^{N_i} (v_{ij}[k] - Avg_{ik})^2}{N_i}} \qquad \text{(Equation 4)}$$

$$MDAvg_k = \frac{\sum_{i=1}^{M} DAvg_{ik}}{M} \qquad \text{(Equation 5)}$$

$$Con(x_k) = 1 - MDAvg_k \qquad \text{(Equation 6)}$$

Genetic algorithm-based feature selection

Since the pre-miRNA has 48 dimension features, there are 2^{48} feature subsets. It is not feasible to find the optimal feature subset with an exhaustive method. Therefore, we propose a feature selection method based on a genetic algorithm.

The process of feature selection is shown in Figure 2. Given the original complete feature set composed of 48 features, each individual represents a feature subset. First, 200 (the size of population) feature subsets are created as an initial population. The fitness of each feature subset is then calculated. The feature subsets with greater fitness are selected to participate in the crossover and mutation operations. The child feature subsets from crossover and mutation operations and the better feature subsets from the current population are used to generate the new population. The fitness of each feature subset is calculated and the new population is generated iteratively until the stopping condition is satisfied. The best feature subset in the iteration process is the result of our feature selection.

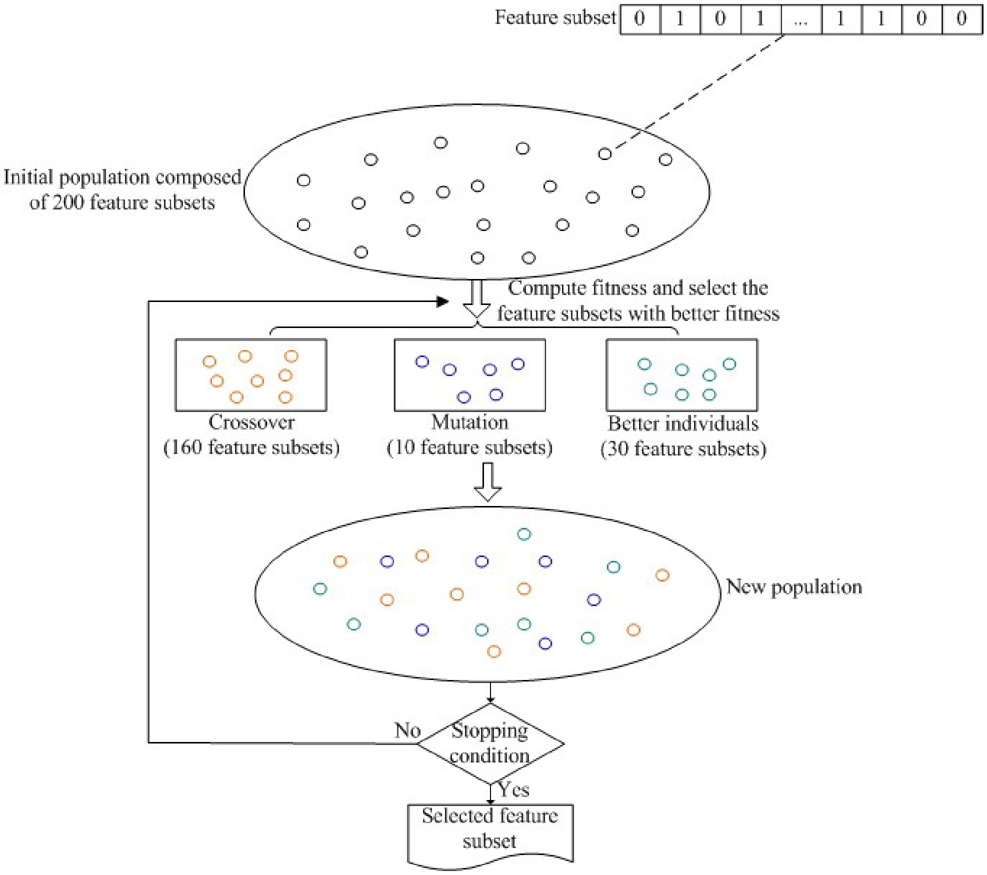
Feature subset
0 1 0 1 ... 1 1 0 0
Initial population composed of 200 feature subsets
Compute fitness and select the feature subsets with better fitness
Crossover (160 feature subsets)
Mutation (10 feature subsets)
Better individuals (30 feature subsets)
New population
No
Stopping condition
Yes
Selected feature subset

Figure 2. Procedure of feature selection based on genetic algorithm.

Individual encoding

Each individual is represented with a binary vector of *L* dimensions and *L* is the size of the complete feature set. Each bit in the binary vector represents whether a feature is included in the current feature subset. $x_i = 1$ means the i^{th} feature is included in the current feature subset. Otherwise, the value of x_i is 0. For instance, the complete feature set is composed of six features, including feature 1 to feature 6. A vector such as (1, 1, 0, 1, 1, 0) means a selected feature subset including feature 1, feature 2, feature 4, and feature 5. *L* is 48 for the 48-dimensional pre-miRNA dataset, which is composed of real pre-miRNAs and pseudo pre-miRNAs.

Population initializing

In order to increase the diversity of individuals, the genetic algorithm typically randomly initializes the population. However, random initializing population would result in the very slow convergence rate of the genetic algorithm. Therefore, we chose random initialization and preference initialization to initialize the first population, which contributes to improving convergence rate and guarantees the diversity of individuals. In random initialization, each bit of an individual is set to 1 with a probability that is selected randomly from the probability set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. After this procedure, 100 feature subsets are created. Second, each bit is set to 1 with the probability of 0.5, and 50 feature subsets are created. In preference initialization, if a feature has greater information gain, the corresponding bit in the individual is preferably 1. The probability of it being set to 1 is according to the rate where the information gain of current feature accounts for the sum of information gains about all features. Thus, another 50 feature subsets are created and the total is 200 individuals in a population.

Fitness calculating

Fitness describes the optimality of an individual (feature subset) so that a particular feature subset may be ranked against all the other feature subsets. In order to generate the next population, the genetic algorithm selects the individuals with greater fitness to participate in the process of crossover and mutation. As a better feature subset, first, each feature should have greater information gain with respect to the class. Second, a selected feature should reflect the conservation of stem well. Third, the redundant features should be avoided for selection.

Suppose a group of features are $x_1, x_2, x_3, x_4, \ldots$, and x_m, k features are selected and the vector of a feature subset is $X' = \{x'_1, x'_2, x'_3, x'_4, \ldots x'_k\}$. x'_i is the i^{th} selected feature in vector X'. $E(X')$ estimates the whole contribution of feature subset X' for classification of real pre-miRNAs and pseudo pre-miRNAs. When calculating $E(X')$, information gain weight (*IGW*), the conservation weight (*ConW*), and the redundancy weight (*RduW*) should be considered.

1) Information gain weight. Assume the class is *c*. *c* is positive ($c = 1$) if the candidate sequence is classified to be real pre-miRNA. *c* is negative ($c = -1$) if the candidate sequence is classified to be pseudo pre-miRNA. $IG(c, x'_i)$ represents the information gain of feature x'_i relative to *c*, which is used to measure the ability of discriminating real and pseudo pre-miRNAs in the training dataset by x'_i.

The minimum, maximum and average value of $IG(c, x'_i)$ are 0.0087, 0.886 and 0.188, respectively. As shown in Equation 7, when $IG(c, x'_i)$ is more than 0.08 (0.08 is determined according to prior experience (Sewer et al., 2005; Ng and Mishra, 2007), $IGW(x'_i)$ is $1+IG[c,x'_i]$. Otherwise, feature x'_i is of little benefit for classification. Thus, $IGW(x'_i)$ is assigned to $-(1-IG[c,x'_i])$.

$$IGW(x_i^{'}) = \begin{cases} 1 + IG(c, x_i^{'}) & \textit{if } IG(c, x_i^{'}) >= 0.08 \\ -(1 - IG(c, x_i^{'})) & \textit{otherwise} \end{cases} \qquad \text{(Equation 7)}$$

2) Conservation weight. $Con(x'_i)$ represents the conservation degree of feature x'_i. The feature conservation ranges from 0 to 1 and its average value is 0.65. As shown in Equation 8, when the feature conservation is more than 0.65 (0.65 is determined according to Sewer et al., 2005 and Ng and Mishra, 2007), the feature contributes to improving the prediction accuracy of the classifier. Therefore, $ConW(x'_i)$ is assigned a plus score. Otherwise, $ConW(x'_i)$ is assigned a minus score.

$$ConW(x_i^{'}) = \begin{cases} Con(x_i^{'}) & \text{if } Con(x_i^{'}) >= 0.65 \\ -(1 - Con(x_i^{'})) & \text{otherwise} \end{cases} \qquad \text{(Equation 8)}$$

3) Redundancy weight. The feature selection should avoid selecting the redundant features, which are strongly relevant to other features. The similarity between two features x'_i and x'_j is measured with $Sim(x'_i, x'_j)$. When $Sim(x,y)$ is greater than the threshold, the features x and y are redundant. The minimum of the $Sim(x'_i, x'_j)$ between the feature pair x'_i and x'_j is 0.0039, the maximum is 0.696, and the averaged value is 0.071. When two features are redundant, selecting both features is not beneficial for classification. The assignment of *RduW* is shown in Equation 9.

$$RduW(x_i^{'}, x_j^{'}) = \begin{cases} Sim(x_i^{'}, x_j^{'}) & \text{if } Sim(x_i^{'}, x_j^{'}) >= 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Equation 9)}$$

$E(X')$ is calculated as Equation 10. Since the *IGW* is more important than the *ConW*, *ConW* is multiplied by 1/2 to coordinate the proportion between *IGW* and *ConW*.

$$E(X') = \sum_{i=1}^{k} \left(IGW(x_i^{'}) + \frac{ConW(x_i^{'})}{2}\right) - \sum_{i=1}^{k} \sum_{j=i+1}^{k} RduW(x_i^{'}, x_j^{'}) \qquad \text{(Equation 10)}$$

As the value of $E(X'_i)$ is greater, the fitness of the i^{th} feature subset X'_i in the population should be greater. Thus, the fitness of X'_i could be estimated by Equation 11,

$$f(X_i^{'}) = \frac{E(X_i^{'})}{\sum_{j=1}^{pop\_size} E(X_j^{'})} \qquad \text{(Equation 11)}$$

where *pop_size* is the number of feature subsets in a population.

Genetic operation

In order to create the next population, a set of feature subsets with greater fitness should be selected to participate in the process of crossover and mutation. The selection strategy, crossover strategy and mutation strategy are described respectively as following.

1) Selection strategy. Selection operation is applied to determine which feature subsets within a population would participate in crossover and mutation, as well as which feature subsets can survive to the next generation. The roulette wheel selection is applied to choose feature subsets according to the proportion of their fitness. Generally, the feature subsets with greater fitness should have a greater chance of survival than the weaker ones. However, the weaker ones are not without any chance. In addition, the roulette wheel selection is helpful in decreasing the probability of the genetic algorithm, achieving local optimal resolution.

2) Crossover strategy. Crossover is a genetic operator that combines two individuals (parents) to produce the new individuals (offspring). In the current population, two feature subsets are chosen with the roulette wheel selection method, referred to as $Parent_1$ and $Parent_2$. There are two crossover strategies, as shown in Figure 3A and B.

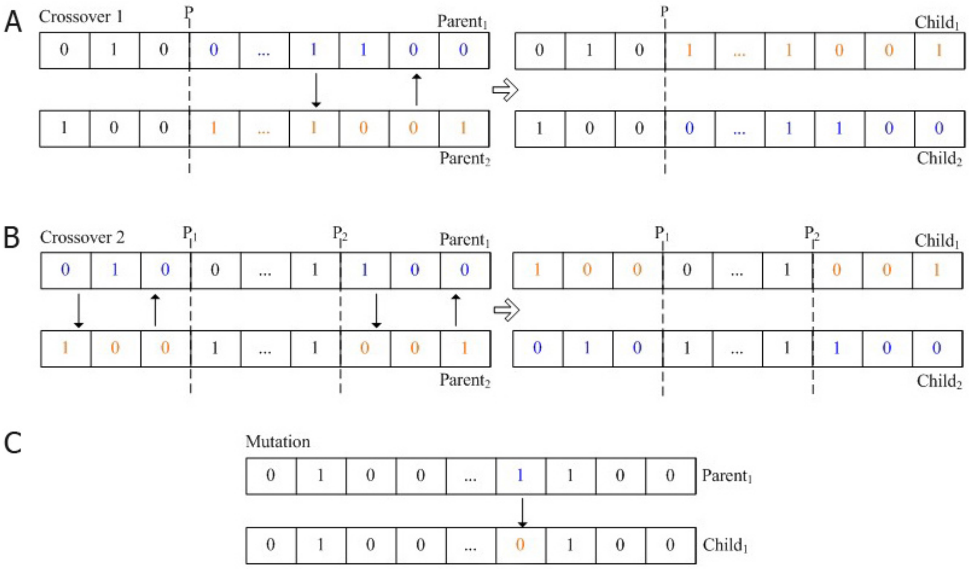

Figure 3. Strategies of crossover and mutation. **A.** Crossover strategy 1. **B.** Crossover strategy 2. **C.** Mutation strategy.

First, a point P is randomly selected. The bits before P in $Parent_1$ and $Parent_2$ are maintained, and the bits after P are exchanged to generate two new offspring feature subsets. Second, two points are randomly selected and represented as P_1 and P_2. The bits between P_1 and P_2 of two parent feature subsets are maintained and the rest of bits are exchanged. In the iteration process of the genetic algorithm, these two crossover strategies are alternately used, which helps to increase the diversity of individuals. The crossover rate is set to 80%. Thus, 80% of individuals (160 feature subsets), which are selected according to individual fitness, would participate in the crossover process.

3) Mutation strategy. Mutation operation could be helpful in maintaining the diversity of individuals in a population. As shown in Figure 3C, a point P is randomly chosen and the value of P is reversed. That is, if the value of P is 1, the value would be set to 0. If the value is 0, it would be set to 1. The 5% selected individuals (10 feature subsets) would participate in the mutation process. In addition, the 15% individuals (30 feature subsets) with greater fitness of the current population would be directly added to the next population, which contributes to protecting the better individuals of each population.

Stopping conditions

There are two conditions to stop the genetic iteration process. First, if the average fitness of the whole population is not changed in the recent *N* iteration or if the difference is smaller than a certain threshold, it means that the evolutionary trend of the population is very slow. Second, the maximum number of iterations is set to finish the genetic algorithm. When either condition is satisfied, the iteration process would be stopped.

Evaluation method

The feature subset could be used to construct an SVM classifier to evaluate our method. The performance of the classifier is measured with three parameters: sensitivity (*SE*), specificity (*SP*), and geometric mean ($G_m = \sqrt{SE \times SP}$). *SE* is the proportion of positive samples (real pre-miRNAs) correctly classified, and *SP* is the proportion of the negative samples (pseudo pre-miRNAs) correctly classified.

Implementation

Our feature selection method is implemented as *GAFeatureSelect* in Java JDK 1.6. *GAFeatureSelect* can be used in any OS with JVM, including Windows, Linux, Unix, etc. After feature selection, the SVM classifier is created with the libSVM2.9 package (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Discretization of feature value could help to compute the entropy value and information gain value of each feature. We discretize a group of values of each feature with the discretization package supported by Weka 3.7.0.

Our *GAFeatureSelect* offers a tool for the high-dimensional data about pre-miRNAs to select a representative feature subset. The selected feature subset contributes to improving classification performance. The feature vectors corresponding to known real pre-miRNAs and pseudo pre-miRNAs are put into the *GAFeatureSelect* as input. The output is the feature subset that could be used to construct an SVM classifier to classify real pre-miRNAs and pseudo hairpin sequences.

RESULTS AND DISCUSSION

Data collection

A classifier of pre-miRNA should distinguish real human pre-miRNAs from both pseudo hairpins and other non-coding RNAs (ncRNAs). Therefore, the positive dataset should be composed of known human pre-miRNAs, while the negative dataset should be composed of both pseudo hairpins and other ncRNAs.

Positive dataset

In order to compare our feature selection method with the *microPred*’s method, we use the same positive dataset with *microPred*. Therefore, the dataset includes 695 human pre-miRNA sequences published in miRBase12.0 (Griffiths-Jones et al., 2008) instead of the current version miRBase15.0. After the redundant sequences have been filtered out, there are 691 non-redundant sequences. A total 660 of these sequences are folded into hairpin secondary structures, and the remaining 31 sequences have multi-branched loops folded with the RNAfold program. In order to identify multiple types of pre-miRNAs, all of these 691 non-redundant pre-miRNA sequences are used as the positive dataset.

Negative dataset

We select 8494 human pseudo hairpin sequences, which are extracted from the protein coding regions and have been previously used in *triple-SVM*, *MiPred*, *miPred*, and *microPred*. The criteria for selecting the pseudo miRNAs are: a minimum of 18 base pairings on the stem of hairpin structure, maximum of -15 kcal/mol free energy of secondary structure, and no multiple loops, which ensure that the extracted pseudo pre-miRNAs are similar to real pre-miRNAs. In addition, the negative dataset also includes 754 other ncRNAs collected by *microPred*, where some ncRNAs have multiple loops, totaling 9248 sequences.

Positive and negative training dataset

Because the classification performance of SVM is also impacted by the training dataset, we propose a two-stage clustering method to select training samples. In the first stage, the samples (real pre-miRNAs or pseudo pre-miRNAs) are clustered according to stem similarity. The samples with similar stems are gathered into a cluster. The clustering method in this stage is the same as the one used in the “Feature conservation extraction” section. The clustered result is as initial status of the second stage. Since 22 features have been selected with our *GAFeatureSelect*, the real or pseudo pre-miRNAs are further clustered according to their positions in 22-dimensional sample space. The classical K-Means clustering algorithm is used in the second stage. However, the distance between a sample and the central point of each cluster is calculated with a new distance formula. Suppose the 22-dimensional feature vector of a sample is x and there are N initial clusters. The vector set of central points is $M = \{m_1, m_2, ..., m_N\}$, where m_i represents the feature vector of the central point in the i^{th} cluster. The distance between the sample x (real pre-miRNA or pseudo pre-miRNA) and the central point of the i^{th} cluster is defined as Equation 12,

$$d_{xm_i} = 1 - \frac{x^t \cdot m_i}{x^t \cdot x + m_i^t \cdot m_i - x^t \cdot m_i} \qquad \text{(Equation 12)}$$

where x^t means the transposed vector x. The greater the value of d is, the farther the distance between the sample and the central point of the i^{th} cluster is. Thus, the samples, which are close together, are merged into the same cluster. Suppose the selection rate is *1/n* and the size of the i^{th} cluster is N_i. Next, we randomly select N_i/n individuals from the i^{th} cluster and add them to the training dataset. A total 333 pre-miRNAs are selected from 691 known real pre-miRNAs as the positive training set, and 442 pseuso pre-miRNAs are selected from 9248 pseudo pre-miRNAs as the negative training set. The selected training dataset is referred to as ***775 training dataset***.

Positive and negative testing dataset

Two groups of positive and negative testing datasets are created. The first group is composed of 350 real pre-miRNAs and 350 pseudo pre-miRNAs. The 350 real pre-miRNAs are randomly selected from the remaining dataset excluding the 333 training pre-miRNAs known pre-miRNAs, and the 350 pseudo pre-miRNAs are selected from 8494 pseudo pre-miRNAs, which is referred to as ***700 random testing dataset***. In the second group, the positive testing dataset is composed of 691 known real pre-miRNAs, and the negative dataset consists of 754 ncRNA sequences. The second group is referred to as ***1445 real and ncRNA testing dataset***. It is well known that some ncRNAs are often wrongly classified by many classifiers as real pre-miRNAs. Therefore, we select all the ncRNAs to create the second negative testing dataset.

Comparison with other methods

The *microPred* applied the following feature subset selection methods for searching the feature space, including *Divergence* (*D*), *Transformed Divergence* (*TD*) and *Jeffries-Matusita distance* (*J-M*). *Jeffries-Matusita distance* achieved the best classification performance in *microPred*. Thus, we compare our feature selection method with the method based on *J-M*. The 22 features are selected with our *GAFeatureSelect*, and their information gain and conservation are listed as follows. First, as shown in Table 1, 12 features (bold) overlap with the 21 features selected by *microPred*. Second, when evaluated statistically onto the 691 non-redundant pre-miRNAs, four pairs of attributes are strongly correlated: *dQ* versus *dD*, *dQ* versus *zQ*, *zQ* versus *zD*, and *dD* versus *zD*. *zQ* is selected due to its higher information gain and conservation compared to *dQ*, *dD* and *zD*. There is a very consensus result in *miPred*, which indirectly certificates our selected feature subset. Third, we found two new strongly correlated pairs of attributes: *dH* versus *dS*, and *dH/L* versus *dS/L*, where *dH* and *dH/L* are selected, respectively.

Table 1. Selected features ranked according to their information gain.

Rank	AttrName	IG(c, attr)	Con(x'_i)	Rank	AttrName	IG(c, attr)	Con(x'_i)
1	**Diversity**	0.886	0.579	11	Tm	0.277	0.429
2	Freq	0.732	0.698	12	**EAFE**	0.242	0.949
3	**MFEI1**	0.596	0.897	13	ZF	0.222	0.673
4	ZG	0.575	0.581	14	**(A-U)%/stems**	0.212	0.624
5	dP	0.427	0.546	15	dH/L	0.147	0.999
6	ZP	0.409	0.588	16	**dF**	0.144	0
7	ZQ	0.357	0.655	17	dH	0.139	0.929
8	**Avg_Bp_Stem**	0.32	0.44	18	**Diff**	0.134	0.78
9	**|A-U|/L**	0.29	0.816	19	UA%	0.114	0.932
10	**dG**	0.287	0.808	20	**MFEI4**	0.114	0.824
11	**MFEI2**	0.279	0.678	21	**(G+C)%**	0.087	0.852

Twelve features (in bold) overlap with the 21 features selected by *microPred*.

The 22 features are used to create the SVM classifiers respectively, referred to as *miPredGA*. *5-Fold cross-validation* is performed on the training data to compare the performance of the two classifiers. We performed 10 repeated evaluations as above for each testing dataset and averaged the results as shown in Table 2. The experimental results indicate that

Table 2. Classification results with different feature selection methods.

Feature selection methods	Number of selected features	Dataset	Classification results (%)		
			SE	SP	G_m
J-M (*microPred*)	21	Training dataset generated by *microPred* (*5-fold cross-validation*)	90.02	97.28	93.58
Feature selection based on GA	22	*775 training dataset* (*5-fold cross-validation*)	99.40	99.34	99.37
J-M (*microPred*)	21	*700 random testing dataset*	90	77.43	83.48
Feature selection based on GA	22		99.71	99.14	99.42
J-M (*microPred*)	21	*1445 real and ncRNA testing dataset*	90.16	77.59	83.64
Feature selection based on GA	22		99.57	98.28	98.93

SE = sensitivity; SP = specificity; G_m = geometric mean; GA = genetic algorithm.

our classifier *miPredGA* outperforms *microPred*. The result of *5-fold cross-validation* with *J-M* was obtained from the publication on *microPred*. Other testing results of *microPred* were obtained through accessing the web server of *microPred* (http://web.comlab.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm). First, *SE* increased by 9.5% on average with our method. The improvement of *SE* could benefit for detecting more new pre-miRNAs. Second, there was an average increase in *SP* of 14.82%. Specificity is usually more important than sensitivity in genome analysis because slight increases in specificity values can greatly decrease false predictions because of the large size of genome sequences. Therefore, the improvement of *SP* is a very significant increase. Thus, our *miPredGA* achieved higher and, especially, much more reliable classification results than *microPred* in terms of both sensitivity and specificity.

Almost all the pre-miRNAs with multiple loops in the testing dataset could be classified correctly, which indicates that, unlike previously reported methods, our method could be sensitive enough to identify pre-miRNAs with multi-loops. There were 4 pre-miRNAs, which are easily misjudged in the positive testing dataset composed of 691 known pre-miRNAs. There are multiple big loops in the precursor of hsa-mir-375. All the precursors of hsa-mir-1308, hsa-mir-1469, and hsa-mir-1825 only include 15 bp. Thus, the minimum of free energy of their secondary structure is higher. The description above could explain why our classifier could not classify these 4 pre-miRNAs.

Complexity analysis

Our feature selection algorithm includes conservation statistics, population initialization and genetic iteration. In the conservation calculation step, when the number of real pre-miRNAs is l, the total running time of conservation statistic is $O(l^2)$. In the initialization step, m individuals are constructed. The total running time of initialization is $O(m)$. The genetic iteration consists of calculating the individual fitness, the crossover and the mutation operation. Suppose k is the number of average selected features in all the feature subsets of a population, and the algorithm iterates n times. Thus, the average running time of the iteration step is $O(n*m* k^2)$. The average total running time of feature selection is $O(l^2+m+n*m* k^2)\approx O(n*m* k^2)$.

In addition, in order to construct an SVM classification model, we just need to select the feature subset only once, and there is no need to select many times. The classification model could then be used to classify real pre-miRNAs and pseudo pre-miRNAs again and again. Therefore, it is worth selecting a representative feature subset to improve classification performance.

CONCLUSIONS

In this paper, we proved that the information gain, conservation and feature redundancy are all important for efficient feature selection. The feature selection algorithm *GAFeatureSelect* obtained the representative feature subset composed of 22 features. The classifier *miRNAPred* trained with the selected samples achieved higher sensitivity and specificity. Further analysis indicated that the improvement of classification accuracy was due to the representative features. In addition, the feature selection algorithm can be used to select informative feature subset for other SVM classifiers, such as *triplet-SVM*, *MiPred*, *miPred*, to increase their classification performance.

ACKNOWLEDGMENTS

Research supported in part by the Chinese Natural Science Foundation (under grant #60932008 and #60871092), the Fundamental Research Funds for the Central Universities (under grant #HIT.ICRST.2010 022), the China Postdoctoral Science Special Foundation (under grant #201003446), and the Returned Scholar Foundation of Educational Department of Heilongjiang Province in China (under grant #1154hz26).

REFERENCES

Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116: 281-297.

Batuwita R and Palade V (2009). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 25: 989-995.

Berezikov E, Guryev V, van de Belt J, Wienholds E, et al. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120: 21-24.

Bushati N and Cohen SM (2007). microRNA functions. *Annu. Rev. Cell Dev. Biol.* 23: 175-205.

Chang DT, Wang CC and Chen JW (2008). Using a kernel density estimation based classifier to predict species-specific microRNA precursors. *BMC Bioinformatics* 9 (Suppl 12): S2.

Chatterjee S and Grosshans H (2009). Active turnover modulates mature microRNA activity in *Caenorhabditis elegans*. *Nature* 461: 546-549.

Fera D, Kim N, Shiffeldrim N, Zorn J, et al. (2004). RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* 5: 88.

Freyhult E, Gardner PP and Moulton V (2005). A comparison of RNA folding measures. *BMC Bioinformatics* 6: 241.

Gan HH, Fera D, Zorn J, Shiffeldrim N, et al. (2004). RAG: RNA-As-Graphs database - concepts, analysis, and features. *Bioinformatics* 20: 1285-1291.

Griffiths-Jones S, Saini HK, van Dongen S and Enright AJ (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36: D154-D158.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, et al. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie/Chemical Monthly* 125: 167-188.

Jiang P, Wu H, Wang W, Ma W, et al. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35: W339-W344.

Moulton V, Zuker M, Steel M, Pointon R, et al. (2000). Metrics on RNA secondary structures. *J. Comput. Biol.* 7: 277-292.

Nam JW, Shin KR, Han J, Lee Y, et al. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* 33: 3570-3581.

Ng KL and Mishra SK (2007). *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics* 23: 1321-1330.

Quinlan JR (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo.

Schultes EA, Hraber PT and LaBean TH (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* 49: 76-83.

Seffens W and Digby D (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* 27: 1578-1584.

Sewer A, Paul N, Landgraf P, Aravin A, et al. (2005). Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics* 6: 267.

Xue C, Li F, He T, Liu GP, et al. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6: 310.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, et al. (2006). Combining multi-species genomic data for microRNA identification using a naive Bayes classifier. *Bioinformatics* 22: 1325-1334.

Yousef M, Jung S, Showe LC and Showe MK (2008). Learning from positive examples when the negative class is undetermined - microRNA gene identification. *Algorithms Mol. Biol.* 3: 2.

Zhang BH, Pan XP, Cox SB, Cobb GP, et al. (2006). Evidence that miRNAs are different from other RNAs. *Cell Mol. Life Sci.* 63: 246-254.