BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 8 2009, pages 989–995
doi:10.1093/bioinformatics/btp107

Genome analysis

microPred: effective classification of pre-miRNAs for human miRNA gene prediction

Rukshan Batuwita* and Vasile Palade*

Oxford University Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Received on November 27, 2008; revised and accepted on February 18, 2009

Advance Access publication February 20, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: In this article, we show that the classification of human precursor microRNA (pre-miRNAs) hairpins from both genome pseudo hairpins and other non-coding RNAs (ncRNAs) is a common and essential requirement for both comparative and non-comparative computational recognition of human miRNA genes. However, the existing computational methods do not address this issue completely or successfully. Here we present the development of an effective classifier system (named as *microPred*) for this classification problem by using appropriate machine learning techniques. Our approach includes the introduction of more representative datasets, extraction of new biologically relevant features, feature selection, handling of class imbalance problem in the datasets and extensive classifier performance evaluation via systematic cross-validation methods.

Results: Our *microPred* classifier yielded higher and, especially, much more reliable classification results in terms of both sensitivity (90.02%) and specificity (97.28%) than the exiting pre-miRNA classification methods. When validated with 6095 non-human animal pre-miRNAs and 139 virus pre-miRNAs from *miRBase*, *microPred* resulted in 92.71% (5651/6095) and 94.24% (131/139) recognition rates, respectively.

Availability: The *microPred* classifier, the datasets used, and the features extracted are freely available at
http://web.comlab.ox.ac.uk/people/ManoharaRukshan.Batuwita/microPred.htm.

Contact: manb@comlab.ox.ac.uk; vasile.palade@comlab.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The microRNA (miRNA) is an important type of non-coding RNA (ncRNA) genes, which participates in post-transcriptional gene regulation. It has been estimated that 20–30% of human genes could be controlled by miRNAs (Kim and Nam, 2006). Very useful associations between miRNA expression levels and human diseases, like different types of cancers and mental retardations, such as Fragile X Syndrome, have been identified (Chang and Mendell, 2007; Croce and Calin, 2005).

Although it has been estimated that there can be thousands of miRNA genes in the human genome (Chang and Mendell, 2007; Kim and Nam, 2006; Miranda *et al*., 2006), only 695 of them have been identified so far according to *miRBase12* (*September 2008*) (Griffiths-Jones *et al*., 2006). The identification of miRNAs by traditional experimental methods, such as direct cloning, suffer from low sensitivity due to the temporal, spatial and low level expression patterns of most miRNAs (Bartel, 2004; Berezikov *et al*., 2006). As an alternative, computational prediction methods dedicated for the discovery of novel miRNA genes by analyzing the genomic DNA play a very crucial role. The main signal used in the computational methods has been the hairpin secondary structure of precursor miRNAs (pre-miRNAs) (Bartel, 2004; Kim and Nam, 2006). The miRNA genes are transcribed as long primary miRNAs which are then processed into ∼80 nt pre-miRNAs folding into hairpin secondary structures. Pre-miRNAs are then cleaved into ∼22 nt mature miRNAs which eventually participate in gene regulation (Kim and Nam, 2006). Figure 1 shows the hairpin secondary structure of human pre-miRNA *hsa-mir-520b* (from *miRBase12*), which was predicted by the *RNAfold* program (Hofacker, 2003).

The available computational methods for human miRNA gene recognition have been developed in two directions, as comparative methods and non-comparative methods. The rationale behind comparative methods is the prediction of genome sequences, which fold into pre-miRNA-like hairpin secondary structures and are conserved in closely related genomes, as novel pre-miRNAs. The corresponding genomic locations are then identified as candidate locations for miRNA genes. Several variations of comparative methods for human miRNA prediction are discussed in *MiRscan* (Lim, 2003), *DIANA-microH* (Szafranski *et al*., 2006), *RNAmicro* (Hertel and Stadler, 2006) and (Berezikov *et al*., 2005). Although these conservation-dependent comparative methods are powerful in genome-wide screening of well-conserved pre-miRNAs among closely related species, they can suffer from low sensitivity with respect to different evolutionary distances (Berezikov *et al*., 2005). That is, these methods could miss novel pre-miRNAs for which close homologous cannot be found due to the limitation of current data, unreliability of alignment algorithms (Loong and Mishra, 2007), or especially due to the availability of rapidly evolving and species-specific miRNAs (Loong and Mishra, 2007; Xue *et al*., 2005). Bentwich *et al*. (2005) has emphasized that the non-conserved miRNAs in human genome, which are missed by comparative methods, can be many and yet to be recognized.

The other approach, non-comparative computational recognition, does not rely on the phylogenetic conservation signal. Therefore, these methods have the capability of recognizing non-conserved/species-specific miRNAs, and miRNAs that can be

*To whom correspondence should be addressed.

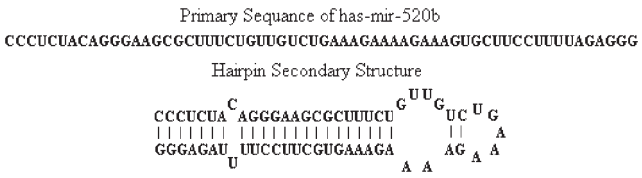

Fig. 1. Human pre-miRNA *has-mir-520b* and its hairpin secondary structure predicted by the *RNAfold* program under the default parameters.

missed due to the limitations of comparative data and methods. The main idea of this approach has been the effective identification of pre-miRNAs among the hairpin secondary structures predicted from the human genome. This is a very challenging task as human genome consists of a vast number of sequences folding into hairpin secondary structures, which are not pre-miRNAs. These structures are called 'pseudo hairpins' (Bentwich *et al*., 2005). Bentwich *et al*. (2005) presents an initial non-comparative method which first screened about 11 million hairpin structures from human genome. Then it combined bioinformatics predictions with microarray analysis and sequence-directed cloning to detect 89 novel human miRNAs, 53 of which are not conserved beyond primates. Following this inaugural work, several classifier systems have been developed as non-comparative prediction methods to distinguish human pre-miRNA hairpins from pre-miRNA-like pseudo hairpins. Xue *et al*. (2005) presents a Support Vector Machine (SVM)-based classifier called *triplet-SVM*, which classifies human pre-miRNAs from pseudo hairpins based on 32 'structure-sequence triplet features'. An extension of *triplet-SVM* method called *MiPred*, which used the random forest algorithm to improve the classification results, is presented in (Jiang *et al*., 2007). Another SVM-based classification method, *miRabela*, which focused on recognizing novel pre-miRNA candidates closely located around known miRNAs genes in human genome, is published in (Sewer *et al*., 2005). The *miPred* (Loong and Mishra, 2007) is also an SVM-based method for the classification of human pre-miRNAs from genome pseudo hairpins based on a set of 29 'global and intrinsic' features.

Presently, the next generation sequencing techniques, such as 'deep-sequencing', have made it possible to discover the tissue-specific and development stage-specific miRNAs and miRNAs expressed at low levels with high sensitivities (Friedlander *et al*., 2008; Ruby *et al*., 2006). However, in addition to the flaws of these high-end sequencing techniques such as sequencing errors, polymorphisms and RNA editing and splicing, these also do present a great deal of computational challenges such as the separating of miRNAs from other sequenced small RNAs or degradation products, and mapping deep-sequencing reads to genomic positions (Friedlander *et al*., 2008). The *miRDeep* (Friedlander *et al*., 2008) is a non-comparative computational method developed for the identification of miRNAs from a pool of sequenced RNA transcripts resulted by deep-sequencing experiments. This method first aligns the transcript reads to genomic locations and selects the genomic sequences from those locations, which can form hairpin secondary structures. Then these hairpin secondary structures are scored by using a probabilistic model, which is based on deep-sequencing signals and Dicer processing features of real pre-miRNAs together with some other features, to identify real pre-miRNAs by rejecting non-pre-miRNA (negative) hairpins.

1.1 Addressing the limitations of the existing computational methods

By analyzing the available datasets, previously published results and capabilities of the existing comparative and non-comparative human miRNA prediction methods, we have identified some limitations associated with them. According to our knowledge, these issues related to human miRNA recognition have not been considered together or completely addressed before.

We first consider the limitations of the existing comparative methods. As mentioned above, there is a vast number of sequences in human genome (∼11 million) that can fold into pre-miRNA-like hairpin secondary structures (Bentwich *et al*., 2005). Most of these are pseudo hairpins, but can have different origins and a variety of other functions (reviewed in Lindow and Gorodkin, 2007; Pearson *et al*., 1996). Due to their different functionalities, it is reasonable to argue that these pre-miRNA-like pseudo hairpins can also be conserved in closely related genomes. Moreover, it has been identified that hairpin secondary structures are common motifs in other types of ncRNAs (Clote *et al*., 2005; Hertel and Stadler, 2006; Zhang *et al*., 2005). Importantly, we observed that 129 other types of ncRNA sequences, which are present in the other ncRNA dataset considered in this study (in Section 2.1), were completely folded into pre-miRNA-like hairpin secondary structures by the *RNAfold* program under the default parameters (at 37°C). These sequences are presented in Table S1 in 'Supplementary Materials and Methods'. Most of these human other ncRNAs can also be conserved in related genomes. Due to these reasons, we can argue that there can be many conserved hairpins in the human genome, which are not pre-miRNAs. Therefore, a proper comparative method for human pre-miRNA recognition should first effectively distinguish whether a genomic sequence folding into a hairpin structure is a real pre-miRNA or not (a pseudo hairpin or a ncRNA-derived hairpin), in addition to its conservation analysis. Among the available comparative methods, *RNAmicro* and *DIANA-microH* methods have partially considered these issues. *RNAmicro* has considered the classification of conserved pre-miRNAs from conserved other ncRNAs, but has not considered the classification pre-miRNAs from pseudo hairpins. On the other hand, although *DIANA-microH* has considered the classification of conserved pre-miRNAs from conserved pseudo hairpins, it has not considered the classification of pre-miRNAs from other ncRNAs.

Now we consider the existing non-comparative methods. As pointed out above, human genome consists of a vast number of pseudo hairpins, and hairpin structures can be found among the complete secondary structures of other types of ncRNAs and their motifs. Therefore, a proper non-comparative approach for novel human pre-miRNA recognition should effectively distinguish real pre-miRNA hairpins from both genome pseudo hairpins and other ncRNAs. However, the existing non-comparative methods (*triplet-SVM*, *MiPred*, *miRabela*, *miPred*) were mainly developed to distinguish real pre-miRNAs from pseudo hairpins only. Although *miRabela* method considered some other ncRNAs (some tRNAs and rRNAs) in the negative training dataset, this dataset was not much representative. In *miPred* method, the classifier trained for the classification of pre-miRNAs from pseudo hairpins was tested for the

classification of another ncRNA dataset. However, the recognition rate obtained was as low as 76.15%.

1.2 Common classification requirement and *microPred* classifier

As discussed above, both comparative and non-comparative computational methods for human miRNA recognition require a suitable method for the effective classification of real pre-miRNA hairpins from both pseudo hairpins and other ncRNAs. In this article, we present a systematic development of a classifier system satisfying this classification requirement by using effective machine learning techniques. Our approach includes the use of a more complete and representative ncRNA and pseudo hairpin dataset as the negative dataset for the classifier development, introduction of new biologically relevant features, feature selection, application of class imbalance learning methods and extensive and systematic training and testing of classifier systems. We name the classifier system developed in this research as '*microPred*'.

As discussed previously under the *miRDeep* method, the classification of pre-miRNA hairpins from non-pre-miRNA hairpins is also an essential requirement in identifying miRNA transcripts in deep-sequencing data. Therefore, *microPred* can also be used in the *miRDeep* method together with its probabilistic model, for example, in a multi-classifier environment, or as an alternative method to distinguish real pre-miRNA hairpins from negative hairpins.

2 MATERIALS AND METHODS

2.1 Biological datasets

The proposed m*icroPred* classifier system should classify real human pre-miRNA hairpins from both pseudo hairpins and other ncRNAs. Therefore, the positive training dataset for the classifier development should be composed of known human pre-miRNAs, while the negative training dataset should be composed of both pseudo hairpins and human other ncRNAs. The datasets selected in this study are introduced below.

2.1.1 Positive dataset-human pre-miRNAs: We retrieved 695 human pre-miRNA sequences published in *miRBase12* (http://microrna.sanger.ac.uk/sequences/) (Griffiths-Jones *et al*., 2006). Then the redundant sequences were filtered out. This retained 691 non-redundant sequences. Out of these, 660 sequences were folded into hairpin secondary structures; while the remaining 31 were folded into structures having multi-branched loops by the *RNAfold* program under the default parameters at 37˚C. These 31 pre-miRNA sequences with their predicted secondary structures are given in Table S2 in 'Supplementary Materials and Methods'. We considered all of these 691 non-redundant pre-miRNA sequences as our positive dataset. The minimum, maximum and average lengths of these sequences were 53, 137 and 89 nt, respectively.

2.1.2 Negative dataset-Pseudo hairpins: we obtained 8494 non-redundant human pseudo hairpin sequences which have been previously used in *triple-SVM*, *MiPred* and *miPred* methods. Originally, these pseudo hairpins were extracted from human RefSeq genes (Pruitt and Maglott, 2001) without undergoing any experimentally validated alternative splicing event. Therefore, it is more likely that these pseudo hairpin sequences do not contain any annotated or un-annotated pre-miRNA sequences. The minimum, maximum and average lengths of these sequences were 62, 119 and 85 nt, respectively.

Human other ncRNAs: ideally, the other ncRNA dataset should be composed of all human other ncRNAs recognized so far except miRNAs. However, a complete human ncRNA dataset is not readily available so far in any RNA database to extract. Although *miPred* method presented an ncRNA dataset, it is not purified due to its containment of animal ncRNAs in addition to human ncRNAs. Therefore, we did not consider that dataset in this study. We obtained the manually annotated human ncRNA dataset discussed in (Griffiths-Jones, 2007), which was originally published in Lander *et al*. (2001). This dataset was formed by starting with the automatic prediction methods, and then carefully removing the predicted pseudogenes manually. Therefore, this dataset is regarded as the best currently available ncRNA predictions for the human genome according to (Griffiths-Jones, 2007). The original dataset contained 1020 ncRNA sequences (except miRNAs) whose sequence lengths ranged from 48 to 548 nt. After removing the redundant sequences and sequences longer than 150 bases (in order to be comparable with human pre-miRNA and pseudo hairpin datasets) 754 sequences were recovered. This dataset included 327 tRNAs, 5 5S-rRNAs, 53 snRNAs, 334 snoRNAs, 32 YRNAs and 3 other miscellaneous RNAs. The updated sequences of snoRNAs were obtained from *snoRNABase* database (Lestrade and Weber, 2006). The average length of a sequence in this dataset was 89 nt. As mentioned in Section 1.1, 129/754 ncRNA sequences in this dataset were folded into hairpin secondary structures by the *RNAfold* program. The remaining 625 sequences were folded into structures with multi-branched loops, most having hairpin motifs. We included all 754 other ncRNA sequences into the negative dataset. We believed that the inclusion of the ncRNA sequences forming multi-branched loop secondary structures (as previously done in *RNAmicro* method) would enrich the negative dataset by providing the additional information representing their hairpin motifs.

2.2 Features

One of the main challenges in machine learning-based classifier development is the extraction of an appropriate set of features on which a classifier is trained to identify each class effectively. In this problem, we had to choose a set of global features that can be extracted regardless of type of the secondary structures of sequences, since our dataset contained both hairpin secondary structures and structures having multi-branched loops.

We first looked into the features used by the existing pre-miRNA classification methods, and considered the 29 'global and intrinsic' features introduced in the *miPred* approach, which can be calculated regardless of the type of the secondary structures of sequences. These features included 17 sequential features [16 dinucleotide features ($AA\%, AC\%, \ldots, UU\%$), and ($\%C+G$)] calculated from the primary sequence itself, 6-folding measures (dG, dP, dQ, dD, $MFEI_1$, $MFEI_2$) and one topological descriptor (dF) calculated from the secondary structure of the sequence, and five normalized variants of dG, dP, dQ, dD and dF, i.e. zG, zP, zQ, zD and zF. When calculating zG, zP, zQ, zD and zF, for each original sequence 1000 random sequences were generated. Here we adopted the same symbols used in *miPred* to denote these 29 features. In order to calculate these features, we used the scripts developed in *miPred*, which are available at http://web.bii.a-star.edu.sg/∼stanley/Publications/Supp_materials/06-002-supp.html.

2.2.1 Newly Introduced Features: In addition to the above features, we newly considered the following 19 features. Let L be the length of a sequence.

New Minimum Free Energy (MFE)-related features:

- MFE Index 3: $MFEI_3 = dG/n\_loops$; where *n _loops* is the number of loops in the secondary structure, and $dG = MFE/L$.
- MFE Index 4: $MFEI_4 = MFE/tot\_bases$; where *tot_bases* is the total number of base pairs in the secondary structure.

RNAfold-related features: these features were extracted using the *RNAfold* program with '$-p$' option at 37˚C, which calculates the partition function and the base paring probability matrix following the algorithms presented in (McCaskill, 1990).

- Normalized Ensemble Free Energy (*NEFE*).
- The frequency of the MFE structure (*Freq*).
- The structural diversity (*Diversity*).

- Related to these features, we introduced the following feature: *Diff* = |*MFE* - *EFE*|/*L*; where *EFE* is the ensemble free energy.

Mfold-related features: these thermodynamical features were calculated using the *UNAfold* program http://dinamelt.bioinfo.rpi.edu/twostate-fold.php (Markham and Zuker, 2005) in the *Mfold* web server package (Zuker, 2003).

- Structure Entropy *dS*, and *dS/L*.
- Structure Enthalpy *dH*, and *dH/L*.
- Melting Energy of the structure *Tm*, and *Tm/L*.

Base pair-related features: these features were calculated by the scripts written by us.

- $|A-U|/L, |G-C|/L, |G-U|/L$; where $|X-Y|$ is the number of $(X-Y)$ base pairs in the secondary structure, $(X-Y) \in \{(A-U), (G-C), (G-U)\}$.
- Average base pairs per stem (*Avg_BP_Stem*): *Avg_BP_Stem* = *tot_bases/n_stems*; where *n_stems* is the number of stems in the secondary structure; *stem* is a structural motif of the secondary structure, which contains more than three contiguous stack of base pairs as defined in *miPred*.
- $\%(A-U)$*/n_stems*, $\%(G-C)$*/n_stems*, $\%(G-U)$*/n_stems*.

All these 48 features are explained in more detail in 'Supplementary Materials and Methods'. When calculating these features, the secondary structures of the sequences were predicted by the *RNAfold* program under the default parameters at 37°C.

2.3 Choice of SVM classifier and model selection

SVM is a supervised machine learning paradigm for solving linear and non-linear classification and regression problems (Burges, 1998). We chose SVM as our classification paradigm in this research due to its high generalization capability (Burges, 1998), ability to find global classification solutions (Burges, 1998) and successful application in bioinformatics and other practical domains.

The model selection for SVMs involves the selection of a kernel function and its parameters which yield the optimal classification performance for a given dataset (Burges, 1998). Among the available kernel functions, the Radial Basis Function (RBF) is the most popular and widely used one due to its higher reliability in finding optimal classification solutions in most practical situations (Keerthi and Lin, 2003). The problems associated with other kernels (Sigmoid, Polynomial, etc.) are discussed in (Burges, 1998; Keerthi and Lin, 2003). Interestingly, it has been found that the Linear kernel could be seen as a special case of RBF and this relationship could be used to ease the parameter selection under RBF (Keerthi and Lin, 2003). We used this method of model selection to train SVM models in this study, which is described in 'Supplementary Materials and Methods'. The performance of the classifier at each parameter point is evaluated by 5-fold cross-validation performance on the training dataset using the *Geometric mean* (*Gm*) metric. The reason for using this metric and its definition are given in Section 2.5. After finding the best parameters giving the highest cross-validation G_m value for the training dataset, a new SVM model was trained using the complete training dataset at those parameters. Then a separate testing dataset was used to measure the performance of the developed classifier. The *matlab* interface of *libsvm2.86* (Chang and Lin, 2001) package was chosen as the SVM training program. All the SVM training experiments in this research were programmed in *matlab*. Before training the SVM classifier systems, the complete dataset was scaled into (−1,+1) interval.

2.4 Feature selection

Our complete feature set consisted of 48 features as introduced in Section 2.2. However, selecting the most discriminative set of features would increase the performance, efficiency and comprehensibility of a classifier system by reducing its complexity. There are basically two types of feature selection methods presented in the machine learning literature: wrapper methods and filter methods (Guyon and Elisseeff, 2003). In wrapper methods, the true classification results given by a learning algorithm is used to evaluate the goodness of feature subsets. However, in this research, the initial attempts to apply wrapper approach for SVMs failed due to the large cross-validation training time required to train SVMs with our large training dataset. Therefore, we focused on filter methods for the selection of the best subset of features.

Filter methods select features prior to training a classifier system based on some discriminative measures. It has been reported that feature subset selection filter methods that consider the interactions among the features are more superior than the feature ranking filter methods that evaluate each feature separately (Guyon and Elisseeff, 2003). Therefore, we applied the following feature subset selection filter methods, which were previously considered in (Kovzoglu and Mather, 2002), with the backward elimination algorithm for searching the feature space: Divergence (*D*), Transformed Divergence (*TD*) and Jeffries–Matusita distance $(J-M)$. These filter methods are briefly explained in 'Supplementary Materials and Methods'.

2.5 Class imbalance problem

The main problem encountered in the dataset selected in this research (introduced in Section 2.1) was its imbalance. That is, the positive dataset (691 pre-miRNAs) was largely outnumbered by the negative dataset (9248 = 8494 pseudo hairpins + 754 other ncRNAs). The ratio of the positive to negative dataset was 1:13.4. It has been well studied in machine learning research that training a classifier system with such an imbalance positive and negative dataset can result in poor classification performance with respect to the minority class (Weiss, 2004)—in this case it would be with respect to the positive (pre-miRNA) class. Generally, a classifier should result in high performance with respect to both positive and negative classes for it to be used for the real-world predictions with high confidence. This problem is known as class imbalance learning problem in machine learning literature. It has been found that SVM classifiers can also be sensitive to class imbalance (Akbani *et al*., 2004; Veropoulos *et al*., 1999).

The solutions developed to overcome this problem are called class imbalance learning methods which can be divided into two main categories: external/data processing methods and internal/algorithmic methods (Weiss, 2004). External methods are independent from the learning algorithm being used, and basically involve in pre-processing of training data to make them balanced. Random over/under-sampling (Weiss, 2004), *SMOTE* (Chawla, 2002) and multi-classifier system (*MCS*) training (Molinara *et al*., 2007) were the external imbalance learning methods considered in this research. Generally, internal methods engage in the modification of the learning algorithm to remove its bias for the majority class. Different error costs (*DEC*) (Akbani *et al*., 2004; Veropoulos *et al*., 1999) and *zSVM* (Imam *et al*., 2006) methods have been developed for SVMs as internal imbalance learning methods. More crucially, it has been found that the best imbalance learning technique which would give the highest performing classifier is domain and dataset dependent (Weiss, 2004). Therefore, we applied all these mentioned external and internal imbalance learning methods for SVMs in order to develop a better performing classifier with our dataset. These imbalance learning methods are briefly described in 'Supplementary Materials and Methods'.

It has been well studied that the most commonly used performance metric 'Accuracy' (Acc = the percentage of correctly classified instances) could not be used to measure the performance of a classifier precisely when the class imbalance problem is present, as it does not reveal the true classification performance with respect to the positive and negative classes separately (Akbani *et al*., 2004; Weiss, 2004). Therefore, we used sensitivity (SE = proportion of the positive examples correctly classified), specificity (SP = proportion of the negative examples correctly classified) and *Geometric mean* ($G_m = \sqrt{\text{SE} \times \text{SP}}$) to measure the performances of the classifiers in this

research, as commonly used in class imbalance learning research (Akbani *et al*., 2004).

3 RESULTS AND DISCUSSION

3.1 Feature selection results

As the first experiment, we trained an SVM classifier with the complete imbalanced dataset to observe the classification performance by using all 48 features. Here, the complete dataset was randomly divided into five equally sized partitions. We used stratified random sampling such that each partition contained the same ratio of positive and negative examples. Then four partitions were used together as the training dataset to train an SVM classifier following the model selection method mentioned in Section 2.3. Next, the resulted model was tested for its classification performance on the fifth data partition. This procedure was repeated five times with different combinations of training (four partitions) and testing (the remaining partition) datasets in an outer 5-fold cross-validation loop, and the classification results on the testing datasets were averaged. We used this systematic cross-validation method for classification performance evaluation throughout this research. Hereafter we refer to this method as '*outer-5-fold-cv*' method. This initial experiment conducted with all 48 features produced the following average test classification results: SE= 80.32%, SP= 98.71% and $G_m = 89.04\%$.

Next we considered only the 29 features introduced in *miPred*, and evaluated the classification results for our dataset using the aforementioned *outer-5-fold-cv* method. This experiment produced SE = 71.98%, SP=98.55% and $G_m = 84.22\%$, which were much lower than the results obtained by using all the features. This showed that the new features introduced by us have a significant influence in more accurate classification of our datasets. Then we applied the filter feature selection methods introduced in Section 2.4 to select the best subset of features from all 48 features, which would give the highest classification results for our dataset. We evaluated the effectiveness of the feature subsets selected by these different filter methods by comparing the true classification results obtained for our dataset on those feature subsets via *outer-5-fold-cv* method. The true classification results obtained subjected to different feature subsets selected in these experiments are summarized in Table 1. Both *D* and *TD* methods resulted the same feature subset.

From these results we chose the feature subset selected by $J-M$ filter method, which yielded the highest classification G_m (90.84% depicted in bold face in Table 1), as the best feature set to be used in the development of the proposed classifier. This feature set contained the following 21 features: $(\%C+G)$, $MFEI_1$, $MFEI_2$, $MFEI_3$, $MFEI_4$, *dG*, *dQ*, *dF*, *zD*, *Diversity*, *NEFE*, *Diff*, *dS*, *dS/L*, $|A-U|/L$, $|G-C|/L$, $|G-U|/L$, *Avg_BP_Stem*, $\%(A-U)$*/n_stem*, $\%(G-C)$*/n_stem*, $\%(G-U)$*/n_stem*. This feature subset retained seven (one sequential and six structural features) out of 29 *miPred* features, and interestingly, 14 out of 19 newly introduced structural features by us. These findings also indicated that the structural features introduced by us and *miPred* method have higher discriminating power for separating pre-miRNAs from negative hairpins than the sequential features. This selected feature subset with less number of features not only gave the highest classification results, but also immensely reduced the large cross-validation training time taken by SVMs, specially, when executing class imbalance learning experiments (e.g. over-sampling) presented in the Section 3.2.

Table 1. True classification results obtained through *outer-5-fold-cv* method with respect to different feature subsets selected

Feature selection methods	Number of features selected	True classification results (%)		
		SE	SP	G_m
All features	48	80.32	98.71	89.04
miPred features	29	71.98	98.55	84.22
J–M	**21**	**83.36**	**99.00**	**90.84**
D and TD	8	67.59	99.44	81.99

3.2 Class imbalance learning results

From the highest classification results obtained with respect to the best feature subset selected in the last section (SE=83.36%, SP=99.00%), it was clear that the resulted classifiers performed poorly with respect to the positive class compared with the negative one. That is, these classifiers developed with our imbalanced dataset (691 positives and 9248 negatives) were biased towards the majority negative class (SP >> SE). If this type of a classifier is used for real-life prediction, due to its lower sensitivity, the chance of miss-detecting the valuable novel pre-miRNAs by it would be quite high. Therefore, these results provided a good evidence for us to apply class imbalance learning methods in this problem for the development of a better performing classifier with respect to both positive and negative classes.

We first considered the external imbalance learning methods. The re-sampling methods (random over/under-sampling and *SMOTE*) were applied until the positive and negative datasets were balanced. In *MCS* training, the negative dataset was randomly divided into 13 sub- datasets based on the negative to positive dataset ratio (∼13.4). Then a set of 13 classifiers were developed such that each one trained on the same positive dataset and one of the negative sub datasets. The majority voting function was used to combine the results of the ensemble. Next we focused on internal imbalance learning methods for SVMs. First, the *DEC* method was applied on the imbalanced dataset with different negative to positive error cost ratios which were in the range $r = (C^-/C^+) = \{0.01, 0.02, \ldots, 0.1\}$. This also includes $r = 0.0747$ which is equivalent to one over the negative to positive class ratio. Under this method, the classifier giving the highest G_m was found at $r = 0.0747$, which agreed with the findings reported in (Akbani *et al*., 2004). As the last imbalance learning method, the *zSVM* method was applied.

These imbalance learning experiments were also conducted through the *outer-5-fold-cv* method. That is, first, an SVM model was trained by applying a particular internal/external imbalance learning method on a training dataset containing four-fifth of the complete dataset. Then its performance was tested on the remaining imbalanced one-fifth of the dataset. This procedure was repeated five times with different combinations of training and testing datasets, and finally, the test results were averaged. Table 2 presents the average test classification results obtained through these class imbalance learning experiments.

From these results, it was observed that all these imbalance learning methods improved the SE by a significant amount (on average by ∼7%) in the expense of reducing some amount of

SP (on average by ∼3.5%), when compared with the preliminary classification results (i.e. SE=83.36%, SP=99.00%). Out of these methods, the *SMOTE* method gave the best performing classifiers for our dataset (with respect to both the classes) by resulting the highest average G_m (93.58%) with SE= 90.02% and SP=97.28%. This method increased the SE by 6.66% by reducing the SP only by 1.72%. Therefore, we chose the best classifier developed under the *SMOTE* method as the final *microPred* classifier. This classifier is publicly available at http://web.comlab.ox.ac.uk/people/ManoharaRukshan.Batuwita/microPred.htm.

We validated the *microPred* predictions on the other animal (non-human) and viral pre-miRNAs published in the *miRBase12*, and obtained a high sensitivity. Out of 6095 other animal pre-miRNAs across 49 species, *microPred* identified 5651 correctly with 92.71% of recognition rate. Out of 139 viral pre-miRNAs across 12 species, 131 were predicted correctly with 92.24% of recognition rate. The prediction results for separate species are given in Table S3 and Table S4 in 'Supplementary Materials and Methods'.

3.3 Comparisons of the existing non-comparative classifiers with *microPred*

When we compared the ways in which the exiting classifiers (*triplet-SVM*, *MiPred*, *miRabela and miPred*) have been developed to the systematic procedure followed to developed *microPred* classifier, we found the following problems of the existing methods.

First, we could clearly observe that the datasets considered in the development of these existing classifiers suffered from class imbalance problem (larger negative dataset compared with positive dataset—see Table 3 under the column 'Complete dataset'). However, surprisingly, none of these methods have considered a proper class imbalance learning analysis for classifiers development. Although it has been mentioned that the *DEC* method was considered for the development of the *miRabela* method, how the training and testing was done has not been given clearly. The *triplet-SVM*, *MiPred* and *miPred* methods chose a random positive and negative more balanced dataset from the complete imbalanced dataset as the training dataset. After training a classifier on this training dataset, its performance was tested on the remaining positives and another randomly chosen negative testing dataset. Table 3 compares the sizes of the complete datasets available for these methods with the sizes of the chosen positive and negative training and testing datasets. Choosing only a small portion of negatives randomly by discarding the rest would neglect the valuable information encoded by those negatives, which could have been more useful for the development of these predictors. In contrast, in the development of *microPred* classifier in this research, we considered the complete available dataset via different class imbalance learning techniques effectively.

Second, none of these existing methods have applied a systematic cross-validation scheme through different training and testing datasets to validate their classification results. In other words, the primary training and testing datasets selected from the complete dataset in these methods (*triplet-SVM*, *MiPred* and *miPred*) to train the classifiers and then to validate their performances were fixed (see Table 3, under the columns 'Training dataset' and 'Testing dataset'). On the contrary, we took a more systematic approach by using different training and testing datasets, which cover the complete dataset, through the *outer-5-fold-cv* method to validate the classification results thoroughly in all the experiments carried out in this research. Therefore, we can state that the classification results reported in our research are much more reliable than the results reported in those existing pre-miRNA classification methods.

Table 2. Classification results obtained through different class imbalance learning methods. The best results are depicted in bold face

Imbalance learning method	SE (%)	SP(%)	G_m(%)
None (imbalanced data)	83.36	99.00	90.84
Over-sampling	91.89	95.20	93.53
Under-sampling	91.03	94.70	92.85
SMOTE	90.02	97.28	**93.58**
MCS	91.46	95.21	93.32
DEC	90.30	93.28	91.78
zSVM	87.70	97.29	92.37

4 CONCLUSION

In this article, we showed that both comparative and non-comparative human miRNA gene recognition approaches require a suitable method for the classification of human pre-miRNA hairpins from both pseudo hairpins and other ncRNAs. Then we presented the systematic development of a classifier system (*microPred*) for this classification requirement by using effective machine learning methods. Our *microPred* classifier obtained higher and more reliable classification results than the existing pre-miRNA classification methods.

Table 3. Comparison of the sizes of complete, training and testing datasets of the existing classifiers with those of the *microPred* classifier developed in this research, which are given in bold face

Methods	Complete dataset		Training dataset		Testing dataset		Classification results (%)		
	#Pos.	#Neg.	#Pos.	#Neg.	#Pos.	#Neg.	SE	SP	G_m
triplet-SVM	193	8494	163	168	30	1000	93.30	88.10	90.66
MiPred	426	8494	163	168	263	265	89.35	93.21	91.26
miPred	323	8494	200	400	123	246	84.55	97.97	91.01
miRabela	178	5395	*Not given clearly in the article*				71.00	97.00	82.99
microPred	**691**	**9248**	***SMOTE + outer-5-fold-cv***				**90.02**	**97.28**	**93.58**

#Pos. = number of positive examples, #Neg. = number of negative examples.

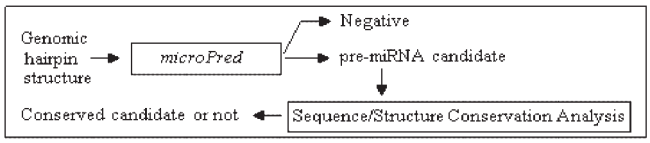

Fig. 2. Use of *microPred* for comparative prediction of pre-miRNAs.

The *microPred* classifier could be used to predict novel human pre-miRNAs in both comparative and non-comparative ways. The non-comparative prediction is straightforward, while the comparative prediction requires additional conservation analysis. As shown in Figure 2, under comparative prediction, *microPred* can be first used to predict whether a genomic sequence falling into a hairpin secondary structure is a real pre-miRNA candidate or not. If it is predicted as a real pre-miRNA hairpin, then it could be further examined for its sequence/structure conservation in other closely related genomes. For sequence conservation analysis, a popular sequence similarity search tool like *BLASTN* (Altschul *et al*., 1990) could be used. In order to find the structure conservation, an RNA structure homology search tool like *INFERNAL* developed in *Rfam* research (Griffiths-Jones, 2005) could be adopted.

Importantly, it would be worth trying to incorporate the advanced features used in the *microPred* with the deep-sequencing data and signals used in the *miRDeep* probabilistic model to develop a better computational method for miRNA discovery. This fact has also been suggested in (Friedlander *et al*., 2008). One way to do this would be through a *MCS*.

As discussed in Section 2.1, 31/674 human pre-miRNAs are folded into secondary structures having multi-branched loops by the *RNAfold* program. There can be many such pre-miRNAs to be recognized. Since our *microPred* classifier can handle the structures with multi-branched loops, this could also be used to screen novel pre-miRNAs folding into structures having multi-branched loops. In this case, however, further investigations have to be carried out for the ways of reducing false positive predictions.

ACKNOWLEDGEMENT

We wish to thank Oxford e-Research Centre for providing us with access to MS-cluster facility to execute our parallel matlab programs.

Conflict of Interest: none declared.

REFERENCES

Akbani,R. *et al*. (2004) Applying support vector machines to imbalanced datasets. In *Proc. of 15th ECML*. Italy, Springer, pp. 39–50.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

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

Bentwich,I. *et al*. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.

Berezikov,E. *et al*. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.

Berezikov,E. *et al*. (2006) Approaches to microRNA discovery. *Nat. Genet.*, **38**, 2–7.

Burges,C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, **2**, 121–167.

Chang,C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date August 01, 2008).

Chang,T.C. and Mendell,J.T. (2007) Roles of microRMAs in vertebrate physiology and human disease. *Annu. Rev. Genomics Hum. Genet.*, **8**, 215–239.

Chawla,N.V. *et al*. (2002) SMOTE: synthetic minority over-sampling technique. *Artif. Intell. Res.*, **16**, 321–357.

Clote,P. *et al*. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**. 578–591.

Croce,C.M. and Calin,G.A. (2005) miRNAs, cancer, and stem cell division. *Cell*, **122,** 6–7.

Friedlander,M.R. *et al*. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Griffiths-Jones,S. *et al*. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, 121–124.

Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34** (*Database Issue*), D140–D144.

Griffiths-Jones,S. (2007) Annotating noncoding RNA genes. *Annu. Rev. Genomics Hum. Genet*., **8**, 279–298.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Hertel,J. and Stadler,P.F. (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, 197–202.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Imam,T. *et al*. (2006) z-SVM: an SVM for improved classification of imbalanced data. In *Proc. of 19th AUS-AI*. Australia, Springer, pp. 264–273.

Jiang,P. *et al*. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, 339–344.

Keerthi,S. and Lin,C.-J. (2003) Asymptotic behaviours of support vector machines with Gaussian kernel. *Neural Comput.*, **15**, 1667–1689.

Kim,V.N. and Nam,J. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.

Kovzoglu,T. and Mather,P.M. (2002) The role of feature selection in artificial neural network applications. *Int. J. Remote Sensing*, **23**, 2919–2937.

Lander,E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lestrade,L. and Weber,M. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*., **34**, 158–162.

Lim,L.P. *et al*. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.

Lindow,M. and Gorodkin,J. (2007) Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol.*, **26**, 339–351.

Loong,K. and Mishra,S. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.

Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, 577–581.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, **29**, 1105–1119.

Miranda,K.C. *et al*. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Molinara,M *et al*. (2007) Facing imbalance classes through aggregation of classifiers. In *Proc. of 14th ICIAP*. IEEE Comp. Soc., Italy, pp. 43–48.

Pearson,C. *et al*. (1996) Inverted repeates, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell Bio-Chem*., **63**, 1–22.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.

Ruby,J.G. *et al*. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, **127**, 1193–1207.

Sewer,A. *et al*. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267–282.

Szafranski,K. *et al*. (2006) Support vector machines for predicting microRNA hairpins. In *Proc. of BIOCOMP*. CSREA Press, Las Vegas, USA, pp. 270–276.

Veropoulos,K. *et al*. (1999) Controlling the sensitivity of support vector machines. *In Proc. of IJCAI*. IJCAII Organization, Sweden, pp. 55–60.

Weiss,G. (2004) Mining with rarity: a unifying framework. *SIGKDD Expl.*, **6**, 7–19.

Xue,C. *et al*. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310–317.

Zhang,B.H. *et al*. (2005) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246–254.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.