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

ABSTRACT

Motivation: In this article, we show that the classification of human precursor microRNA (pre-miRNA) 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/Monfrini/Rukshan.Batuwita/microPred.html.

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-microR (Szataniksa 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
These structures are called ‘pseudo hairpins’ (Bentwich et al., 2005). The main idea of this approach has been the effective identification of pre-miRNAs from other sequenced small RNAs which do present a great deal of computational challenges such as errors, polymorphisms and RNA editing and splicing, these also of these high-end sequencing techniques such as sequencing ‘deep-sequencing’, have made it possible to discover the tissue-specific and development stage-specific miRNAs and miRNAs ‘global and intrinsic’ features. 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 sequences folding 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). Important, 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.

Fig. 1. Human pre-miRNA has-mir-520b and its hairpin secondary structure predicted by the RNAfold program under the default parameters.
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-mRNA 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 microPred 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 rRNAs, 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 ([6 dinucleotide features (A4%, AC%, AC%, CC%, ..., UU%), and (%C + %G) calculated from the primary sequence itself, 6-folding measures (dG, dD, dQ, dF, MFEI, MFEI), and one topological descriptor (dF) calculated from the secondary structure of the sequence, and five normalized variants of dG, dP, dq, dp, and dF; i.e. ≤dG, ≤dP, ≤dq, ≤dp, and ≤dF]. When calculating ≤dG, ≤dP, ≤dq, ≤dp, and ≤dF, 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 = dG/n_loops, where n_loops is the number of loops in the secondary structure, and dG = MFEI.

- MFE Index 4: MFEI = MFEI/nt_bases, where nt_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 pairing 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).
After finding the best parameters giving the highest cross-validation performance, we used the default parameters at 37°C (Markham and Zuker, 2005) in the RNAfold program (http://diana.ccu.edu.tw/RNAfold/html/index.html; Zuker, 2003).

Structure Entropy \( ds \), and \( dS \).
Structure Entropy \( dh \), 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.

\[
\begin{align*}
[A-Y]/L, [G-C]/L, [G-U]/L; \text{where } [X-Y] & \text{ is the number of } (X-Y) \text{ base pairs in the secondary structure, } [X-Y](A-U), (G-C), (G-U).
\end{align*}
\]

Average base pairs per stem \( \text{Avg}_BP_{Stem} = \frac{\text{tot bases}}{n_{stems}} \) 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 mfold.

\[
\% (A-Y)/s_{stems}, \% (G-C)/s_{stems}, \% (G-U)/s_{stems}.
\]

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 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, Polynornial, 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 (Geo) 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 Geo 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 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 (Kovazgu 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 mRNAs). 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 (Molina 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 SVM (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) and specificity (SP = proportion of the negative examples correctly classified) and Geometric mean \( \text{Geo} = \sqrt{SE \times SP} \) to measure the performances of the classifiers in this
As the first experiment, we trained an SVM classifier with all features. The average test classification results obtained through these class imbalance learning methods improved the SE by a significant amount (on average by ~7%) in the expense of reducing some amount of classification power for separating pre-miRNAs from negative hairpins than the sequential features. 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.

### 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 missing 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 MC3 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 \( r = \frac{C^+}{C^-} = \{0.01, 0.02, \ldots, 0.1\} \). 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 classification 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

<table>
<thead>
<tr>
<th>Feature selection methods</th>
<th>True classification results (%)</th>
<th>SE</th>
<th>SP</th>
<th>Gm</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>48</td>
<td>80.32</td>
<td>98.71</td>
<td>89.04</td>
</tr>
<tr>
<td>miPred features</td>
<td>29</td>
<td>71.98</td>
<td>96.55</td>
<td>84.22</td>
</tr>
<tr>
<td>J−M</td>
<td>21</td>
<td>83.36</td>
<td>99.00</td>
<td>90.84</td>
</tr>
<tr>
<td>D and TD</td>
<td>8</td>
<td>67.59</td>
<td>99.44</td>
<td>81.99</td>
</tr>
</tbody>
</table>

MFEI1, MFEI2, MFEI3, MFEI4, \( dG, dQ, dF, zD, \) Diversity, NEFE, Diff, \( dS, dS/L, |A−U|/L, |G−C|/L, |G−U|/L, Avg_BP_Stem, %\((A−U)/\text{stem}\), %\((G−C)/\text{stem}\). 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.
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 \( G_m \) (95.58%) 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 existing classifiers (tripllet-SVM, MiPred, miRabela and miPred) have been developed to the systematic procedure followed to develop 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 predictions. 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.

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

<table>
<thead>
<tr>
<th>Imbalance learning method</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>( G_m ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (imbalance data)</td>
<td>83.36</td>
<td>99.00</td>
<td>90.84</td>
</tr>
<tr>
<td>Over-sampling</td>
<td>91.89</td>
<td>95.20</td>
<td>93.53</td>
</tr>
<tr>
<td>Under-sampling</td>
<td>91.03</td>
<td>94.70</td>
<td>92.85</td>
</tr>
<tr>
<td>SMOTE</td>
<td>90.02</td>
<td>97.28</td>
<td>93.58</td>
</tr>
<tr>
<td>MCS</td>
<td>91.46</td>
<td>95.21</td>
<td>93.32</td>
</tr>
<tr>
<td>DEC</td>
<td>90.30</td>
<td>93.28</td>
<td>91.78</td>
</tr>
<tr>
<td>( \delta )SVM</td>
<td>87.70</td>
<td>97.29</td>
<td>92.37</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Complete dataset</th>
<th>Training dataset</th>
<th>Testing dataset</th>
<th>Classification results (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Pos.</td>
<td>#Neg.</td>
<td>#Pos.</td>
<td>#Neg.</td>
</tr>
<tr>
<td>tripllet-SVM</td>
<td>193</td>
<td>8494</td>
<td>163</td>
<td>168</td>
</tr>
<tr>
<td>MiPred</td>
<td>426</td>
<td>8494</td>
<td>163</td>
<td>168</td>
</tr>
<tr>
<td>miRabela</td>
<td>323</td>
<td>8494</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>microPred</td>
<td>691</td>
<td>9248</td>
<td>Not given clearly in the article</td>
<td>SMOTE + outer-5-fold-cv</td>
</tr>
</tbody>
</table>

\#Pos. = number of positive examples, \#Neg. = number of negative examples.
The non-comparative prediction is straightforward, while the
pre-miRNAs folding into structures having multi-branched loops.
As a result, further investigations have to be carried out
suggested in (Friedlander et al., 2005) could be adopted.
Importantly, it would be worth trying to incorporate the advanced
features used in the microPred with the deep-sequence 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
Berezikov E. et al. (2005) Identification of hundreds of conserved and nonconserved
Berezikov E. et al. (2005) Phylogenetic shadowsing and computational identification of
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 micro RNAs in vertebrate physiology and
Clote P. et al. (2005) Structural RNA has lower folding energy than random RNA of
the same dinucleotide frequency. RNA, 11, 578–591.
Friedlander M.R. et al. (2008) Discovering microRNAs from deep sequencing data
Griffiths-Jones S. et al. (2005) Rfam: annotating non-coding RNAs in complete
Griffiths-Jones S. et al. (2006) miBase: microRNA sequences, targets and gene
Genet., 8, 279–298.
using random forest prediction model with combined features. Nucleic Acids Res.,
35, 339–344.
Keetela S. and Lin C.-J. (2005) Asymmetric behaviours of support vector machines with
409, 860–921.
from genomic pseudo hairpins using global and intrinsic folding measures.
Bioinformatics, 23, 1321–1330.
Markham N.R. and Zelen M. (2005) DINA/MC web server for nucleic acid melting
McCaskill J.S. (1990) The equilibrium partition function and base pairing probabilities
Mirenda K.C. et al. (2006) A pattern-based method for the identification of
Parrish C. et al. (1996) Inverted repeats, stem-loops, and cruciforms: significance for
Ruby J.G. et al. (2006) Large-scale sequencing reveals 21U-RNAs and additional
Sever A. et al. (2005) Identification of clustered microRNAs using an ab initio
prediction method. BMC Bioinformatics, 6, 267–282.
Xue C. et al. (2005) Classification of real and pseudo microRNA precursors using
local structure-sequance features and support vector machine. BMC Bioinformatics,
6, 310–317.
Zhang B.H. et al. (2005) Evidence that microRNAs are different from other RNAs.