An Improved Non-Comparative Classification Method for Human microRNA Gene Prediction

Rukshan Batuwita, Vasile Palade

***Abstract*—micoRNA (miRNA) is a vital class of non-coding RNA genes, which participates in post-transcriptional gene regulation in eukaryotic cell. Interestingly, some close relationships between miRNA expression levels and several human diseases like cancers have been recently uncovered. Difficulties of identifying miRNAs via direct experimental method due to their special and temporal expression patterns make the computational prediction methods paramount important. Specially, non-comparative computational methods would have the advantage of recognizing species-specific miRNAs that can be missed by comparative methods. In this paper we present a systematic development of an improved classifier system for non-comparative human miRNA gene recognition using effective machine learning techniques.**

I. Introduction

microRNA (miRNA) is an important class of non-coding RNAs (ncRNAs). ncRNAs work in a cell as RNA molecules without ever being translated into proteins. miRNAs regulate translation process of messenger RNAs (mRNAs) into proteins. That is, through specific base pairing with mRNAs, miRNAs induce mRNA degradation or translation repression, or both, which are collectively known as 'post-transcriptional gene regulation' [1],[2]. miRNAs operate in highly complex regulatory networks, and control many functions in eukaryotic cell [1],[2]. It has been estimated that 20-30% of human genes could be controlled by miRNAs [2]. Interestingly, very close relationships between miRNA expression levels and human diseases like different types of cancers and mental retardations such as Fragile X Syndrome have been recently identified [3],[4]. These findings have already offered the prospect of using miRNA expression profiles for diagnosis of cancers [5].

Although it has been estimated in [2],[3] that there has to be at least about 1000 conserved and non-conserved miRNA genes in human genome, only 678 of them have been identified so far according to *miRBase11* [6]. *miRBase* [6] is the major miRNA database currently available. The discovery of novel miRNAs and understanding their regulatory networks would provide an opportunity for identifying their functionalities such as the associations with other human diseases. However, the identification of novel miRNA genes by direct experimental methods alone is a very tedious task due to their temporal and special expression patterns [1],[2],[7]. Therefore, proper computational prediction methods are paramount important in the discovery of novel miRNA genes in human and other genomes.

Generally, computational ncRNA gene prediction is a far more difficult task compared to protein coding gene prediction due to the lack of availability of proper signals that can be extracted from ncRNA genes [8],[9]. Nevertheless, the main signals used in the existing methods for ncRNA gene recognition are the features related to RNA secondary structures [8],[9]. Similarly, the major signal used in the miRNA gene prediction is the hairpin (stem-loop) secondary structure of precursor miRNAs (pre-miRNAs) [1],[2]. Pre-miRNA is a vital sub-state of miRNA biogenesis pathway, generally folding into hairpin secondary structures. miRNA genes are transcribed as long primary miRNAs which are then processed into ~90nt pre-miRNAs. Pre-miRNAs are then cleaved into ~22nt mature miRNAs. miRNA biogenesis pathway is described in detail in [1],[2]. Fig. 1 depicts the hairpin secondary structure of human pre-miRNA *hsa-mir-520b* (from *miRBase11*), which is predicted by the *RNAfold* [10] program.

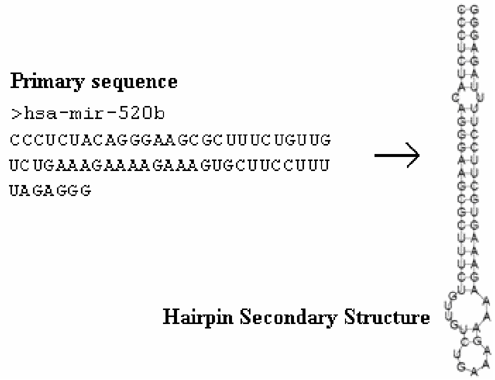
Primary sequence
>hsa-mir-520b
CCCUCUACAGGGAAGCGCUUUCUGUUG
UCUGAAAGAAAAGAAAGUGCUUCCUUU
UAGAGGG
Hairpin Secondary Structure

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

The available computational methods for human miRNA gene recognition have been developed in two directions as comparative methods and non-comparative methods. The main rationale behind comparative methods is the prediction of genome sequences that can be folded into pre-miRNA-like hairpin secondary structures and are conserved in one or

Manuscript received June 26, 2008.

R. Batuwita is with the Oxford University Computing Laboratory, Oxford University, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK. (e-mail: manb@comlab.ox.ac.uk)

V. Palade is with the Oxford University Computing Laboratory, Oxford University, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK. (e-mail: vasile.palade@comlab.ox.ac.uk)

more 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 [11]-[15]. Although these phylogenetic conservation-dependent comparative methods are powerful in genome-wide screening of well-conserved pre-miRNAs among closely related species, these could miss novel miRNAs for which close homologous cannot be found due to the limitation of current data, unreliability of alignment algorithms [16], or especially due to the availability of rapidly evolving (non-conserved) and species-specific miRNAs [16],[17]. Reference [18] has emphasized that the non-conserved miRNAs in human genome (which are missed by comparative methods) are still high and yet to be recognized.

The other approach, non-comparative computational recognition, does not rely on the phylogenetic conservation signal. Therefore, they have the advantage of recognizing non-conserved/species-specific miRNAs, and miRNAs that can be missed due to the limitations of comparative data and methods. The main idea of non-comparative methods developed so far is the effective identification of pre-miRNAs among the hairpin secondary structures predicted from the human genome. This is not a simple task as human genome consists of many random genomic sequences that can fold into pre-miRNA-like hairpin secondary structures, which are not real pre-miRNAs [18]. These are called ‘pseudo hairpins’ [16]-[19]. Reference [18] presents an initial non-comparative method which first screened about 11 million hairpin structures from human genome, most of which were pseudo hairpins. Then it combined bioinformatics predictions with microarray analysis and sequence-directed cloning to detect a set of novel human miRNAs that had been missed by comparative methods earlier. This method predicted 89 novel human miRNAs, interestingly, 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. Reference [17] presents a Support Vector Machine (SVM)-based method called *3SVM*, which classified human pre-miRNAs from pseudo hairpins based on 32 ‘structure-sequence triple features’. Another SVM-based classification method, *miRabela*, which focused on recognizing new miRNA candidates closely located around known miRNAs in human genome, is presented in [19]. *miPred* [16] is also an SVM-based method developed for the classification of human pre-miRNAs based on a set of 29 ‘global and intrinsic’ features.

In this paper we present a systematic development of a novel SVM-based classifier system that could be more useful for the non-comparative prediction of human pre-miRNAs than the exiting methods. In Section 2 we reformulate the classification problem associated with the miRNA gene recognition based on some interesting findings uncovered during this research and by others. There, we also show that the existing non-comparative classification methods have been not developed to meet these classification requirements, and hence, the importance of the proposed method. Section 3 introduces the best possible dataset available to develop the proposed classifier, and the features extracted from the datasets. Section 4 explains the choice of SVM classification paradigm and an efficient technique for SVM model selection. Section 5 discusses the effect of class imbalance problem associated with our datasets and the solutions to overcome it. Section 6 presents the classification results obtained in this research. Finally the paper concludes in section 7.

II. The proposed method

As pointed out above, human genome consists of a vast number of pre-miRNA-like pseudo hairpin sequences [18]. Moreover, it has also been found that hairpin secondary structures are common motifs in other types of ncRNAs [14],[20],[21]. Importantly, we identified 129 other types of ncRNA sequences, which are present in our other ncRNA dataset described in section 3, were completely folded into pre-miRNA-like hairpin secondary structures by the *RNAfold* program under the default parameters (at $37^{\circ}C$). Therefore, it is clear that the effective classification of pre-miRNA hairpins not only from genome pseudo hairpins, but also from other ncRNAs is essential in the non-comparative prediction of novel human pre-miRNAs.

On the other hand, we observed that 31 out of 674 non-redundant human pre-miRNA sequences available in *miRBase11* were folded into secondary structures having multi-branched loops like the structures of most of the other ncRNAs. This folding was also observed by the *RNAfold* program under the default parameters (at $37^{\circ}C$). *miRBase11* IDs of these 31 pre-miRNAs are presented in Table 1. There can be many such pre-miRNAs to be recognized. Therefore, it may be useful to consider the sequences folded into structures with multi-branched loops too when finding novel human pre-miRNAs.

Table 1. *miRBase11* IDs of 31 human pre-miRNA sequences folded into structures with multi-branched loops by the *RNAfold* program. X = ‘has-let’. Y = ‘has-miR’.

IDs of Human pre-miRNA sequence folding into structures with multi-branched loops
X-7a-1, X-7b, X-7d, X-7f-2, Y-7-2, Y-151, Y-181a-2, Y-181b-1, Y-194-2, Y-204, Y-212, Y-217, Y-220c, Y-339, Y-425, Y-453, Y-551a, Y-557, Y-566, Y-572, Y-598, Y-657, Y-744, Y-1224, Y-1225, Y-1227, Y-1236, Y-320b-2, Y-1202, Y-1302-2, Y-1302-3.

However, the existing non-comparative methods [16]–[19] were mainly developed to distinguish real pre-miRNAs from genome pseudo hairpins only. In *miPred* [16] method, the classifier trained for the classification of pre-miRNAs from pseudo hairpins was tested for the classification of an animal ncRNA dataset. However, the recognition rate obtained was low as 76.15%. Although *miRabela* [19] method considered some other ncRNAs (some tRNAs and rRNAs) in its negative dataset, this dataset was not complete. Therefore, these exiting non-comparative methods could predict the hairpin structures of other ncRNAs and their motifs incorrectly as candidate human pre-miRNAs resulting more false positives. Moreover, except *miPred* method, these exiting classifiers can not be used to recognize the pre-miRNAs folding into structures with multi-branched loops.

This paper presents the systematic development of a proper classifier system for the classification of pre-miRNA hairpins from both pseudo hairpins and other ncRNAs by using effective machine learning techniques. This classifier can be used for the prediction of pre-miRNAs folding into hairpin structures as well as pre-miRNAs folding into structures with multi-branched loops.

III. Data and Features

A. *Data*

The proposed classifier system should classify real human pre-miRNAs 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 are introduced below.

1) Positive dataset

human pre-miRNAs: We retrieved 678 human pre-miRNA sequences published in *miRBase11* (http://microrna.sanger.ac.uk/sequences/) [6]. Then we filtered 674 non-redundant pre-miRNA sequences to be used as the positive dataset. The minimum, maximum and average lengths of these pre-miRNAs were 53nt, 137nt and 89nt, respectively.

2) Negative dataset

Pseudo hairpins: We obtained 8,494 non-redundant human pseudo hairpin sequences which have been previously used in *3SVM* [17] and *miPred* [16] methods. Originally these pseudo hairpins were extracted from human *RefSeq* genes [22] 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 62nt, 119nt and 85nt, 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 a manually annotated human ncRNA dataset discussed in [23], which was originally published in [24]. This dataset was created 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 [23]. The original dataset contained 1,020 ncRNA sequences (except miRNAs) whose sequence lengths varied from 48nt to 548nt. 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 to use as the other-ncRNA dataset in this study. 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 (http://www-snorna.biotoul.fr/) [25]. The average length of a sequence in the selected ncRNA dataset was 89nt.

B. *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, one should choose a proper set of features that can be equally extracted from both genomic sequences folding into hairpin secondary structures and sequences folding into structures with multi-branched loops.

In this research we focused on the features used by the existing human miRNA classification methods. Out of these, we selected the 29 global features used in *miPred* [16], which can be calculated regardless of the type of the secondary structure. These features include 17 sequential features (16 dinucleotide features [*AA*%, *AC*%, …, *UU*%], and [*%C+G*]) calculated from the primary sequence itself, 6 folding measures (*dG*, *dP*, *dD*, *dQ*, $MFEI_1$, $MFEI_2$) and 1 topological descriptor (*dF*) calculated from the secondary structure of the sequence, and 5 normalized variants of *dG*, *dP*, *dQ*, *dD* and *dF*, i.e., *zG*, *zP*, *zQ*, *zD* and *zF*. Here we adopted the same symbols used in *miPred* to denote these

features. The secondary structures of the sequences were predicted by the *RNAfold* program with the default parameters at $37°C$. We extracted these features on our positive and negative datasets by adopting the scripts written in *miPred* method, which are available at http://web.bii.a-star.edu.sg/~stanley/Publications.

IV. Choice of SVM Classifier Paradigm and Model Selection

SVM is a supervised machine learning paradigm for solving linear and non-linear classification and regression problems [26]. We chose SVM as the classification paradigm in this research due to its high generalization capability [27], ability to find global classification solutions [27], and successful application in bioinformatics and other practical domains including the previous pre-miRNA classification research [16],[17],[19].

A. *Model Selection of SVM*

The model selection of SVM involves the selection of a kernel function and its parameters, which yield the optimal classification performance for a given dataset [27]. 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 [28],[29]. The problems associated with other kernels (Sigmoid, Polynomial, *etc.*) are discussed in [27]-[29]. 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 [28]. In this method, first, a linear parameter search is conducted under the Linear kernel and the optimal value for the parameter *C* is found. Let's call that value as $\tilde{C}$. Then, the range of one parameter (say γ) under the RBF is fixed. The corresponding best value of the other parameter (*C*) with respect to each value in the range of γ can be calculated by (1). The derivation of this relationship is explained in [28].

$$\log_2 C = \log_2 \tilde{C} - (1 + \log_2 \gamma) \tag{1}$$

Now the parameter search of RBF has become linear which is more efficient than the usual grid search specially with large datasets as ours. We used this method of model selection to train SVM models in this research. The performance of the classifier at each parameter point is evaluated by 5-fold cross-validation training on the training dataset using *G-mean* metric. The reason for using this metric with its definition is given in the next section. Following the above method, we first considered the Linear kernel function and conducted a coarse parameter search with the value of $\log_2 C = [-5, -4, ..., 15]$. Say we found the highest value for cross-validation *G-mean* at $\log_2 C = a$. Then we conducted a narrow parameter search in the space $\log_2 C = [a - 0.75, a - 0.5, ..., a + 0.75]$, found the optimal value for $\log_2 C$, and fixed it as the value of $\log_2 \tilde{C}$. Then the RBF kernel was considered. We fixed the range $\log_2 \gamma = [-15, -14, ...5]$. Then the corresponding value of $\log_2 C$ for each $\log_2 \gamma$ value was found by (1), and a coarse parameter search with each $(\log_2 C, \log_2 \gamma)$ was conducted. If we found the best value for cross-validation *G-mean* at $\log_2 \gamma = b$, again a narrow parameter search was conducted in the range $\log_2 \gamma = [b - 0.75, b - 0.5, ..., b + 0.75]$ with the corresponding $\log_2 C$ values found by (1). After finding the best parameters giving the highest cross-validation *G-mean* 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* [30] package was chosen as the SVM training program. All the experiments in this research were run in *matlab*. Before training the SVM classifier systems, we scaled the complete dataset into the range [-1,+1] following the guidelines in [29].

V. Class Imbalance Problem

One of the main problems encountered in our dataset was its imbalance. That is, the positive dataset (674 pre-miRNAs) was largely outnumbered by the negative dataset (9248 = 8494 pseudo hairpins + 754 other ncRNAs). The ratio between the positive and negative dataset was ~1:13.7. 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 [31], - 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 data imbalance problem is generally known as class imbalance learning problem in machine learning literature. It has been found that SVM classifiers can also be sensitive to class imbalance [32], [33].

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 [31]. External methods are independent from the learning algorithm being used, and basically involved in pre-processing of training data to make them balanced. Random over/under-sampling [31], SMOTE [34] and multi-classifier training [31] were the external imbalance learning methods considered in this research. In random under-sampling the examples from the majority class are removed randomly until a particular class ratio is met [31]. In random over-sampling the examples in majority class are duplicated [31]. SMOTE [34] is an over-sampling technique which introduces new synthetic examples in the neighborhood of minority examples instead of directly duplicating them. In multi-classifier system (MCS) training, the negative (majority) training dataset is randomly divided into several sub datasets each having the similar number of examples as the positive dataset. Then a set of classifiers are developed, each with a different negative dataset and the same positive one. The predictions of the ensemble of the classifiers are combined using a particular combination function.

Generally, internal imbalance learning methods engage in the modification of the learning algorithm to remove its bias for the majority class. Different error costs (DEC) method [32],[33] has been used for SVMs as an internal imbalance learning method. DEC method uses two error cost values in SVM training, such that C^+ for the positive class and C^- for the negative class [33]. Assigning a higher miss classification error cost for the positive class than the negative one would make the classifier less favorable for the negative class under the data imbalance.

More crucially, it has been found that the best imbalance learning technique which would give the highest performing classifier is domain and dataset dependent [31]. Therefore, we applied the above mentioned external and internal imbalance learning methods for SVMs in order to find out the best classification results in this problem. The results obtained are discussed in the next section. It has been well studied that the most commonly used performance metric 'Accuracy' (Acc = the percentage of correctly classify instances) could not be used to measure the performance of a classifier precisely when the class imbalance problem is presented, as it does not reveal the true classification performance with respect to the positive and negative classes separately [31],[32]. Therefore, we used sensitivity (SE = proportion of the positive examples correctly classified), specificity (SP = proportion of the negative examples correctly classified) and Geometric mean $(G-mean = \sqrt{SE*SP})$ to measure the performances of the classifiers developed in this work as used in other class imbalance learning research [32].

VI. RESULTS AND DISCUSSION

We first trained an SVM classifier with the complete imbalance dataset to get an idea about the classification performance that could be obtained. Here, the complete imbalance 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 described in section 4. Next, the resulted model was tested for its classification performance on the fifth dataset. This procedure was repeated five times with different combinations of training and testing partitions in an outer 5-fold cross validation loop and the classification results on the testing datasets were averaged. This experiment gave the following averaged test classification results: Gm=86.36%, SE=75.23%, and SP=99.13%. From these results it was clear that the classifier developed with the imbalance dataset was biased towards the majority negative class (SP >> SE), and this provided a good evidence for the requirement of applying class imbalance learning methods for the development of a proper classifier in this problem.

Under class imbalance learning, we first considered the external imbalance learning methods. The re-sampling methods (random over/under sampling and SMOTE) were applied with 50% and 100% re-sampling rates. As an example, in 50% under-sampling, only 50% of the additional examples in the majority class were randomly removed. In SMOTE algorithm the number of nearest neighbors (k) was used as 14. In MCS training, the negative dataset was divided into 14 subsets based on the positive to negative dataset ratio (~13.7), and subsequently the same number of classifiers were trained. The majority voting function was used to combine the results of the ensemble. Next the DEC method was applied with the imbalance dataset. Following the findings of [32] we chose the ratio between positive error cost (C^+) and negative error cost (C^-) to be equal to one over the class ratio (0.073).

These imbalance learning experiments were also conducted using the 5-fold outer loop cross validation method described at the beginning of this section. That is, first, an SVM model was trained by a particular imbalance learning method with the training dataset containing 4/5th of the complete dataset. Then its performance was tested on a separate testing dataset containing the remaining 1/5th of the dataset. This procedure was repeated 5 times with different training and testing datasets and the test results were averaged. Table 2 summarizes the classification results obtained by these class imbalance learning methods. According to these results the DEC method produced the highest classification results (Gm=92.66%) in this problem with SE=90.80% and SP=94.56%.

TABLE 2. CLASSIFICATION RESULTS OBTAINED BY DIFFERENT CLASS IMBALANCE LEARNING METHODS.

Learning Method	G-mean (%)	SE (%)	SP (%)
Imbalance Data	86.36	75.23	99.13
Over-Sampling – 100%	92.64	90.66	94.66
Over-Sampling – 50%	92.17	87.99	96.55
Under-Sampling – 100%	92.63	91.60	93.67
Under-Sampling – 50%	88.28	78.79	98.93
SMOTE - 100%	92.38	89.91	94.91
SMOTE – 50%	90.95	85.93	96.27
MCS	92.49	90.50	94.52
DEC	**92.66**	90.80	94.56

The results obtained in this research for the classification of human pre-miRNAs from both pseudo hairpins and other ncRNAs outperform the results reported in the existing non-comparative classification methods (presented in [16],[17],[19]) developed solely for the classification of pre-miRNAs from pseudo hairpins (Table 3).

Specially, it was observed that the datasets used by these existing classification methods [16],[17],[19] suffered from class imbalance problem (very large pseudo hairpin dataset compared to pre-miRNA dataset), but, surprisingly, none of these methods have considered using proper class imbalance learning methods for classifiers training. Moreover, the training and testing methods used for the development of these classifiers were not systematic in the sense that none of these methods have used different training and testing datasets in a cross-validation scheme to validate the classification results.

TABLE 3. COMPARISON OF THE BEST CLASSIFICATION RESULTS OBTAINED WITH THE RESULTS OF EXISTING NON-COMPARATIVE METHODS.

Method	Classification of pre-miRNAs from	G-mean (%)	SE (%)	SP (%)
Our	pseudo hairpins + other ncRNAs	92.66	90.80	94.56
3SVM	pseudo hairpins only	90.66	93.30	88.10
miPred	pseudo hairpins only	91.01	84.55	97.97
miRabela	pseudo hairpins only	82.99	71.00	97.00

VII. CONCLUSION

In this paper we presented the development of an improved classifier system for non-comparative human miRNA gene recognition using effective machine learning techniques in a systematic way. This included the introduction of a new ncRNA training dataset, the application of class imbalance learning methods, and the use of systematic cross validation training for classifier development and performance evaluation.

However, the best classification results obtained in the research (SE=90.80% and SP=94.56%) might be further improved before this classifier is applied for the real-world prediction of human miRNA genes. This could be possible by further experimenting with class imbalance learning methods that increase SE without sacrificing a significant amount of SP, and by extracting new biologically relevant features better representing the datasets. Moreover, when the real-world prediction is carried out, one can focus on the neighborhood of known miRNA genes based on the observation made in [19] that human miRNA genes can be found in clusters in the genome. However, investigations should be carried out for the ways of reducing false positive predictions when predicting the human pre-miRNAs folding into structures with multi-branched loops.

REFERENCES

[1] D. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function”, *Cell*, vol. 116, pp. 281-297, 2004.

[2] V. N. Kim and J. Nam, “Genomics of microRNA”, *Trends in Genetics*, vol. 22, pp. 165-173, 2006.

[3] T. C. Chang and J. T. Mendell, “Roles of microRNAs in vertebrate physiology and human disease”, *Annu. Rev. Genomics Hum. Genet*, vol. 8, pp. 215-239, 2007.

[4] A. Esquela-Kerscher and F. J. Slack, “Oncomirs - microRNAs with a role in cancer”, *Nat. Rev. Cancer*, vol. 6, pp. 259-269, 2006.

[5] J. Lu *et al*., “MicroRNA expression profiles classify human cancers”, *Nature*, vol. 435, pp. 834-838, 2005.

[6] S. Griffiths-Jones, R. Grocock, S. Dongen, A. Bateman and A. Enright, “miRBase: microRNA sequences, targets and gene nomenclature”, *Nucleic Acids Research Database Issue*, vol. 34, pp. 140-144, 2006.

[7] E. Berezikov, E. Cuppen and R. Plasterk, “Approaches to microRNA discovery”, *Nature Genetics*, vol. 38, pp. 2 – 7, 2006.

[8] S. R. Eddy, “Computational genomics of noncoding RNA genes”, *Cell*, vol. 109, pp. 137-140, 2002.

[9] A. Huttenhofer, P. Schattner and N. Polacek, “Non-coding RNAs: hope or hype?”, *Trends. Genet*., vol. 21, pp. 289-297, 2005.

[10] L. Hofacker *et al*., “Fast folding and comparison of RNA secondary structures", *Monatshefte f. Chemie*, vol. 125, pp. 167-188, 1994.

[11] L. P. Lim, M. Glasner, S. Yekta, C. Burge and D. Bartel, “Vertebrate microRNA genes”, *Science*, vol. 299, pp.1540, 2003.

[12] E. Berezikov *et al*., “Phylogenetic shadowing and computational identification of human microRNA genes”, *Cell*, vol. 120, pp. 21-24, 2005.

[13] K. Szafranski, M. Megraw, M. Rescnko and A. Hatzigeorgiou, “DIANA-microH: support vector machines for predicting microRNA hairpins”. *In Proc. of Int. Conf. on Bioinformatics & Computational Biology*, Las Vegas, USA, 2006.

[14] J. Hertel and P. F. Stadler, “Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data”, *Bioinformatics*, vol. 22, pp. 197-202, 2006.

[15] S. C. Li, Y. C. Pan and W. C. Lin, “Bioinformatics discovery of microRNA precursors from human ESTs and introns”, *BMC Genomics*, vol. 7, pp.164-175, 2006.

[16] K. Loong and S. Mishra, “De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures”, *Bioinformatics*, vol. 23, pp.1321-1330, 2007.

[17] C. Xue *et al*., “Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine”, *BMC Bioinformatics*, vol. 6, pp. 310-317, 2005.

[18] I. Bentwich *et al*., “Identification of hundreds of conserved and nonconserved human microRNAs”, *Nature Genetics*, vol. 37, pp. 766-770, 2005.

[19] A. Sewer *et al.* “Identification of clustered microRNAs using an ab initio prediction method”, *BMC Bioinformatics*, vol. 6, pp.267-282, 2005

[20] P. Clote, F. Ferre, E. Kranakis and D. Krizanc, “Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency”, *RNA*, vol. 11, pp. 578-591, 2005.

[21] B. Zhang, X Pan, S. Cox, G. Codd and T. Anderson. “Evidence that miRNAs are different from other RNAs”, Cellular and Molecular Life Sciences, vol 63, 2005, pp. 246-254.

[22] K. D. Pruitt and D. R. Maglott, “RefSeq and locuslink: NCBI gene-centered resources”, *Nucleic Acids Res*., vol 29, pp. 137-140, 2001.

[23] S. Griffiths-Jones, “Annotating noncoding RNA genes”, *Annu. Rev. Genomics Hum. Genet*., vol. 8, pp. 279-298, 2007.

[24] E. Lander, “Initial sequencing and analysis of the human genome”, *Nature*, vol. 409, pp. 860-921, 2001.

[25] L. Lestrade and M. Weber, “snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.” *Nucleic Acids Res*., vol. 34, pp. 158-162, 2006.

[26] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, 1995.

[27] C. Burges, “A tutorial on support vector machines for pattern recognition”. *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[28] S. Keerthi and C.–J. Lin, “Asymptotic behaviours of support vector machines with Gaussian kernel”, *Neural Computation*, vol. 15, pp. 1667-1689, 2003.

[29] C.-W. Hsu, C.-C. Chang and C.-J. Lin, “A practical guide to support vector classification”, 2007, Available http://www.csie.ntu.edu.tw/~cjlin.

[30] C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines”, 2001, Available http://www.csie.ntu.edu.tw/~cjlin/libsvm

[31] G. M. Weiss, “Mining with rarity: a unifying framework”, *Special issue on learning from imbalance data sets, SIGKDD Expl*., vol. 6, pp. 7-19. 2004.

[32] R. Akbani, S. Kwek and N. Japkowicz, “Applying support vector machines to imbalanced datasets”, *In Proc. of 15th Euro. Conf. on Machine Learning*, Italy, September, 2004, p. 39-50.

[33] K. Veropoulos, Cristianini N. and C. Campbell, “Controlling the sensitivity of support vector machines”. *In Proc. of the Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, 1999, pp. 55-60.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, *Artificial Intelligence Research*, vol. 16, pp. 321- 357, 2002.