IEICE TRANS. INF. & SYST., VOL.E94–D, NO.3 MARCH 2011

PAPER Special Section on Knowledge Discovery, Data Mining and Creativity Support System

MicroRNA Expression Profiles for Classification and Analysis of Tumor Samples*

Dang Hung TRAN[†a)], ***Nonmember***, **Tu Bao HO**[††b)], ***Member***, **Tho Hoan PHAM**[†c)], ***Nonmember***, ***and*** **Kenji SATOU**[†††d)], ***Member***

SUMMARY One kind of functional noncoding RNAs, microRNAs (miRNAs), form a class of endogenous RNAs that can have important regulatory roles in animals and plants by targeting transcripts for cleavage or translation repression. Researches on both experimental and computational approaches have shown that miRNAs indeed involve in the human cancer development and progression. However, the miRNAs that contribute more information to the distinction between the normal and tumor samples (tissues) are still undetermined. Recently, the high-throughput microarray technology was used as a powerful technique to measure the expression level of miRNAs in cells. Analyzing this expression data can allow us to determine the functional roles of miRNAs in the living cells. In this paper, we present a computational method to (1) predicting the tumor tissues using high-throughput miRNA expression profiles; (2) finding the informative miRNAs that show strong distinction of expression level in tumor tissues. To this end, we perform a support vector machine (SVM) based method to deeply examine one recent miRNA expression dataset. The experimental results show that SVM-based method outperforms other supervised learning methods such as decision trees, Bayesian networks, and backpropagation neural networks. Furthermore, by using the miRNA-target information and Gene Ontology annotations, we showed that the informative miRNAs have strong evidences related to some types of human cancer including breast, lung, and colon cancer.

key words: *microRNA, gene regulation, cancer, support vector machine, feature selection*

1. Introduction

MicroRNAs (miRNAs) are a class of small functional noncoding RNAs (20-24nt) that can play important regulatory roles in animals and plants. They regulate the expression of target genes by binding to specific sites in the 3'UTR of the messenger RNAs (mRNAs) [1], [2]. Each miRNA can bind to many different transcripts and down-regulates protein expression of multiple target genes. Through experimental approaches and bioinformatics applications, thousands of miRNAs have been identified in complex eukaryotic genomes. Together with this, greater than one third of all human genes have been predicted to be miRNA targets [3]–[5]. Therefore, miRNAs are an abundant and important class of regulatory molecules. However, the cellular function of most mammalian miRNAs is still unknown [6]. Hence, it is necessary to understand biological mechanisms that miRNAs are involved.

Recent studies have been shown that several miRNAs were directly involved in human cancers (including lung, breast, brain, liver, and colon cancer) [7]–[10]. While some miRNAs play functions as oncogenes, other ones as tumor suppressors. This is because more than 50% of miRNA genes are located in cancer-associated genomic regions or fragile sites [11]. This evidence also suggests that miRNAs may play a more important role in the human cancers than was previously thought. However, it is still not clear which miRNAs contribute the most information to the specific cancer diseases.

Recently, the high-throughput microarray technology is used as a powerful technique to measure the expression level of miRNAs at biological molecules. While traditional methods only allow one or a few miRNAs to be examined at once, the microarray techniques measure the expression level of thousands of miRNAs simultaneously. Analyzing this expression data can allow us to determine the functional roles of miRNAs in the living cells. Furthermore, investigating of miRNA expression data at the level of biological modules, rather than individual genes, is recognized as an important factor for understanding the cancer regulatory mechanisms [12].

In 2005, Zheng et al. [13] used a Discrete Function Learning (DFL) algorithm to find the subset of miRNAs that shows strong distinction of expression levels in normal and tumor tissues. The DFL algorithm is based on a theorem of information theory. The advantage of using the DFL algorithm is to remove the irrelevant and redundant features so that the induction algorithms may produce better prediction accuracies. To do this, the algorithm needs to examine all possible subsets of features, however, it is a NP-hard problem. Hence, in their paper they proposed several heuristics to reduce the searching space for the DFL algorithm. It thus makes their results are unreliable. Other method, proposed by Kim et al. [14], used a random hypernetwork to identify the gene modules associated with cancers from miRNA microarray data. Hypernetwork is a generalization of the hypergraph by assigning weights to its hyperedges. However,

Manuscript received May 7, 2010.
Manuscript revised September 30, 2010.

[†]The authors are with the Hanoi National University of Education, Vietnam.
[††]The author is with the Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.
[†††]The author is with the Kanazawa University, Kanazawa-shi, 920–1192 Japan.

*The early version of this paper was presented at the 3rd International Conference on Knowledge, Information and Creativity Support Systems

a) E-mail: hungtd@hnue.edu.vn
b) E-mail: bao@jaist.ac.jp
c) E-mail: hoanpt@hnue.edu.vn
d) E-mail: ken@t.kanazawa-u.ac.jp

DOI: 10.1587/transinf.E94.D.416

to construct a hypernework, they faced a combinatorial explosion problem. In order to solve this problem, they generated a hypernetwork by repeating a random hypergraph process. Thus, their hypernetwork was strongly depended on the initial process of generating a first hypergraph, however - it is a random process. Though the experimental results showed that their method provided a competitive performance to BNN and SVM, but they did not proclaim the value of parameters of SVM classifiers that they used to report the results.

In this paper, we present a computational method to (1) predicting the tumor tissues using high-throughput miRNA expression profiles; (2) finding the informative miRNAs that show strong distinction of expression level in tumor tissues. To this end, we perform a supervised learning method to deeply examine one recent miRNA expression dataset [15]. Specifically, we present a support vector machine (SVM) classifier to predicting and analyzing tumor tissues. An SVM is one of the most popular machine learning algorithms and it has good performance in classification problems. In fact, the experimental results show that the SVM-based method outperforms other methods such as decision trees, Bayesian networks, and backpropagation neural networks.

Moreover, to answer the question of which miRNA is important for discriminating between normal and tumor samples, we used a two-step feature selection method to find a subset of informative miRNAs that have strong relevance to the tumor class. The investigation into the biological significance of target genes of informative miRNAs reveals strong evidences related to some types of human cancer, including breast, lung, and colon cancer.

2. Method

2.1 Support Vector Machines for Binary Classification

The support vector machine (SVM) is a learning technique based on statistical learning theory, which from a set of positively and negatively labeled training vectors learns a classifier that can be used to classify new unlabeled test samples. SVM learns the classifier by mapping the input training samples into a possibly high-dimensional feature space, and seeking a hyperplane in this space which separates the positive examples from the negative ones with the largest possible margin (Fig. 1). If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a tradeoff between good classification and large margin.

The implementation of SVM is as follows. Let (x_i, y_i), $i = 1, \ldots, \ell$, be a training dataset, where x_i is a vector and $y_i = \pm 1$ is a class attribute. SVM training solves the following primal problem:

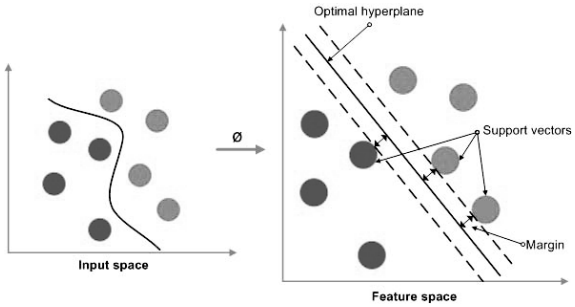

Fig. 1 An illustration of the SVM training method. Red and green circles indicate positive and negative samples to be classified.

$$\begin{cases} \min\limits_{w,b,\xi} \dfrac{w^T w}{2} + C \sum\limits_{i=1}^{\ell} \xi_i \\ y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, \ell \\ \xi_i \geq 0, \quad i = 1, \ldots, \ell \end{cases} \tag{1}$$

Its dual is a quadratic optimization problem:

$$\begin{cases} \min\limits_{\alpha} \dfrac{\alpha^T Q \alpha}{2} - e^T \alpha \\ 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, \ell \\ y^T \alpha = 0 \end{cases} \tag{2}$$

where e is the vector of all ones, $C > 0$ is an error penalty parameter, $y = \{y_i\}_{i=1,\ldots,\ell}$, $Q_{ij} = y_i y_j K(x_i, x_j)$, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is a kernel function, and $\phi(x_i)$ maps x_i into a higher (maybe infinite) dimensional space. So $K(x_i, x_j)$ is a symmetric positive definite function that reflects the similarity between examples x_i and x_j. In this research, we employed a linear function $K(x_i, x_j) = x_i.x_j$, a polynomial function $K(x_i, x_j) = (x_i.x_j + 1)^d$, and a radial basis function (RBF) $K(x_i, x_j) = \exp(-\gamma(x_i - x_j)^2)$ as kernel functions. The SVMs classification function, once trained, has the following form:

$$f(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{3}$$

where $\alpha = \{\alpha_i\}_{i=1,\ldots,\ell}$ is the solution of the above dual problem and b is in the solution of the primal problem. Based on Karush-Kuhn-Tucker theory [16], the solutions of the primal and dual problems satisfy the following equation:

$$\alpha_i \left\{ y_i(w^T \phi(x_i) + b) - 1 + \xi_i \right\} = 0 \tag{4}$$

Therefore, if $\alpha_i \neq 0$ for some i, then $y_i(w^T \phi(x_i) + b) - 1 + \xi_i = 0$. In this case, x_i is called a *support vector* (see Fig. 1).

SVMs have a solid theoretical background, a good performance in practice, and a guaranteed global optimum. It can also handle large datasets and is easier to implement and train than a neural network. A more detailed description of SVMs can be found in [17], [18].

2.2 Ranking Informative Feature Using Fisher Criterion and Linear SVM

Ranking informative (discriminative) features is of fundamental and practical interest in data mining and knowledge discovery. The aim here is to select a subset of relevant features available from the dataset that most contribute to distinguishing instances from different classes. In this research, we use two feature ranking methods to select the informative miRNAs that contribute more information to tumor class. First, we rank all miRNAs based on their Fisher scores and then use a SVM-based feature selection method for ranking miRNAs.

Fisher method. Fisher criterion is one of statistical criteria that is simple, effective and independent of the choice of classification methods. In this criterion, the discriminative strength of each feature is defined as follows. Given a dataset X with two classes, denote instances in class 1 as X^1 and those in class 2 as X^2. Suppose $\bar{x}_j^k$ is the average of the *jth* feature in X^k, the Fisher score of the *jth* feature is:

$$F(j) = \frac{(\bar{x}_j^1 - \bar{x}_j^2)^2}{(s_j^1)^2 + (s_j^2)^2} \tag{5}$$

where

$$(s_j^k)^2 = \sum_{x \in X^k} (x_j - \bar{x}_j^k)^2 \tag{6}$$

The numerator indicates the discrimination between two classes and the denominator indicates the scatter within each class. The larger the Fisher score is, the more likely this feature is more discriminative.

SVM-based method. SVM has been successfully applied to feature selection [19]–[21]. When SVM uses a linear kernel, it finds an optimal hyperplane that separates the positive from the negative class in the original space (not mapping into a higher dimensional space). This optimal hyperplane has then the following form (replacing $K(x, y) = x.y$ in Eq. (3)):

$$f(X = (f_1, f_2, \ldots, f_m)) = \sum_{i=1}^{m} w_i f_i + b \tag{7}$$

We can change the sign of the weights w_i, $i = 1, \ldots, m$, and b in the above function such that if $f(X) > 0$ then X would be classified as a positive example and otherwise, as a negative example. It can be clearly seen that if w_i is positive, then feature i would support the positive class. Otherwise, this feature would support the negative class (or prevent the positive class), and the larger the absolute value of w_i, the stronger feature i supports (or prevents) the respective class. From this remark, we define the weight w_i as the *support* of feature i.

2.3 Validating of Target Genes of Informative miRNAs Using Gene Ontology

With the current knowledge of combinatorial coregulation, it is hard for us to directly validate the predicted target genes of the informative miRNAs. Fortunately, using Gene Ontology (GO) [22] we can validate the target genes of each miRNA with respect to biological processes, cellular components and molecular functions. This validation can be achieved by searching for statistically significant GO terms associated with genes.

3. Results and Discussions

3.1 Datasets

In this work, we used a microarray dataset, which contains expression profiles of miRNAs in human. The original experimental dataset was obtained from Lu et al. [15]. This includes the expression profiles of 151 miRNAs on 223 samples. Of these 223 samples, 166 samples are normal samples from six different tissues, including colon, kidney, prostate, uterus, lung, and breast. The remaining 57 samples are tumor samples from the same six different tissues. We used normal samples as positive data and tumor samples as negative data in our classification problem. To validate the biological significance of target genes of miRNAs, we obtained a set of computationally predicted human miRNA target genes from Krek et al. [23]. The current miRNA target prediction methods are mainly based on the principle of miRNA-target interactions, and the accuracy of these methods has been confirmed by experimental validation of randomly selected miRNA targets [24] and by large-scale gene expression profiling studies [25]. Up to 90% of the randomly selected miRNA targets from the predictions by Krek et al. [23] has been validated as true targets [24].

3.2 Prediction Results

In general, a 10-fold cross-validation is good enough for evaluating the predictive accuracy of classification methods. However, in case of small data (as the dataset used in this study), a leave-one-out cross-validation (LOOCV) performs better than the 10-fold cross-validation. LOOCV simulates the performance of a classification algorithm on unseen samples. In LOOCV, the algorithm is repeatedly retrained, leaving out one sample in each round, and testing each sample on a classifier that was trained without this sample. In our experiments, we used SVMs with three kernel functions (RBF, Linear, and Polynomial kernels) to perform LOOCVs on the miRNA expression dataset (Sect. 3.1). Three popular criteria of *Precision*, *Recall*, and $F1$ are used to evaluate the results. They are defined as follows:

$$Precision = TP/(TP + FP) \tag{8}$$

$$Recall = TP/(TP + FN) \tag{9}$$

$$F1 = 2 * (Precision * Recall)/(Precision + Recall) \tag{10}$$

Where TP, TN, FP, and FN are the number of true positive, true negative, false positive and false negative examples, respectively. $F1$ is the harmonic mean of *Precision*

and *Recall*, it is maximized when *Precision* and *Recall* are maximized at the same time. We also use the area under the curve (AUC) of a receiver operating characteristic (ROC) curve for describing classification performance. The AUC gives a good indication for the overall performance of classifier and whether one classifier performs better than the other classifiers. The bigger value is the better than small ones.

We used LIBSVM (version 2.84) [26] for training and test our SVM classifiers. For preprocessing data, we developed a C++ program to convert the dataset from delimited file format to the format of the LIBSVM. In the other hand, we also conducted a simple scaling on the data by transforming each attribute-value to the range of $[0, 1]$. Three kernel functions (RBF, Linear, and Polynomial) were used in our experiments to validate the classification ability of the SVM. For RBF kernel, we tried several values of C and γ for finding good parameters. We found that the best accuracy was reached when $C = 1.0$ and $\gamma = 0.001$. The performance of the method is shown in Table 1. As can be seen that, in our experiment, we can obtain the highest result of *Precision* = 0.92, *Recall* = 0.98, $F1 = 0.95$, and $AUC = 0.98$ when using SVM with RBF kernel. This indicates that the SVM-based method with RBF kernel is suitable for distinguishing the tumor samples from normal samples when using miRNA expression data.

To make a comparison of the SVM-based method to other classification methods, we used the Weka (version 3.5) [27] to evaluate the performance of backpropagation neural network (BNN), decision tree (DT), k-nearest neighbor (kNN), and bayesian network (BN) methods. According to making a fair comparison, we carry out all experiments by using LOOCV on the same dataset that mentioned in Sect. 3.1. We also carefully selected appropriate parameters for each method. The prediction results of all compared methods, including ours, are shown in Fig. 2. It can be seen that SVM (RBF kernel) classifier is better than other classification methods on all critical measures. For example, SVM classifier gave the $F1 = 0.95$ while BNN, DT (C4.5 algorithm), kNN, and BN had the $F1$ equal to 0.93, 0.85, 0.79, and 0.75, respectively. The highest results are detected from the SVM method with $F1 = 0.95$ and $AUC = 0.98$. While the lowest results are detected from BN with 0.2 and 0.09 lower values on $F1$ and AUC, respectively.

Table 1 The results of SVM classifiers on the miRNA expression dataset.

Kernel function	Precision	Recall	F1	AUC
RBF kernel ($C = 1.0$, $\gamma = 0.001$)	**0.92**	**0.98**	**0.95**	**0.98**
Linear kernel	0.95	0.95	0.95	0.97
Polynomial kernel ($d = 3$)	0.93	0.95	0.94	0.96

3.3 Informative miRNAs Supporting for Tumor Tissues

In this paper, we used feature selection methods to investigate which might play a more dominant role in tumor tissues. Such methods are then used to improve the performance of a classifier and to help understand the problem. Our intention is to determine which features (miRNAs) contribute more information to the tumor tissue class. As described in Sect. 2.2, we used a two-step feature selection method to find the important miRNAs. In the first step, we calculated the Fisher score of each feature (miRNA) and ranked in descending order by their scores. Then, only miRNAs, which have a Fisher score equal or greater than 0.1, were selected for the next step. Those features were evaluated using the SVM classifier with a linear kernel in the second step. When applying to the miRNA expression dataset (see Sect. 3.1), the top 20 contributing features (informative miRNAs) are shown in Table 2.

Of these informative miRNAs, hsa-miR-205 has the first rank with a weight of 0.34. Hsa-miR-125b has the second rank with a weight of 0.27. Three members of let-7 family (hsa-let-7c, hsa-let-7a, and hsa-let-7i) also appear in that list. Other ones, including hsa-miR-146, hsa-miR-145, hsa-

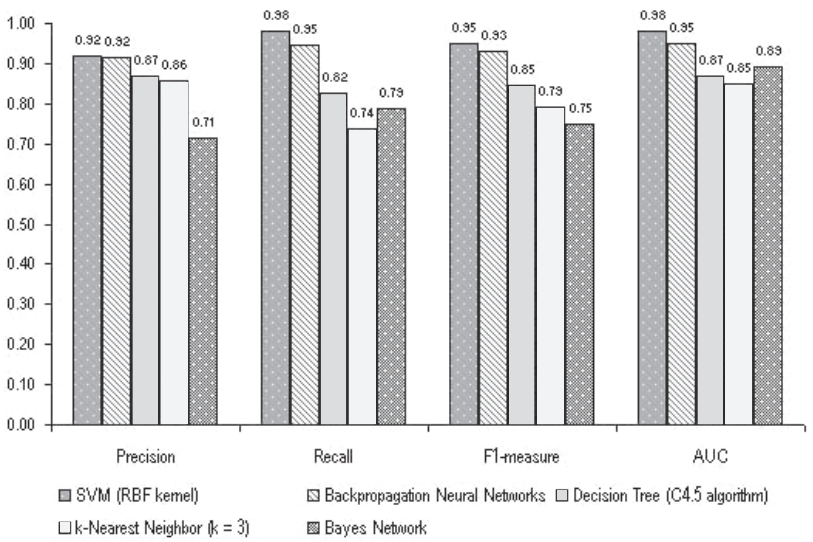

Fig. 2 Comparison results of SVMs and conventional methods on the miRNA expression dataset.

Table 2 The top 20 miRNAs contributing to tumor tissues (negative class) obtained from the trained linear SVM model and their corresponding high-confidence target genes (with *PicTar*_score ≥ 4.5).

Rank	miRNA	|*weight*|	High-confidence target genes
1	hsa-miR-205	0.34	DYRK1A, EIF4E, LHFPL2, MAPK9, YES1
2	hsa-let-7c	0.27	BZW1, CDC25A, CDC34, ERCC6, MAP3K3, QARS, RGS16
3	hsa-miR-125b	0.24	ANPEP, ATP5G2, DVL3, EDEM1, LNPEP, MLF2, PHOX2B, PRDM2, PTPN18, ST6GAL1, TAZ, TDG, TNFSF4
4	hsa-miR-181c	0.22	ADAM11, ARHGEF7, ATXN1, FKBP4, LMO1, MAP3K10
5	hsa-miR-183	0.22	BZW1, EPHA4, FGF9, SAFB, SLC35A1, THBS3, TRAM1, UCHL3
6	hsa-miR-182	0.18	ADD3, ARF4, ARHGEF7, BCL2, EIF5, EPAS1, F13A1, FOXF2, HAS2, HES1, ITPR1, MTM1, PBEF1
7	hsa-miR-146	0.15	NOVA1, TLN2
8	hsa-let-7a	0.12	BZW1, CDC25A, CDC34, ERCC6, MAP3K3, QARS, RGS16
9	hsa-miR-145	0.11	ADD3, AKAP12, FLI1, SEMA3A
10	hsa-miR-222	0.11	CTCF, DYRK1A, INA, MAP3K10, WNT1
11	hsa-miR-33	0.10	CAMK2G, MLLT3, NPY, TSG101, YES1
12	hsa-miR-17-5p	0.10	ATXN1, BCL2L2, EPHA5, HIF1A, KPNA3, MAP3K14, MTMR3, NPAS2, PKD2, PKIA, SLC2A4, TRIP11, TXNIP
13	hsa-miR-181b	0.10	ADAM11, ARHGEF7, ATXN1, EPS8, FKBP4, LMO1, MAP3K10
14	hsa-miR-152	0.10	CCKBR, DNMT1, EDG1, EPAS1, ITPK1, PPP1R10, SFRS2IP, TXNIP, WNT1
15	hsa-miR-153	0.09	DOC2A, CLCN5, EPHA4, ADD1, APC, ITPR1
16	hsa-miR-200a	0.09	MAP3K3, ATXN1, MAP2K4, EXOC5, CUL3, MYH10, UBE3A, TFRC, EPHA7
17	hsa-miR-25	0.08	MAP2K4, LBX1, LHFPL2, HNRPM
18	hsa-let-7i	0.08	BZW1, CDC25A, CDC34, ERCC6, MAP3K3, QARS, RGS16
19	hsa-miR-199b	0.08	NPAS2, ST6GAL1, ZNF238, PXN, EPB41L1, NCOA2, AKAP1, CDH2, PPP1R2
20	hsa-miR-98	0.08	CDC34, BZW1, CDC25A, RGS16, QARS, MAP3K3, ERCC6

miR-181b, and hsa-miR-7-5p are conserved in some mammalian species. Interestingly, we can see that in the top 20 miRNAs contributing to tumor tissues, some of them have been confirmed to be related to several types of human cancer. For example, hsa-miR-205 is located at the region amplified in lung cancer. It had a low expression level in the lung, breast, colorectal, and prostate cancer samples [28]. Besides, Iorio et al. [29] reported that hsa-miR-125b and hsa-miR-145 were indeed involved in human breast cancer. While hsa-miR-125b was down-regulated, hsa-miR-145 was up-regulated in human breast cancer. Their analysis suggested that these miRNAs may potentially act as tumor suppressors. Furthermore, expression of hsa-miR-145 was found at a low level in lung cancer samples compared to normal samples [28]. Based on the target prediction and expression level of hsa-miR-145 in human cancers, Akao et al. [30] also suggested that this miRNA may suppress genes involved in signal transduction and oncogenesis. The expression level of hsa-miR-181b was investigated in the study of Xi et al. [31]. Their analysis revealed that hsa-miR-181b was strongly associated with the mutation status of the *p*53 in tumor.

To determine the biological significance of informative miRNAs, we analyzed target genes that are regulated by these miRNAs. Though there are several available miRNA target prediction methods such as PicTar, miRanda, and TargetScan. A recent study indicated that PicTar had the highest success rate in target gene prediction [32]. We thus utilized PicTar algorithm [23] for obtaining predicted target genes of each informative miRNA. High-confidence target genes (*PicTar*_score ≥ 4.5) of each informative miRNAs are listed in the last column of Table 2.

To test if the target genes for each informative miRNA might be enriched functionally based on arbitrary Gene Ontology (GO) terms [22], we performed GO annotation and significance analysis using GOstat [33]. We observed terms associated significantly with the target genes included in the GO gene-association database (goa_human and Affymetrix HG_U95AV2 Human known genes). In order to find significantly overrepresented GO terms, GOstat calculates a *p*-value upon assuming hyper-geometric distribution of annotated GO terms. Table 3 shows the shared GO terms of target genes of two first informative miRNAs (hsa-miR-205 and hsa-let-7c). We examine the significant terms with *p*-value ≤ 0.005 and *p*-value ≤ 0.001 for hsa-miR-205 and hsa-let-7c, respectively. It can be seen that target genes of hsa-miR-205 and hsa-let-7c belong to biologically functional categories, which are related to post-transcription, protein modification, and regulation of metabolic processes.

Table 4 presents the target genes of hsa-miR-205 and hsa-let-7c in detail. It shows the functional description of each target gene. Interestingly, all these genes function as oncogenes in some types of human cancer. For instance, DYRK1A is a member of a conserved family of serine kinases which a activated by intramolecular tyrosine phosphorylation. Amplification of the DYRK1A has been observed in several different types of cancer. YES1 is an oncogene with kinase activity in a number of solid tumors, including breast and colon [34]. The MAP3K3 gene encodes a transduction protein. More interestingly, the oncogene YES1 and the transduction protein MAP3K3 are potential targets of both hsa-miR-145 and hsa-miR-155, which are known as

Table 3 The GO terms associated to target genes of hsa-miR-205 and hsa-let-7c.

GOId	Biological processes	Genes	*p*-value
hsa-miR-205			
GO:0046777	Protein amino acid autophosphorylation	DYRK1A, YES1	5.86E-04
GO:0006468	Protein amino acid phosphorylation	DYRK1A, YES1, MAPK9	2.24E-03
GO:0016310	Phosphorylation	DYRK1A, YES1, MAPK9	3.14E-03
GO:0006793	Phosphorus metabolic process	DYRK1A, YES1, MAPK9	4.01E-03
GO:0044267	Cellular protein metabolic process	EIF4E, DYRK1A, YES1, MAPK9	4.72E-03
GO:0044260	Cellular macromolecule metabolic process	EIF4E, DYRK1A, YES1, MAPK9	4.72E-03
GO:0043687	Post-translational protein modification	DYRK1A, YES1, MAPK9	4.72E-03
hsa-let-7c			
GO:0050789	Regulation of biological process	BZW1, RGS16, CDC34, MAP3K3, CDC25A, ERCC6	9.27E-03
GO:0006283	Transcription-coupled nucleotide repair	ERCC6	9.27E-03
GO:0044267	Cellular protein metabolic process	BZW1, CDC34, MAP3K3, QARS, CDC25A	9.56E-03
GO:0043283	Biopolymer metabolic process	BZW1, CDC34, MAP3K3, QARS, CDC25A, ERCC6	9.56E-03
GO:0044260	Cellular macromolecule metabolic process	BZW1, CDC34, MAP3K3, QARS, CDC25A	9.56E-03
GO:0065007	Biological regulation	BZW1, RGS16, CDC34, MAP3K3, QARS, CDC25A	9.96E-03
GO:0019538	Protein metabolic process	BZW1, CDC34, MAP3K3, QARS, CDC25A	9.96E-03

Table 4 Description of target genes of hsa-miR-205 and hsa-let-7c.

Ensembl ID	Gene name	Annotation
hsa-miR-205		
ENSG00000157540	DYRK1A	Dual specificity tyrosine-phosphorylation-regulated kinase 1A
ENSG00000151247	EIF4E	Eukaryotic translation initiation factor 4E
ENSG00000145685	LHFPL2	Lipoma HMGIC fusion partner-like 2 protein
ENSG00000050748	MAPK9	Mitogen-activated protein kinase 9
ENSG00000176105	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
hsa-let-7c		
ENSG00000082153	BZW1	Basic leucine zipper and W2 domains 1
ENSG00000164045	CDC25A	Cell division cycle 25 homolog A
ENSG00000099804	CDC34	Cell division cycle 34 homolog
ENSG00000032514	ERCC6	Excision repair cross-complementing rodent repair deficiency
ENSG00000198909	MAP3K3	Mitogen-activated protein kinase 3
ENSG00000172053	QARS	Glutaminyl-tRNA synthetase
ENSG00000143333	RGS16	Regulator of G-protein signaling 16

tumor suppressors in breast cancer [29]. Thus, it is reasonable for us to conclude that the method presented in this research can find informative miRNAs that contribute much information to tumor tissues.

4. Conclusions

We have presented a computational method based on SVMs to analyze microRNA expression profiles from a wet lab experiment. Our prediction results indicated that support vector machines are able to classify tissues base on this data and outperforms other classification methods, such as decision trees, Bayesian networks, and backpropagation neural networks.

Furthermore, relied on a two-step feature selection method using Fisher criterion and SVM with linear kernel, we found a subset of informative miRNAs which contribute more information for discriminating between normal and tumor samples. An analysis of predicted target genes of these miRNAs allowed us to determine the functional roles of these miRNAs in the living cells. This analysis revealed that informative miRNA are involved in several types of cancer and their corresponding target genes indeed share common roles in biological processes.

Acknowledgment

The authors would like to thank Dr. Chih-Jen Lin from National Taiwan University for providing the LIBSVM tool and Prof. Ian Witten from University of Waikato for providing the Weka software. This work was supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED Project No. 102.03.21.09).

References

[1] V. Ambros, "The functions of animal microRNAs," Nature, vol.431, pp.350–355, 2004.
[2] D.P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," Cell, vol.116, pp.281–297, 2004.
[3] P. Maziere and A.J. Enright, "Prediction of microRNA targets," Drug Discovery Today, vol.12, no.11, pp.452–459, 2007.
[4] D.P. Bartel, "MicroRNAs: Target regcognition and regulatory functions," Cell, vol.136, pp.215–233, 2009.
[5] N.D. Mendes, A.T. Freits, and M.F. Sagot, "Current tools for the identification of miRNA genes and their targets," Nucleic Acids Research, vol.37, no.8, pp.2419–2433, 2009.
[6] L. He and G.J. Hannon, "MicroRNAs: Small RNAs with a big role in gene regulation," Nature Review, vol.5, pp.522–531, 2004.
[7] Z. Baohong, P. Xiaoping, P.C. George, and A.A. Todd, "MicroRNAs as oncogenes and tumor suppressors," New Eng. J. Med., vol.353,

pp.1767–1771, 2007.
[8] T. Dalmay, "MicroRNA and cancer," J. Int. Med., vol.263, pp.1365–2796, 2008.
[9] W. Wei, S. Miao, Z. Gang-Ming, and C. Jianjun, "MicroRNA and cancer: Current status and prospective," Int. J. Cancer, vol.120, pp.953–960, 2006.
[10] A. Drakaki and D. Iliopoulos, "MicroRNA gene networks in oncogenesis," Current Genomics, vol.10, pp.35–41, 2009.
[11] G.A. Calin, C. Sevignani, C. Dan, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, and C.M. Croce, "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," Proc. Natl. Acad. Sci. USA, vol.101, pp.2999–3004, 2004.
[12] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller, "From signatures to models: Understangding cancer using microarray," Nat. Genet., vol.37, pp.s38–s45, 2005.
[13] Y. Zheng and C.K. Kwoh, "Cancer classification with microRNA expression patterns found by an information theory approach," J. Comput., vol.1, no.5, pp.30–39, 2005.
[14] S. Kim, S.J. Kim, and B.T. Zhang, "Evolving hypernetwork classifiers for microRNA expression profile analysis," Proc. IEEE Congress on Evolutionary Computation, pp.313–319, 2007.
[15] J. Lu, G. Getz, A.E. Miska, E. Alvarez-SAavedra, J. Lamb, D. Peck, A. Sweet-Cordero, L.B. Ebert, H.R. Mak, A.A. Ferrando, D.T. Jacks, H.R. Horvitz, and R.T. Golub, "MicroRNA expression profiles classify human cancers," Nature, vol.435, pp.834–838, 2005.
[16] K. Lange, Optimization, Springer Texts in Statistics, Springer-Verlag, 2004.
[17] N. Cristianini and J.S. Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[18] V. Vapnik, Statistical Learning Theory, Wiley, 1998.
[19] J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic, "Feature selection using support vector machines," Proc. 3rd Int. Conf. Data Mining Methods and Databases for Engineering, Finance and Other Fields, pp.261–273, 2002.
[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Selection for cancer classification using support vector machines," Mach. Learn., vol.46, no.1/3, pp.389–422, 2002.
[21] T.H. Pham, K. Satou, and T.B. Ho, "Support vector machines for prediction and analysis of beta and gamma-turns in proteins," J. Bioinf. Comput. Biol., vol.3, no.2, pp.343–358, 2005.
[22] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology. the gene ontology consortium," Nature Genetic, vol.25, pp.25–29, 2000.
[23] A. Krek, D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E. Epstein, P. MacMenamin, I. Piedade, K.C. Gunsalus, M. Stoffel, and N. Rajewsky, "Combinatorial microRNA target predictions," Nat. Genet., vol.37, pp.495–500, 2005.
[24] N. Rajewsky, "Microrna target predictions in animals," Nat. Genet., vol.38, pp.S8–S13, 2006.
[25] L.P. Lim, N.C. Clau, P. Garret-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Jonson, "Microarray analysis shows that some microRNAs downregulate large numbers of target mrnas," Nature, vol.433, pp.769–773, 2005.
[26] C.C. Chang and C.J. Lin, Libsvm, A library for support vector machines, http://www.csie.ntu.edu.tw/˜cjlin/papers/libsvm.pdf, 2001.
[27] I.H. Witten and E. Frank, Data Mining, Practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.
[28] N. Yanaihara, N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R.M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, G.A. Calin, C.G. Liu, C.M. Croce, and C.C. Harris, "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis," Cancer Cell, vol.9, no.3, pp.189–198, 2006.
[29] M.V. Iorio, M. Ferracin, C.G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Ménard, J.P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G.A. Calin, P. Querzoli, M. Negrini, and C.M. Croce, "MicroRNA gene expression deregulation in human breast cancer," Cancer Res., vol.65, pp.7065–7070, 2005.
[30] Y. Akao, Y. Nakagawa, and T. Naoe, "MicroRNAs 143 and 145 are possible common onco-micrornas in human cancers," Oncol Rep., vol.16, no.4, pp.845–850, 2006.
[31] Y. Xi, R. Shalgi, O. Fodstad, Y. Pil, and J. Ju, "Differentially regulated micro-RNAs and actively translated messenger RNA transcripts by tumor suppressor p53 in colon cancer," Clin Cancer Res., vol.12, pp.2014–2024, 2006.
[32] D. Grun, Y.L. Wang, D. Langenberger, K.C. Gunsalus, and N. Rajewsky, "MicroRNA target predictions across seven drosophila species and comparison to mammalian targets," PLoS Comput. Biol., vol.1, no.1, p.e13, 2005.
[33] T. Beissbarth and T.P. Speed, "Gostart: Find statistically overrepresented gene ontologies within a group of genes," Bioinformatics, vol.20, pp.1464–1465, 2004.
[34] J. Park, A.I. Meisler, and C.A. Cartwright, "c-yes tyrosine kinase activity in human colon carcinoma," Oncogene, vol.8, no.10, pp.2627–2635, 1993.

Dang Hung Tran received the BS in computer science from Hanoi National University of Education in 2001, MS in computer science from Vietnam National University, Hanoi in 2006, and PhD in knowledge science from Japan Advanced Institute of Science and Technology in 2009. He is currently a lecturer of Hanoi National University of Education. His research interests include machine learning and its applications on molecular biology.

Tu Bao Ho currently is a professor of School of Knowledge Science, Japan Advanced Institute of Science and Technology. He received a BT in applied mathematics from Hanoi University of Technology (1978), MS and PhD in computer science from Pierre and Marie Curie University, Paris (1984, 1987). His research interests include knowledge-based systems, machine learning, knowledge discovery and data mining.

Tho Hoan Pham currently is a lecturer of Hanoi National University of Education. He received the BS in mathematics from Hanoi National University of Education in 1993, MS in computer science from Hanoi University of Technology in 1997, and PhD in knowledge science from Japan Advanced Institute of Science and Technology in 2005. He is interested in machine learning, data mining, and bioinformatics.

Kenji Satou received the B.E., M.E., and D.E. degrees in computer science and communication engineering from Kyushu University, in 1987, 1989, and 1996, respectively. He is currently an associate professor of Kanazawa University, Ishikawa, Japan. His research interests include wide variety of topics in bioinformatics.