International Journal of Nanomedicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Rough sets for in silico identification of differentially expressed miRNAs

International Journal of Nanomedicine downloaded from https://www.dovepress.com/
For personal use only.

This article was published in the following Dove Press journal:
International Journal of Nanomedicine
13 September 2013
Number of times this article has been viewed

Sushmita Paul
Pradipta Maji

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

Abstract: The microRNAs, also known as miRNAs, are the class of small noncoding RNAs. They repress the expression of a gene posttranscriptionally. In effect, they regulate expression of a gene or protein. It has been observed that they play an important role in various cellular processes and thus help in carrying out normal functioning of a cell. However, dysregulation of miRNAs is found to be a major cause of a disease. Various studies have also shown the role of miRNAs in cancer and the utility of miRNAs for the diagnosis of cancer and other diseases. Unlike with mRNAs, a modest number of miRNAs might be sufficient to classify human cancers. However, the absence of a robust method to identify differentially expressed miRNAs makes this an open problem. In this regard, this paper presents a novel approach for in silico identification of differentially expressed miRNAs from microarray expression data sets. It integrates judiciously the theory of rough sets and merit of the so-called B.632+ bootstrap error estimate. While rough sets select relevant and significant miRNAs from expression data, the B.632+ error rate minimizes the variability and bias of the derived results. The effectiveness of the proposed approach, along with a comparison with other related approaches, is demonstrated on several miRNA microarray expression data sets, using the support vector machine.

Keywords: microRNA, feature selection, bootstrap error, support vector machine

Introduction

The microRNAs or miRNAs, a class of short, approximately 22-nucleotide, noncoding RNAs found in many plants and animals, often act posttranscriptionally to inhibit mRNA expression. Hence, the miRNAs are related to diverse cellular processes and are regarded as important components of the gene regulatory network. Multiple reports have noted the utility of miRNAs for the diagnosis of cancer and other diseases. Unlike with mRNAs, a modest number of miRNAs, 200 in total, might be sufficient to classify human cancers.[1,2] Moreover, the bead-based miRNA detection method has the attractive property of being not only accurate and specific, but also easy to implement in a routine clinical setting. In addition, unlike mRNAs, miRNAs remain largely intact in routinely collected, formalin-fixed, and paraffin-embedded specimens.[2] Recent studies have also shown that miRNAs can be detected in serum.[2] These studies offer the promise of utilizing miRNA screening via less invasive blood-based mechanisms. In addition, mature miRNAs are relatively stable. These phenomena make miRNAs superior molecular markers and targets for interrogation and as such, miRNA expression profiling can be utilized as a tool for cancer diagnosis and other diseases.

The functions of miRNAs have regulatory effects in various cellular functions. Just as miRNA is involved in the normal functioning of eukaryotic cells, so has dysregulation

Correspondence: Pradipta Maji
Machine Intelligence Unit, Indian Statistical Institute, 203, B T Road,
Kolkata, 700108, India
Email {sushmita_t,pmaji}@isical.ac.in

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/IJN.S40739

© 2013 Paul and Maji, publisher and licensee Dove Medical Press Ltd. This is an Open Access article which permits unrestricted noncommercial use, provided the original work is properly cited.

of miRNA been associated with disease.[3] This indicates that these miRNAs can prove to be potential biomarkers for developing a diagnostic tool. Hence, in silico identification of differentially expressed miRNAs that target genes involved in diseases is necessary. These differentially expressed miRNAs can be further used in developing effective diagnostic tools. Recently, a few studies were carried out to identify differentially expressed miRNAs.[4–8] However, the absence of a robust method of identification makes this an open problem. Hence, data sets are needed to be explored for understanding the complex biological activities of miRNAs.

A miRNA expression data set can be represented by an expression table or matrix, where each row corresponds to one particular miRNA, each column to a sample or time point, and each entry of the matrix is the measured expression level of a particular miRNA in a sample or time point, respectively. However, for microarray data, the number of training samples is typically very small, while the number of miRNAs is in the thousands. In general, it is not possible to use all available miRNAs to form the prediction rule of any classifier. Further, use of all the miRNAs might allow the noise associated with miRNAs of little or no discriminatory power. In effect, this would also inhibit and degrade the performance of the prediction rule in its application to unclassified or test samples. In other words, although the apparent error rate, which is the proportion of the training samples misclassified by the prediction rule, will decrease as it is formed from more and more miRNAs, its error rate in classifying samples outside of the training set eventually will increase. That is, the generalization error of the prediction rule will be increased if it is formed from a sufficiently large number of miRNAs. Hence, in practice, consideration has to be given to implement some procedure of feature selection for reducing the number of miRNAs to be used in constructing the prediction rule.[9]

The method called significance analysis of microarrays is used in several works[10–15] to identify differentially expressed miRNAs. Different statistical tests are also employed to identify differentially expressed miRNAs.[1,4–8,16–19] Xu et al[20] used the particle swarm optimization technique for selecting important miRNAs that contribute to the discrimination of different cancer types. However, the mutual information[21] or *f*-information[22]-based minimum redundancy-maximum relevance framework can also be used to select a set of nonredundant and relevant miRNAs for sample classification.

One of the main problems in miRNA expression data analysis is uncertainty. Some of the sources of this uncertainty include imprecision in computations and vagueness in class definition. In this context, the rough set theory has gained popularity in modeling and propagating uncertainty. It deals with vagueness and incompleteness and is proposed for indiscernibility in classification, according to some similarity.[23] It has been applied successfully to feature selection of discrete valued data.[24] Given a data set with discretized attribute values, it is possible to find a subset of the original attributes, using rough set theory, that are the most informative; all other attributes can be removed from the data set with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most useful in determining classifications from their values.[23,24] Rough set theory has been successfully applied to microarray data analysis.[25–34]

In general, the performance of the prediction rule generated by a classifier for a subset of selected miRNAs is evaluated by leave-one-out cross-validation (LOOCV) error. Given that the entire set of available samples is relatively small, in practice, one would like to make full use of all available samples in the miRNA selection and training of the prediction rule. But, if the LOOCV is calculated within the miRNA selection process, it has a selection bias when it is used as an estimate of the prediction error. The LOOCV error of the prediction rule obtained during the selection of the miRNAs provides a too optimistic estimate of the prediction error rate. Hence, an external cross-validation should be undertaken subsequent to the miRNA selection process to correct for this selection bias. Alternatively, the bootstrap procedure can be used.[35,36]

Although, the LOOCV error with external cross validation is nearly unbiased, it can be highly variable in the sense that there is no guarantee that the same subset of miRNAs will be obtained as during the original training of the rule, on all the training samples. Indeed, with the huge number of miRNAs available, it generally will yield a subset of miRNAs that has at most, only a few miRNAs in common with the subset selected during the original training of the rule. Suitably defined bootstrap procedures can reduce the variability of the LOOCV error in addition to providing a direct assessment of variability for the estimated parameters in the prediction rule. However, the bootstrap approach overestimates the error. To reduce the weakness of both these approaches, Efron and Tibshirani introduced the concept of B.632+ error for correcting the upward bias in bootstrap error with the downwardly biased apparent error,[35] which is very much applicable for the data sets with small number of training samples and large number of features or miRNAs.

In this regard, this paper presents a novel approach for in silico identification of differentially expressed miRNAs from expression data sets. It integrates the merit of the rough set–based feature-selection algorithm using a maximum relevance maximum significance criterion (RSMRMS)[29] and the concept of the so-called B.632+ error rate.[35] The RSMRMS algorithm selects a subset of miRNAs from a data set by maximizing both relevance and significance of the selected miRNAs. It employs rough set theory to compute both the relevance and significance of the miRNAs. Hence, the only information required in the feature selection method is in the form of equivalence partitions for each miRNA, which can be automatically derived from the given microarray data set. This avoids the need for domain experts to provide information on the data involved and ties in with the advantage of rough sets in that it requires no information other than the data set itself. On the other hand, the B.632+ error rate minimizes the variability and bias of the derived results. The support vector machine is used to compute the B.632+ error rate as well as several other types of error rates, as it maximizes the margin between data samples in different classes. The effectiveness of the proposed approach, along with a comparison with other related approaches, is demonstrated on a set of miRNA expression data sets.

The paper is organized as follows: The next section reports a brief description of several miRNA data sets used in the current study, along with the proposed methodology, which covers an overview of the rough sets, the rough set–based miRNA selection algorithm, fuzzy discretization method, the concept of the B.632+ error rate, and the support vector machine. Implementation details, experimental results, discussion, and a comparison among different algorithms are presented in the following section. Finally, concluding remarks are given.

Material and method

Data sets used

In the current research work, three publicly available miRNA expression data sets are used to establish the effectiveness of the proposed approach. Three miRNA expression data sets with accession numbers GSE17681, GSE17846, and GSE29352 were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). The first data set was generated to detect specific patterns of miRNAs in peripheral blood samples of lung cancer patients. As controls, blood of donors without known affection were tested. The number of miRNAs, samples, and classes in this data set are 866, 36, and two, respectively. The second data set represents the analysis of miRNA profiling in the peripheral blood samples of multiple sclerosis and in the blood of normal donors. It contains 864 miRNAs, 41 samples, and two classes. In the third data set, miRNA expression profiles in pancreatic cystic tumors with low malignant potential (serous microcystic adenomas) and high malignant potential (mucinous cystadenoma and intraductal papillary mucinous neoplasm [IPMN]) have been generated. These expression profiles are further compared in pancreatic ductal adenocarcinoma and carcinoma-ex-IPMN. The data set contains 43 samples, 885 miRNAs, and three classes.

Proposed method

The rough set–based proposed in silico approach is illustrated in Figure 1. It mainly consists of a rough set–based feature selection method (ie, RSMRMS), a support vector machine (SVM), and several types of error analysis parts, namely, apparent error (*AE*), bootstrap error (*B1*), no-information error (γ), and B.632+ error. This section presents each of these topics in detail, along with the basic notions of rough sets.

Rough sets

The theory of rough sets begins with the notion of an approximation space, which is a pair <$\mathbb{U}$, $\mathbb{A}$>, where $\mathbb{U} = \{x_1, \dots, x_\mathrm{i}, \dots, x_\mathrm{n}\}$ is a nonempty set, the universe of discourse, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. *V is* the value domain of A and *f* is an information function f: $\mathbb{U} \times \mathbb{A} \rightarrow V$. An approximation space is also called an information system.[23] Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence, also called indiscernibility, relation $IND(\mathbb{P})$ on $\mathbb{U}$

$$IND(\mathbb{P}) = \{(x_i,x_j) \in \mathbb{U} \times \mathbb{U} | \forall_a \in \mathbb{P}, f(x_{i,}a) = f(x_j,a)\}.$$

If $(x_i, x_j) \in IND(\mathbb{P})$, then x_i and x_j are indiscernible by attributes from $\mathbb{P}$. The partition of $\mathbb{U}$ generated by $IND(\mathbb{P})$ is denoted as

$$U/IND(\mathbb{P}) = \{[x_i]_{\mathbb{P}} : x_i \in U\}, \tag{1}$$

where $[x_i]_{\mathbb{P}}$ is the equivalence class containing x_i. The elements in $[x_i]_{\mathbb{P}}$ are indiscernible or equivalent with respect to knowledge $\mathbb{P}$. Equivalence classes, also termed as information granules, are used to characterize arbitrary subsets of $\mathbb{U}$. The equivalence classes of $IND(\mathbb{P})$ and the empty set ϕ are the elementary sets in the approximation space <$\mathbb{U}$, $\mathbb{A}$>.

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in <$\mathbb{U}$, $\mathbb{A}$>. One may

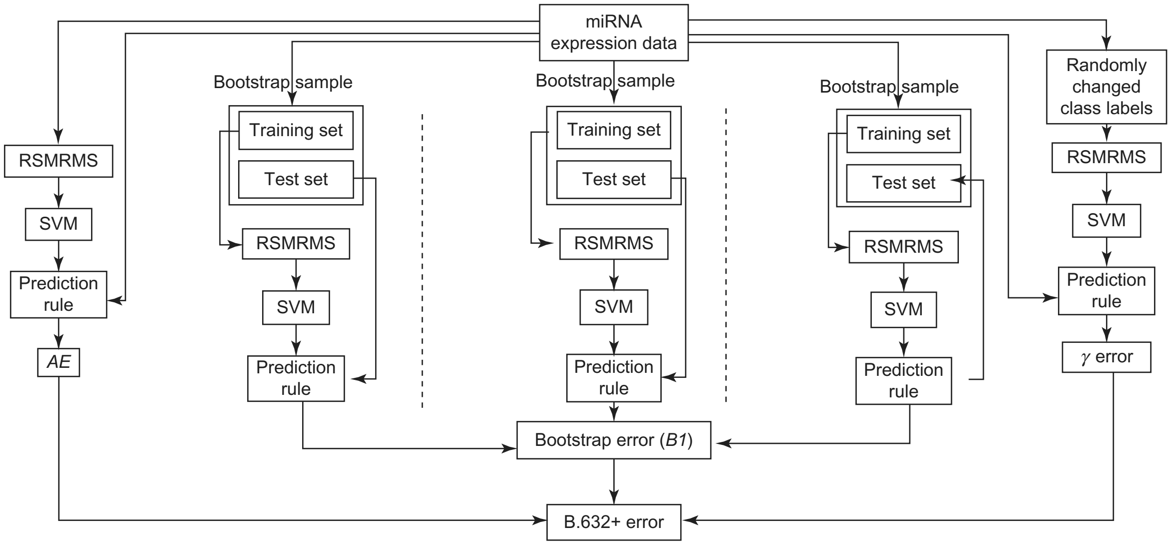

Figure 1 Schematic flow diagram of the proposed in silico approach for identification of differentially expressed miRNAs.
Abbreviations: miRNA, microRNA; RSMRMS, rough set–based maximum relevance maximum significance criterion; SVM, support vector machine; *AE*, apparent error; γ error, no-information error.

characterize X by a pair of lower and upper approximations, defined as follows:[23]

$$\underline{\mathbb{P}}(X) = \cup\{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \subseteq X\} \text{ and}$$
$$\overline{\mathbb{P}}(X) = \cup\left\{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \cap X \neq \phi\right\}. \tag{2}$$

Hence, the lower approximation $\underline{\mathbb{P}}(X)$ is the union of all the elementary sets which are subsets of X, and the upper $\overline{\mathbb{P}}(X)$ approximation is the union of all the elementary sets which have a nonempty intersection with X. The tuple $< \underline{\mathbb{P}}(X),\ \overline{\mathbb{P}}(X) >$ is the representation of an ordinary set X in the approximation space $<\mathbb{U}, \mathbb{A}>$ or simply called the rough set of X. The lower (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ (respectively, $\overline{\mathbb{P}}(X)$) is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called a positive region sometimes, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable or exact in $<\mathbb{U}, \mathbb{A}>$ if $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise X is indefinable and termed as a rough set.

Definition 1: An information system $<\mathbb{U}, \mathbb{A}>$ is called a decision table if the attribute set $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ is the condition attribute set and $\mathbb{D}$ is the decision attribute set. The dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as[23]

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}, \tag{3}$$

where $POS_{\mathbb{C}}(\mathbb{D}) = \cup \underline{\mathbb{C}} X_i$, X_i is the ith equivalence class induced by $\mathbb{D}$ and $|\cdot|$ denotes the cardinality of a set.

Definition 2: Given $\mathbb{C}$, $\mathbb{D}$ and an attribute $\mathcal{A} \in \mathbb{C}$, the significance of the attribute $\mathcal{A}$ is defined as[23]

$$\sigma_C(\mathbb{D}, \mathcal{A}) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-\{\mathcal{A}\}}(\mathbb{D}). \tag{4}$$

The change in dependency that arises when an attribute is removed from the set of condition attributes is a measure of the significance of the attribute. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable.

RSMRMS algorithm

In real data analysis such as microarray data, the data set may contain a number of insignificant features. The presence of such irrelevant and insignificant features may lead to a reduction in the useful information. Ideally, the selected features should have high relevance to the classes and high significance in the feature set. The features with high relevance are expected to be able to predict the classes of the samples. However, if insignificant features are present in the subset, they may reduce the prediction capability and may contain similar biological information. A feature set with high relevance and high significance enhances the predictive capability. Accordingly, a measure is required that can enhance the effectiveness of the feature set. In this work, the rough set theory is used to select the relevant and significant miRNAs from high-dimensional microarray data sets.

Let $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$ be the set of m miRNAs of a given microarray data set and $\mathbb{S}$ is the set of selected miRNAs. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the miRNA $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$, while $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)$ is the significance of the miRNA $\mathcal{A}_j$ with respect to the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The total relevance of all selected miRNAs is as follows:

$$\mathcal{J}_{relev} = \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}), \tag{5}$$

while the total significance among the selected miRNAs is

$$\mathcal{J}_{signf} = \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j), \tag{6}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of relevant and significant miRNAs from the whole set $\mathbb{C}$ of m miRNAs is equivalent to maximize both $\mathcal{J}_{relev}$ and $\mathcal{J}_{signf}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \mathcal{J}_{relev} + \beta \mathcal{J}_{signf}, \tag{7}$$

$$ie, \quad \mathcal{J} = \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}) + \beta \sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j), \tag{8}$$

where β is a weight parameter. To solve the above problem, a greedy algorithm is used.[29] The relevance and significance of individual miRNA are calculated based on the theory of rough sets, using Equations 3 and 4, respectively. The weight parameter β in the RSMRMS algorithm regulates the relative importance of the significance of the candidate miRNA with respect to the already-selected miRNAs and the relevance with the output class. If β is zero, only the relevance with the output class is considered for each miRNA selection. If β increases, this measure is incremented by a quantity proportional to the total significance, with respect to the already-selected miRNAs. The presence of a β value larger than zero is crucial in order to obtain good results. If the significance between miRNAs is not taken into account, selecting the miRNAs with the highest relevance with respect to the output class may tend to produce a set of redundant miRNAs that may leave out useful complementary information.

Fuzzy discretization

In miRNA expression data, the class labels of samples are represented by discrete symbols, while the expression values of miRNAs are continuous. Hence, to measure both relevance and significance of miRNAs using rough set theory, the continuous expression values of a miRNA have to be divided into several discrete partitions to generate equivalence classes.

In this regard, a fuzzy set–based discretization method is used to generate the equivalence classes required to compute both the relevance and significance of the miRNAs. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes. Given a finite set $\mathbb{U}$, $\mathbb{C}$ is a fuzzy condition attribute set in $\mathbb{U}$, which generates a fuzzy equivalence partition on $\mathbb{U}$. If c denotes the number of fuzzy equivalence classes generated by the fuzzy equivalence relation and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are sets of (cn) values $\{\mu_{ij}^{\mathbb{C}}\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $\mathbb{M}_{\mathbb{C}} = \left[\mu_{ij}^{\mathbb{C}}\right]$, which is denoted by

$$\mathbb{M}_{\mathbb{C}} = \begin{pmatrix} \mu_{11}^{\mathbb{C}} & \cdots & \mu_{1n}^{\mathbb{C}} \\ \cdots & \cdots & \cdots \\ \mu_{c1}^{\mathbb{C}} & \cdots & \mu_{cn}^{\mathbb{C}} \end{pmatrix}, \tag{9}$$

where $\mu_{ij}^{\mathbb{C}} \in [0, 1]$ represents the membership of object x_j in the ith fuzzy equivalence partition or class F_i.[27,37]

Each row of the matrix $\mathbb{M}_{\mathbb{C}}$ is a fuzzy equivalence partition or class. In the rough set–based feature selection method, the π function in one dimensional form is used to assign membership values to different fuzzy equivalence classes for the input miRNAs. A fuzzy set with membership function $\pi(x;\bar{c},\sigma)$ represents a set of points clustered around c, where

$$\pi(x;\bar{c},\sigma) = \begin{cases} 2\left(1 - \dfrac{\| x - \bar{c} \|}{\sigma}\right)^2 & \text{for } \dfrac{\sigma}{2} \le \| x - \bar{c} \| \le \sigma \\ 1 - 2\left(\dfrac{\| x - \bar{c} \|}{\sigma}\right)^2 & \text{for } 0 \le \| x - \bar{c} \| \le \dfrac{\sigma}{2} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $\sigma > 0$ is the radius of the π function with c as the central point and $\| \cdot \|$ denotes the Euclidean norm. When the pattern x lies at the central point c of a class, then $\|x - c\| = 0$ and its membership value is maximum, that is, $\pi(\bar{c};\bar{c},\sigma) = 1$. The membership value of a point decreases as its distance from the central point c (ie, $\|x - c\|$) increases. When $\|x - c\| = (\sigma/2)$, the membership value of x is 0.5, and this is called a crossover point.[38] The $(c \times n)$ matrix $\mathbb{M}_{\mathcal{A}i}$, corresponding to the ith miRNA $\mathcal{A}_i$, can be calculated from the c-fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$\mu_{k_j}^{\mathcal{A}_i} = \frac{\pi\left(x_j;\overline{c_k},\sigma_k\right)}{\sum_{l=1}^{c} \pi\left(x_j;\overline{c_l},\sigma_l\right)}. \tag{11}$$

In effect, each position $\mu_{k_j}^{A_i}$ of the matrix $\mathbb{M}_{\mathcal{A}i}$ must satisfy the following conditions:

$$\mu_{k_j}^{A_i} \in [0,1]; \sum_{k=1}^{c} \mu_{k_j}^{A_i} = 1, \forall_j \text{ and for any value of } k, \text{ if}$$

$$s = \arg\max_j \left\{\mu_{k_j}^{A_i}\right\}, \text{then } \max_j \left\{\mu_{k_j}^{A_i}\right\} = \max_l \left\{\mu_{ls}^{A_i}\right\} > 0.$$

After the generation of the matrix $\mathbb{M}_{\mathcal{A}i}$ corresponding to the miRNA $\mathcal{A}_i$, the object x_j is assigned to one of the c equivalence classes, based on the maximum value of memberships of the object in different equivalence classes that follows next:

$$x_j \in F_{\wp}; \text{ where } \wp = \arg\max_k \left\{\mu_{k_j}^{\mathcal{A}_i}\right\}.$$

Each input real valued miRNA in quantitative form can be assigned to different fuzzy equivalence classes in terms of membership values, using the π fuzzy set with appropriate $\bar{c}$ and σ. The centers and radii of the π functions along each miRNA axis are determined automatically from the distribution of the training patterns. In the proposed RSMRMS algorithm, three fuzzy equivalence classes ($c = 3$), namely, low, medium, and high are considered. These three equivalence classes correspond to underexpression, baseline, and overexpression of continuous valued miRNAs, respectively. Corresponding to the three fuzzy sets, low, medium, and high, the following relations hold:

$$\bar{c}_1 = \bar{c}_{low}(\mathcal{A}_i); \bar{c}_2 = \bar{c}_{medium}(\mathcal{A}_i); \bar{c}_3 = \bar{c}_{high}(\mathcal{A}_i);$$

$$\sigma_1 = \sigma_{low}(\mathcal{A}_i);\ \sigma_2 = \sigma_{medium}(\mathcal{A}_i);\ \sigma_3 = \sigma_{high}(\mathcal{A}_i).$$

The parameters $\bar{c}$ and σ of each π fuzzy set are computed according to the following procedure.[39] Let $\overline{m_i}$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith miRNA $\mathcal{A}_i$. Then $\overline{m_{i_l}}$ and $\overline{m_{i_k}}$ are defined as the mean along the ith miRNA of the objects having coordinate values in the range $[(\mathcal{A}_{imin}, m_i)$ and $(m_i, \mathcal{A}_{imax})]$, respectively, where $\mathcal{A}_{imax}$ and $\mathcal{A}_{imin}$ denote the upper and lower bounds of the dynamic range of miRNA $\mathcal{A}_i$ for the training set. For the three fuzzy sets, low, medium, and high, the centers and corresponding radii are computed as follows:

$$\bar{c}_{low}(\mathcal{A}_i) = \overline{m_{i_l}};\ \bar{c}_{medium}(\mathcal{A}_i) = \overline{m_i};\ \bar{c}_{high}(\mathcal{A}_i) = \overline{m_{i_k}}$$

$$\sigma_{low}(\mathcal{A}_i) = 2(\bar{c}_{medium}(\mathcal{A}_i) - \bar{c}_{low}(\mathcal{A}_i));$$
$$\sigma_{high}(\mathcal{A}_i) = 2(\bar{c}_{high}(\mathcal{A}_i) - \bar{c}_{medium}(\mathcal{A}_i));$$
$$\sigma_{medium}(\mathcal{A}_i) = \eta \times \frac{A}{B},$$

where $A = \left\{\sigma_{low}(A_i)\left(A_{i_{\max}} - c_{medium}(A_i)\right) + \sigma_{high}(A_i)\left(c_{medium}(A_i) - A_{i_{\min}}\right)\right\}; B = \left\{A_{i_{\max}} - A_{i_{\min}}\right\},$

where η is a multiplicative parameter controlling the extent of the overlapping. The distribution of the patterns or objects along each miRNA axis is taken into account while computing the corresponding centers and radii of the fuzzy sets. Also, the amount of overlap between the three fuzzy sets can be different along the different axes, depending on the distribution of the objects or patterns.

B.632+ error rate

In order to minimize the variability and bias of derived result, the so-called B.632+ bootstrap approach[35] is used, which is defined as follows:

$$\text{B.632+} = (1-\omega)AE + \omega B1, \tag{12}$$

where AE denotes the proportion of the original training samples misclassified, termed as apparent error rate, and $B1$ is the bootstrap error, defined as follows:

$$B1 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\Sigma_{k=1}^{M} I_{jk} Q_{jk}}{\Sigma_{k=1}^{M} I_{jk}}\right) \tag{13}$$

where n is the number of original samples and M is the number of bootstrap samples. If the sample x_j is not contained in the kth bootstrap sample, then $I_{jk} = 1$, otherwise 0. Similarly, if x_j is misclassified, $Q_{jk} = 1$, otherwise 0. The weight parameter ω is given by

$$\omega = \frac{0.632}{1 - 0.368r}; \tag{14}$$

$$\text{where} \quad r = \frac{B1 - AE}{\gamma - AE}; \tag{15}$$

$$\text{and} \quad \gamma = \sum_{i=1}^{k} p_i(1 - q_i); \tag{16}$$

where K is the number of classes, p_i is the proportion of the samples from the ith class, and q_i is the proportion of them assigned to the ith class. Also, γ is termed as the no-information error rate that would apply if the distribution of the class-membership label of the sample x_j did not depend on its feature vector.

SVM

In the current study, the SVM[40] is used to compute the B.632+ error rate. The SVM is a margin classifier that

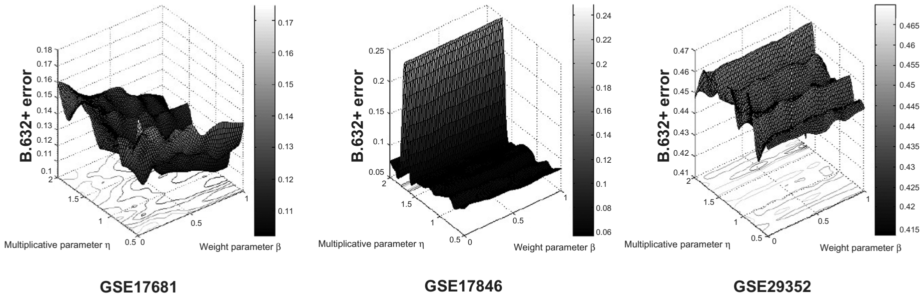

Figure 2 Variation of B.632+ error rate of the SVM with respect to multiplicative parameter η and weight parameter β.
Abbreviation: SVM, support vector machine.

draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between data samples in different classes, therefore leading to good generalization properties. A key factor in the SVM is to use kernels to construct a nonlinear decision boundary. In the present work, linear kernels are used. The source code of the SVM is downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Experimental results and discussions

In this section, the performance of the RSMRMS algorithm is compared with that of the mutual information based minimum redundancy-maximum relevance (mRMR) algorithm,[21] on three miRNA microarray data sets. The fuzzy set–based discretization method is also compared with several other discretization methods.[22,24] The margin classifier SVM[40] is used to evaluate the performance of different algorithms. To compute the different types of error rates obtained using the SVM, the bootstrap approach is performed on each miRNA expression data set. For each training set, a set of differential miRNAs is first generated, and then the SVM is trained with the selected miRNAs. After the training, the information of miRNAs, those selected for the training set, is used to generate a test set, and then the class label of the test sample is predicted using the SVM. For each data set, the 50 ($d = 50$) top-ranked miRNAs are selected for the analysis.

Optimum values of parameters

The rough set–based miRNA selection algorithm uses the weight parameter β to control the relative importance of significance of a miRNA with respect to its relevance. On the other hand, the multiplicative parameter η controls the degree of overlapping between the three fuzzy sets that are used to generate fuzzy equivalence classes. Hence, the performance of the proposed approach very much depends on both the parameters β and η.

The value of β is varied from 0.0 to 1.0, while the parameter η varies from 0.5 to 2.0. Extensive experimental results were obtained for all values of β and η on the three miRNA expression data sets. Figure 2 presents the variation of the B.632+ error rate obtained using the RSMRMS algorithm for different values of β and η on the three miRNA data sets. From the results reported in Figure 2, it is seen that as the value of β increases, the B.632+ error of the SVM decreases. On the other hand, the error rate increases for very high or very low values of η. Table 1 presents the optimum values of β and η for which the minimal B.632+ error rate of the SVM is achieved. From the results reported in Table 1, it is seen that the proposed algorithm with $\beta \neq 0.0$ provides a better result than that of $\beta = 0.0$, in all three cases, which justifies the importance of both the relevance and significance criteria. The corresponding values of η indicate that very large or very small amounts of overlapping among the three equivalence classes of input miRNAs are found to be undesirable for $\beta > 0.0$.

Importance of B.632+ error rate

This section establishes the importance of using the B.632+ error rate over other types of errors, such as *AE*, γ, and *B1*. Different types of errors on each miRNA expression data set are calculated using the SVM for the proposed method. Figure 3 represents the various types of errors obtained by the proposed algorithm on the three miRNA expression data sets. From Figure 3, it is seen that different types of errors decrease as the number of selected miRNAs increases. For all

Table 1 Optimum values of two parameters for three miRNA data

Parameter/data set	GSE17681	GSE17846	GSE29352
Weight parameter β	1.0	0.5	1.0
Multiplicative parameter η	1.7	1.0	1.7

Abbreviation: miRNA, microRNA.

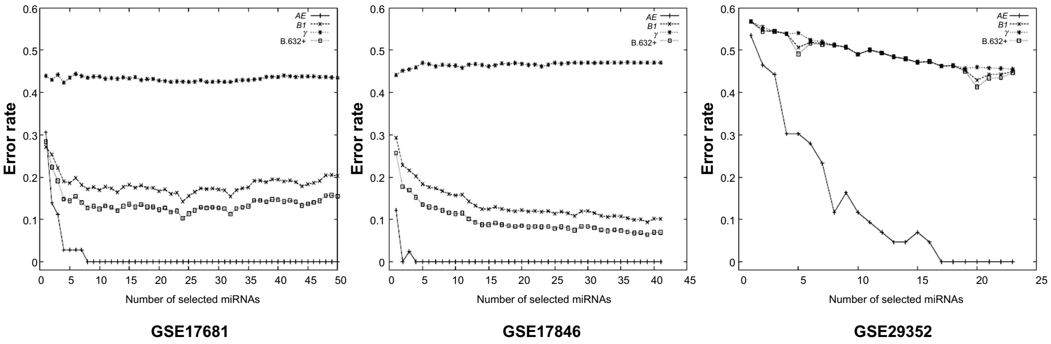

Figure 3 Error rate of the SVM obtained using the RSMRMS algorithm averaged over 50 random splits.
Abbreviations: SVM, support vector machine; RSMRMS, rough set–based maximum relevance maximum significance; miRNA, microRNA; *AE*, apparent error; *B1*, bootstrap error; γ, no-information error; B.632+, B.632+ error.

three data sets, the *AE* consistently attains the lowest value, while γ has highest value. On the other hand, the *B1* has a smaller error rate than γ but is higher than the *AE*. Moreover, the B.632+ estimate has smaller error rate than the *B1* but higher than the *AE*.

Table 2 reports the minimum values of different errors along with the number of required miRNAs to attain these values. From all the results reported in this table, it can be seen that the B.632+ estimator corrects the upward bias of *B1* and downward bias of *AE*. Also, it puts more weight on *B1* in the situation where the amount of overfitting, as measured by ($B1 - AE$), is relatively large. It thus is applicable in the present context where the prediction rule generated by the SVM is overfitted.

Role of fuzzy discretization method

In the current study, the fuzzy set–based discretization method was used to generate equivalence classes or information granules, for computing the relevance and significance of miRNAs using the theory of rough sets. To establish the effectiveness of the fuzzy set–based discretization method over other discretization methods, extensive experimentation was done on three miRNA data sets. The methods compared were the mean and standard deviation–based method,[22] the supervised discretization method,[24] and the unsupervised discretization method.[24] Figure 4 reports the variation of several errors with respect to number of selected miRNAs, while Table 3 presents the minimum error values obtained using the different discretization methods. From all the results reported in Figure 4 and Table 3, it can be seen that the fuzzy set–based discretion method performed better than the other discretization methods, irrespective of the types of errors and miRNA data sets used.

Table 2 Comparative analysis of different errors

Error/no of miRNA	Microarray data sets		
	GSE17681	GSE17846	GSE29352
AE	0.000	0.000	0.000
miRNA	8	2	17
B1 error	0.142	0.093	0.429
miRNA	24	39	20
γ error	0.423	0.441	0.455
miRNA	4	1	23
B.632+ error	0.103	0.064	0.413
miRNA	24	39	20

Abbreviations: *AE* error, apparent error; *B1* error, bootstrap error; γ error, no-information error; B.632+, B.632+ error.

Comparative performance analysis

This section compares the performance of the mRMR and RSMRMS algorithms with respect to the various types of errors. Figure 5 presents the different error rates obtained by the mRMR and RSMRMS algorithms on the three miRNA expression data sets. From the figure, it is seen that in most cases, the different types of error rates were consistently lower for the RSMRMS algorithm compared with the mRMR method.

Finally, Table 4 compares the performance of the rough set–based proposed method with the best performance of the mRMR method. The results are presented based on the error rate of the SVM classifier obtained on the three miRNA microarray data sets. From the results reported in Table 4, it is seen that although the best *AE* for each miRNA data set was same for both algorithms, the RSMRMS achieved this value with a lower number of selected miRNAs than that obtained by the mRMR method. Also, the RSMRMS attained

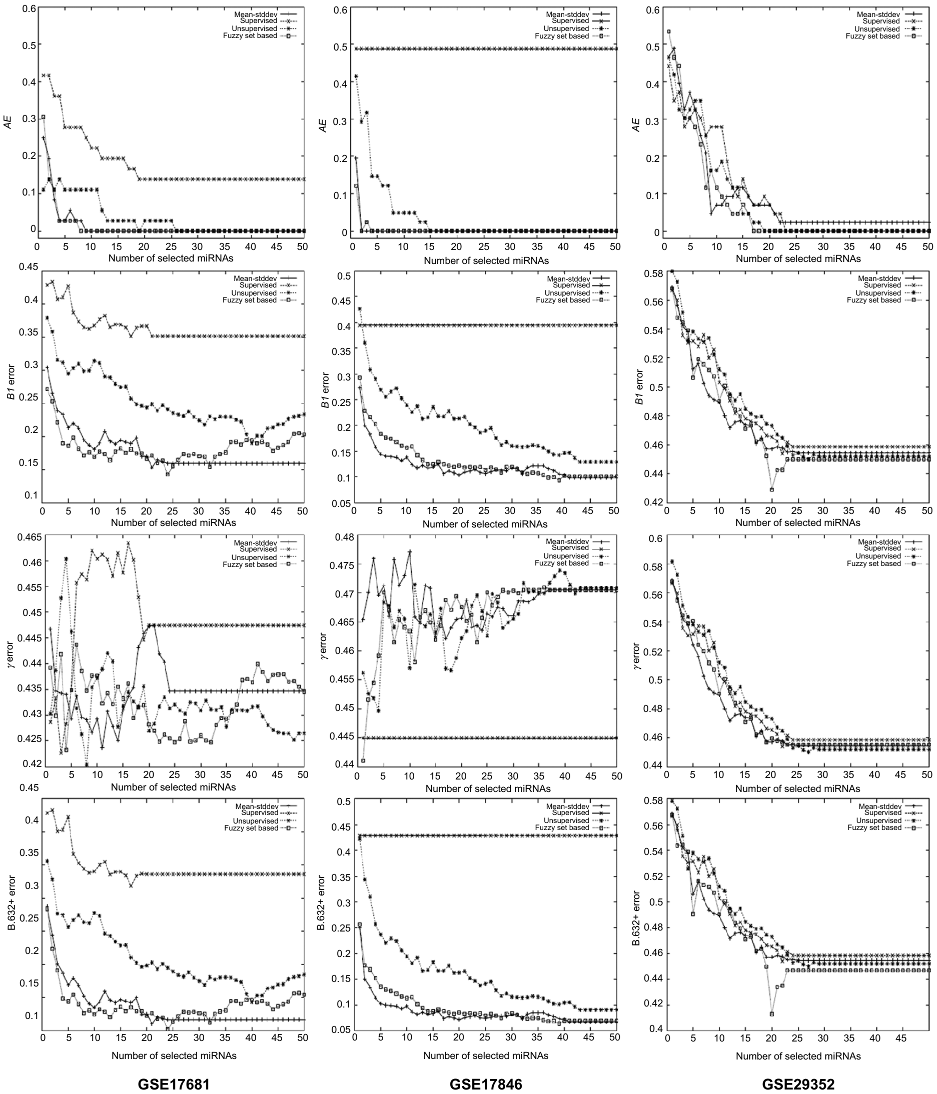

Figure 4 Error rates of the SVM obtained using different discretization methods averaged over 50 random splits.
Abbreviations: SVM, support vector machine; miRNA, microRNA; *AE*, apparent error; *B1*, bootstrap error; *γ*, no-information error; B.632+, B.632+ error.

Table 3 Comparative performance analysis of different discretization methods

Microarray data sets	Discretization methods	*AE*		*B1*		*γ*		*B.632+*	
		Error	miRNAs	Error	miRNAs	Error	miRNAs	Error	miRNAs
GSE17681	Mean-stddev	0.000	9	0.153	21	0.424	11	0.110	21
	Supervised	0.139	20	0.351	17	0.423	3	0.319	17
	Unsupervised	0.000	26	0.190	40	0.420	8	0.143	40
	Fuzzy set based	0.000	8	0.142	24	0.423	4	0.103	24
GSE17846	Mean-stddev	0.000	2	0.098	41	0.462	17	0.067	41
	Supervised	0.338	1	0.394	1	0.445	1	0.429	1
	Unsupervised	0.000	15	0.129	43	0.450	4	0.091	43
	Fuzzy set based	0.000	2	0.093	39	0.441	1	0.064	39
GSE29352	Mean-stddev	0.023	25	0.454	25	0.455	25	0.454	25
	Supervised	0.000	23	0.454	22	0.454	22	0.454	22
	Unsupervised	0.000	19	0.450	27	0.450	27	0.450	27
	Fuzzy set based	0.000	17	0.429	20	0.455	23	0.413	20

Abbreviations: *AE* error, apparent error; *B1* error, bootstrap error; *γ* error, no-information error; B.632+, B.632+ error; miRNA, microRNA.

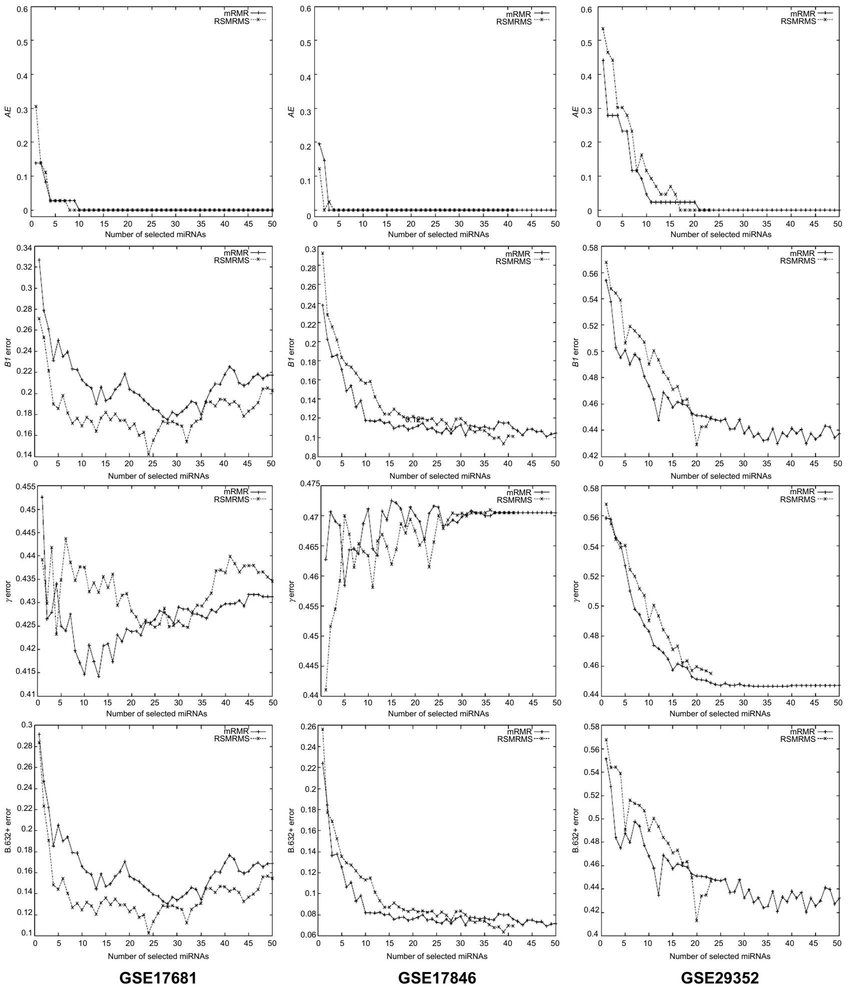

Figure 5 Error rates of the SVM obtained using the mRMR and RSMRMS algorithms averaged over 50 random splits.

Abbreviations: SVM, support vector machine; mRMR, mutual information–based minimum redundancy-maximum relevance; RSMRMS, rough set–based maximum relevance maximum significance; miRNA, microRNA; *AE*, apparent error; *B1*, bootstrap error; *γ*, no-information error; B.632+, B.632+ error.

the lowest B.632+ bootstrap error rate, as well as *B1* error rate, of the SVM classifier for all three miRNA data sets, with a lesser number of selected miRNAs.

The better performance of the RSMRMS algorithm was achieved due to the fact that it uses rough sets for computing both miRNA-class relevance and miRNA-miRNA significance to select differentially expressed miRNAs. The lower and upper approximations of rough sets can effectively deal with incompleteness, vagueness, and uncertainty of the data set.

Table 4 Comparative performance analysis of mRMR and RSMRMS algorithms

Microarray data sets	Methods/ algorithms	*AE*		***B1***		*γ*		**B.632+**	
		Error	**miRNAs**	**Error**	**miRNAs**	**Error**	**miRNAs**	**Error**	**miRNAs**
GSE17681	mRMR	0.000	10	0.175	28	0.414	13	0.130	28
	RSMRMS	0.000	8	0.142	24	0.423	4	0.103	24
GSE17846	mRMR	0.000	3	0.101	48	0.441	1	0.069	49
	RSMRMS	0.000	2	0.093	39	0.458	5	0.064	39
GSE29352	mRMR	0.000	21	0.430	43	0.447	32	0.420	43
	RSMRMS	0.000	17	0.429	20	0.455	23	0.413	20

Abbreviations: mRMR, mutual information–based minimum redundancy-maximum relevance; RSMRMS, rough set–based maximum relevance maximum significance; *AE*, apparent error, *B1*, bootstrap error; *γ*, no-information error; B.632+, B.632+ error; miRNA, microRNA.

Conclusion

This paper presents a novel approach for in silico identification of differentially expressed miRNAs. It integrates judiciously the merits of rough sets, SVM, and the B.632+ error rate for selecting relevant and significant miRNAs, which can classify samples into different classes with minimum error rate. The results obtained on three miRNA data sets demonstrate that the proposed method can bring a remarkable improvement to the miRNA selection problem, and therefore, can be a promising alternative to existing models for the prediction of class labels of samples. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. The new method is capable of identifying effective miRNAs that may contribute to revealing the underlying etiology of a disease, providing a useful tool for exploratory analysis of miRNA data.

Acknowledgment

This work is partially supported by the Indian National Science Academy, New Delhi (grant no SP/YSP/68/2012). The work was done when one of the authors, S Paul, was a Senior Research Fellow of the Council of Scientific and Industrial Research, Government of India. The authors would like to thank Dr Kuntal Ghosh of the Indian Statistical Institute, Kolkata, India for providing helpful comments and valuable criticisms on the original version of the manuscript.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature*. 2005;435(7043):834–838.
2. Budhu A, Ji J, Wang XW. The clinical potential of microRNAs. *J Hematol Oncol*. 2010;3(37).
3. Lehmann U, Streichert T, Otto B, et al. Identification of differentially expressed microRNAs in human male breast cancer. *BMC Cancer*. 2010;l0:109.
4. Blenkiron C, Goldstein LD, Thorne NP, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol*. 2007;8(10):R214.
5. Chen Y, Stallings RL. Differential patterns of microRNA expression in neuroblastoma are correlated with prognosis, differentiation, and apoptosis. *Cancer Res*. 2007;67(3):976–983.
6. Guo J, Miao Y, Xiao B, et al. Differential expression of microRNA species in human gastric cancer versus non-tumorous tissues. *J Gastroenterol Hepatol*. 2009;24(4):652–657.
7. Schrauder MG, Strick R, Schulz-Wendtland R, et al. Circulating microRNAs as potential blood-based markers for early stage breast cancer detection. *PLoS One*. 2012;7(1):e29770.
8. Zhao H, Shen J, Medico L, Wang D, Ambrosone CB, Liu S. A pilot study of circulating miRNAs as potential biomarkers of early stage breast cancer. *PLoS One*. 2010;5(10):e13735.
9. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*. 2002;99(10):6562–6566.
10. Iorio MV, Visone R, Di Leva G, et al. MicroRNA signatures in human ovarian cancer. *Cancer Res*. 2007;67(18):8699–8707.
11. Li S, Chen X, Zhang H, et al. Differential expression of microRNAs in mouse liver under aberrant energy metabolic status. *J Lipid Res*. 2009;50(9):1756–1765.
12. Nasser S, Ranade AR, Sridhart S, et al. Identifying miRNA and imaging features associated with metastasis of lung cancer to the brain. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine; November 1–4, 2009; Washington DC, USA. IEEE Conference Publications; 2009:246–251.
13. Ortega FJ, Moreno-Navarrete JM, Pardo G, et al. MiRNA expression profile of human subcutaneous adipose and during adipocyte differentiation. *PLoS One*. 2010;5(2):e9022.
14. Pereira PM, Marques JP, Soares AR, Carreto L, Santos MA. MicroRNA expression variability in human cervical tissues. *PLoS One*. 2010;5(7):e11780.
15. Raponi M, Dossey L, Jatkoe T, et al. MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. *Cancer Res*. 2009;69(14):5776–5783.
16. Arora S, Ranade AR, Tran NL, et al. MicroRNA-328 is associated with (non-small) cell lung cancer (NSCLC) brain metastasis and mediates NSCLC migration. *Int J Cancer*. 2011;129(11):2621–2631.
17. Dixon-McIver A, East P, Mein CA, et al. Distinctive patterns of microRNA expression associated with karyotype in acute myeloid leukaemia. *PLoS One*. 2008;3(5):e2141.
18. Wang C, Yang S, Sun G, et al. Comparative miRNA expression profiles in individuals with latent and active tuberculosis. *PLoS One*. 2011;6(10):e25832.
19. Zhu M, Yi M, Kim CH, et al. Integrated miRNA and mRNA expression profiling of mouse mammary tumor models identifies miRNA signatures associated with mammary tumor lineage. *Genome Biol*. 2011;12(8):R77.
20. Xu R, Xu J, Wunsch DC 2nd. MicroRNA expression profile based cancer classification using Default ARTMAP. *Neural Netw*. 2009;22(5–6):774–780.
21. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. 2005;3(2):185–205.
22. Maji P. f-Information measures for efficient selection of discriminative genes from microarray data. *IEEE Trans Biomed Eng*. 2009;56(4):1063–1069.
23. Pawlak Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer; 1991.
24. Maji P, Pal SK. *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. Hoboken: Wiley-IEEE Computer Society Press; 2012.
25. Fang J, Grzymala-Busse JW. Mining of microRNA expression data – a rough set approach. *Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology*; July 24–26, 2006; Chongquing, China. Berlin: Springer-Verlag; 2006:758–765.
26. Maji P. Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. *IEEE Trans Syst Man Cybern B Cybern*. 2011;41(1):222–233.
27. Maji P, Pal SK. Fuzzy-rough sets for information measures and selection of relevant genes from microarray data. *IEEE Trans Syst Man Cybern B Cybern*. 2010;40(3):741–752.
28. Maji P, Paul S. Microarray time-series data clustering using rough-fuzzy c-means algorithm. *Proceedings of the 5th IEEE International Conference on Bioinformatics and Biomedicine*; November 12–15, 2011; Atlanta, USA. IEEE Conference Publications; 2011:269–272.
29. Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. *Int J Approx Reason*. 2011;52(3):408–426.

30. Maji P, Paul S. Rough-fuzzy clustering for grouping functionally similar genes from microarray data. *IEEE/ACM Trans Comput Biol Bioinform*. Epub July 24, 2012.
31. Paul S, Maji P. Robust RFCM algorithm for identification of coexpressed miRNAs. *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*; October 4–7, 2012; Philadelphia, USA. IEEE Conference Publications; 2012:520–523.
32. Paul S, Maji P. Rough sets and support vector machine for selecting differentially expressed miRNAs. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops*; October 4–7, 2012; Philadelphia, USA. IEEE Conference Publications; 2012:864–871.
33. Slezak D, Wroblewski J. Roughfication of numeric decision tables: the case study of gene expression data. Proceedings of the 2nd International Conference on Rough Sets and Knowledge Technology; May 14–16, 2007; Toronto, Canada. Berlin: Springer-Verlag; 2007: 316–323.
34. Valdes JJ, Barton AJ. Relevant attribute discovery in high dimensional data: application to breast cancer gene expressions. *Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology*; July 24–26, 2006; Chongquing, China. Berlin: Springer-Verlag; 2006:482–489.
35. Efron B, Tibshirani R. Improvements on cross-validation: The B.632+ bootstrap method. *Journal of the American Statistical Association*. 1997;92(438):548–560.
36. Maji P, Das C. Relevant and significant supervised gene clusters for microarray cancer classification. *IEEE Trans Nanobioscience*. 2012; 11(2):161–168.
37. Maji P, Pal SK. Feature selection using f-information measures in fuzzy approximation spaces. *IEEE Trans Knowl Data Eng*. 2010;22(6): 854–867.
38. Pal SK, Pramanik PK. Fuzzy measures in determining seed points in clustering. *Pattern Recognit Lett*. 1986;4(3):159–164.
39. Pal SK, Mitra S. *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. New York: John Wiley & Sons; 1999.
40. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag; 1995.

International Journal of Nanomedicine

Dovepress

Publish your work in this journal

The International Journal of Nanomedicine is an international, peer-reviewed journal focusing on the application of nanotechnology in diagnostics, therapeutics, and drug delivery systems throughout the biomedical field. This journal is indexed on PubMed Central, MedLine, CAS, SciSearch®, Current Contents®/Clinical Medicine, Journal Citation Reports/Science Edition, EMBase, Scopus and the Elsevier Bibliographic databases. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: http://www.dovepress.com/international-journal-of-nanomedicine-journal