μHEM for identification of differentially expressed miRNAs using hypercuboid equivalence partition matrix

Sushmita Paul1,2 and Pradipta Maji1,2*

Abstract

Background: The miRNAs, a class of short approximately 22-nucleotide non-coding RNAs, often act post-transcriptionally to inhibit mRNA expression. In effect, they control gene expression by targeting mRNA. They also help in carrying out normal functioning of a cell as they play an important role in various cellular processes. However, dysregulation of miRNAs is found to be a major cause of a disease. It has been demonstrated that miRNA expression is altered in many human cancers, suggesting that they may play an important role as disease biomarkers. Multiple reports have also noted the utility of miRNAs for the diagnosis of cancer. Among the large number of miRNAs present in a microarray data, a modest number might be sufficient to classify human cancers. Hence, the identification of differentially expressed miRNAs is an important problem particularly for the data sets with large number of miRNAs and small number of samples.

Results: In this regard, a new miRNA selection algorithm, called μHEM, is presented based on rough hypercuboid approach. It selects a set of miRNAs from a microarray data by maximizing both relevance and significance of the selected miRNAs. The degree of dependency of sample categories on miRNAs is defined, based on the concept of hypercuboid equivalence partition matrix, to measure both relevance and significance of miRNAs. The effectiveness of the new approach is demonstrated on six publicly available miRNA expression data sets using support vector machine. The .632+ bootstrap error estimate is used to minimize the variability and biasedness of the derived results.

Conclusions: An important finding is that the μHEM algorithm achieves lowest .632+ error rate of support vector machine with a reduced set of differentially expressed miRNAs on four expression data sets compare to some existing machine learning and statistical methods, while for other two data sets, the error rate of the μHEM algorithm is comparable with the existing techniques. The results on several microarray data sets demonstrate that the proposed method can bring a remarkable improvement on miRNA selection problem. The method is a potentially useful tool for exploration of miRNA expression data and identification of differentially expressed miRNAs worth further investigation.

Keywords: MicroRNA; Feature selection, Rough hypercuboid, Bootstrap error, Support vector machine

Background

The microRNAs or miRNAs are small non-coding RNAs of length around 22 nucleotides, present in many plants and animals. They repress the expression of a gene post-transcriptionally. In effect, they regulate expression of a gene or protein. The miRNAs are related to diverse cellular processes and regarded as important components of gene regulatory network. Studies into miRNA function have mainly focused on a variety of human diseases, particularly cancer, and mainly related to the use of miRNAs as disease biomarkers and for monitoring drug efficacy. Multiple reports have noted the utility of miRNAs for the diagnosis of cancer and other diseases [1].

Unlike with mRNAs, a modest number of miRNAs might be sufficient to classify human cancers [1]. 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, paraffin-embedded clinical tissues [2]. Recent studies have also shown that miRNAs can be detected in serum. 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 appear to be different in various cellular functions. Just as miRNA is involved in the normal functioning of eukaryotic cells, so has dysregulation of miRNA been associated with disease [3]. It 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, few studies are carried out to identify differentially expressed miRNAs [4-9]. However, absence of robust method makes it an open problem.

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, and each entry of the matrix is the measured expression level of a particular miRNA in a sample, respectively. However, for microarray data, the number of training samples is typically very small, while the number of miRNAs is in the thousands. Hence, the prediction rule formed by any classifier may not be able to be formed by using all available miRNAs. Even if all the miRNAs can be used, the use of all the miRNAs allows the noise associated with miRNAs of little or no discriminatory power, which inhibits and degrades 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 [10].

The method called significance analysis of microarrays is used in several works [11-16] to identify differentially expressed miRNAs. Different statistical tests are also employed to identify differentially expressed miRNAs [1,4-8,17-20]. Xu et al. [21] used particle swarm optimization technique for selecting important miRNAs that contribute to the discrimination of different cancer types. However, 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 background, 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 [22]. It has been applied successfully to feature selection of discrete valued data [23]. 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. The theory of rough sets has also been successfully applied to microarray data analysis in [9,24-35].

However, the real life high dimensional microarray data set may contain a number of irrelevant and insignificant miRNAs [9]. The presence of such miRNAs may lead to a reduction in useful information and degrade the prediction capability. The selected miRNA subset should contain the miRNAs those have high relevance with the classes and high significance in the miRNA set. Such miRNAs are expected to be able to predict the classes of the samples. Accordingly, a measure is required that can assess the effectiveness of a miRNA set [9].

In microarray 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 [9]. However, the inherent error that exists in discretization process is of major concern in the computation of the dependency of real valued features. The rough hypercuboid approach of Wei et al. [36] is found to be suitable for numerical data sets.

In this regard, this paper presents a new miRNA selection method, termed as $\mu$HEM. It employs rough hypercuboid approach to provide a means by which real valued noisy data can be effectively reduced without the need for user-specified information. The proposed method selects a subset of miRNAs from whole miRNA set by maximizing both relevance and significance of the selected miRNAs. Using the concept of hypercuboid equivalence partition matrix, the degree of dependency is calculated for miRNAs, which is used to compute both relevance and significance of the miRNAs. Hence, the only information required in the proposed method is in the form of equivalence classes for each miRNA, which can be automatically derived from the data set. The concept of so-called $B.632+ \text{ error rate}$ [37] is used to minimize the
variability and biasedness 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 several miRNA expression data sets.

**Methods**

**Data sets used**

In the current research work, publicly available six miRNA expression data sets with accession number GSE17681, GSE17846, GSE21036, GSE24709, GSE28700, and GSE31408 are used, which are downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/).

**GSE17681**

This data set has been generated to detect specific patterns of miRNAs in peripheral blood samples of lung cancer patients. As controls, blood of donors without known affection have been tested. The number of miRNAs, samples, and classes in this data sets are 866, 36, and 2, respectively [38].

**GSE17846**

This data set represents the analysis of miRNA profiling in peripheral blood samples of multiple sclerosis and in the blood of normal donors. It contains 864 miRNAs, 41 samples, and 2 classes [39].

**GSE21036**

This data set contains miRNA expression profiles of 218 prostate tumors with primary or metastatic prostate cancer with a median of 5 years clinical follow-up. The number of miRNAs and samples are 373 and 141, respectively [40].

**GSE24709**

It analyzes peripheral miRNA blood profiles of patients with lung diseases. The miRNA expression profiling has been done for patients with lung cancer, chronic obstructive pulmonary disease, and normal controls. It contains total 863 miRNAs, 71 samples, and 3 classes.

**GSE28700**

This data set contains expression profiles of miRNAs from 22 paired gastric cancer and normal tissues. It contains total 44 samples and 470 miRNAs. The samples are grouped into 2 classes [41].

**GSE31408**

It analyzes miRNA expression profiles of cutaneous T-cell lymphomas and benign inflammation of skin. It consists of total 705 miRNAs, 148 samples, and 2 classes [42].
universe generated by an equivalence relation. Each row of the matrix \( \mathbb{H}(A_k) \) is a hypercuboid equivalence partition or class. Here \( h_{ij}(A_k) \in \{0, 1\} \) represents the membership of object \( x_j \) in the \( i \)th equivalence partition or class \( \beta_i \) satisfying following two conditions:

\[
1 \leq \sum_{j=1}^{n} h_{ij}(A_k) \leq n, \forall i; \quad (3)
\]

\[
1 \leq \sum_{i=1}^{c} h_{ij}(A_k) \leq c, \forall j. \quad (4)
\]

The above axioms should hold for every equivalence partition, which correspond to the requirement that an equivalence class is non-empty. However, in real data analysis, uncertainty arises due to overlapping class boundaries. Hence, such a granulation process does not necessarily result in a compatible granulation in the sense that every two class hypercuboids or intervals may intersect with each other. The intersection of two hypercuboids also forms a hypercuboid, which is referred to as implicit hypercuboid. The implicit hypercuboids encompass the misclassified samples or objects those belong to more than one classes. The degree of dependency of the decision attribute set or class label on the condition attribute set depends on the cardinality of the implicit hypercuboids. The degree of dependency increases with the decrease in cardinality. Hence, the degree of dependency of decision attribute on a condition attribute set is evaluated by finding the implicit hypercuboids that encompass misclassified objects. Using the concept of hypercuboid equivalence partition matrix, the misclassified objects of implicit hypercuboids can be identified based on the confusion vector defined next

\[
\mathbb{V}(A_k) = [v_1(A_k), \ldots, v_j(A_k), \ldots, v_n(A_k)] \quad (5)
\]

where \( v_j(A_k) = \min\{1, \sum_{i=1}^{c} h_{ij}(A_k) - 1\} \). \( (6) \)

According to the rough set theory, if an object \( x_j \) belongs to the lower approximation of any class \( \beta_i \), then it does not belong to the lower or upper approximations of any other classes and \( v_j(A_k) = 0 \). On the other hand, if the object \( x_j \) belongs to the boundary region of more than one classes, then it should be encompassed by the implicit hypercuboid and \( v_j(A_k) = 1 \). Hence, the hypercuboid equivalence partition matrix and corresponding confusion vector of the condition attribute \( A_k \) can be used to define the lower and upper approximations of the \( i \)th class \( \beta_i \) of the decision attribute set \( D \).

Let \( \beta_i \subseteq \cup, \beta_i \) can be approximated using only the information contained within \( A_k \) by constructing the \( A \)-lower and \( A \)-upper approximations of \( \beta_i \):

\[
\overline{A}(\beta_i) = \{x_j| h_{ij}(A_k) = 1 \text{ and } v_j(A_k) = 0\}; \quad (7)
\]

\[
\overline{A}(\beta_i) = \{x_j| h_{ij}(A_k) = 1 \text{ and } v_j(A_k) = 1\}; \quad (8)
\]

where equivalence relation \( A \) is induced from attribute \( A_k \).

\[
B_{A_k}(\beta_i) = \{x_j| h_{ij}(A_k) = 1 \text{ and } v_j(A_k) = 1\}. \quad (9)
\]

**Dependency**

Combining (1), (5), and (7), the dependency between condition attribute \( A_k \) and decision attribute \( D \) can be defined as follows:

\[
\gamma_{A_k}(D) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} h_{ij}(A_k) \cap [1 - v_j(A_k)], \quad (10)
\]

that is, \( \gamma_{A_k}(D) = 1 - \frac{1}{n} \sum_{j=1}^{n} v_j(A_k) \), \( (11) \)

where \( 0 \leq \gamma_{A_k}(D) \leq 1 \). If \( \gamma_{A_k}(D) = 1 \), \( D \) depends totally on \( A_k \); if \( 0 < \gamma_{A_k}(D) < 1 \), \( D \) depends partially on \( A_k \), and if \( \gamma_{A_k}(D) = 0 \), then \( D \) does not depend on \( A_k \). The \( \gamma_{A_k}(D) \) is also termed as the relevance of attribute \( A_k \) with respect to class \( D \).

**Significance**

Given two condition attributes \( A_k \) and \( A_l \), the \( c \times n \) hypercuboid equivalence partition matrix corresponding to the set \( \mathbb{H} = \{A_k, A_l\} \) can be calculated from two \( c \times n \) hypercuboid equivalence partition matrices \( \mathbb{H}(A_k) \) and \( \mathbb{H}(A_l) \) as follows:

\[
\mathbb{H}([A_k, A_l]) = \mathbb{H}(A_k) \cap \mathbb{H}(A_l); \quad (12)
\]

where \( h_{ij}([A_k, A_l]) = h_{ij}(A_k) \cap h_{ij}(A_l) \). \( (13) \)

The change in dependency when an attribute is removed from the set of condition attributes, is a measure of the significance of the attribute. To what extent an attribute is contributing to calculate the dependency on decision attribute can be calculated by the significance of that attribute. The significance of the attribute \( A_k \) with respect to the condition attribute set \( \{A_k, A_l\} \) is given by

\[
\sigma_{\mathbb{H}}(D, A_k) = \frac{1}{n} \sum_{j=1}^{n} [v_j(A_k) - v_j(A_k)]; \quad (14)
\]

where \( 0 \leq \sigma_{\mathbb{H}}(D, A_k) \leq 1 \). Hence, the higher the change in dependency, the more significant the attribute \( A_k \) is. If significance is 0, then the attribute is dispensable.

**\( \mu \)-HEM: proposed miRNA selection method**

Let \( \gamma_{A_k}(D) \) be the relevance of the miRNA \( A_k \) with respect to the class labels \( D \) and \( \sigma_{\mathbb{H}(A_k)}(D, A_k) \) is the significance of the miRNA \( A_k \) with respect to another miRNA \( A_l \) \( \in S \), where \( S \) is the set of selected miRNAs. The average relevance of all selected miRNAs is, therefore, given by

\[
J_{\text{relev}} = \frac{1}{|S|} \sum_{A_k \in S} \gamma_{A_k}(D), \quad (15)
\]
while the average significance among the selected miRNAs is as follows
\[
J_{\text{signf}} = \frac{1}{|S|(|S| - 1)} \times \sum_{A_i \neq A_j \in S} \{\sigma_{\{A_i\} \setminus \{A_j\}}(\mathbb{D}, A_i) + \sigma_{\{A_i\} \setminus \{A_j\}}(\mathbb{D}, A_j)\}.
\]

(16)

Therefore, the problem of selecting a set \( S \) of relevant and significant miRNAs from the whole miRNA set \( C \) is equivalent to maximize \( J_{\text{relev}} \) and \( J_{\text{signf}} \), that is, to maximize the objective function \( J \), where
\[
J = \omega J_{\text{relev}} + (1 - \omega) J_{\text{signf}}
\]

(17)

where \( \omega \) is a weight parameter. To solve the above problem, the following greedy algorithm is used.

1. Initialize \( C \leftarrow \{A_1, \ldots, A_{m}\}, S \leftarrow \emptyset \).
2. Generate hypercuboid equivalence partition matrix \( \mathbb{H}(\mathbb{A}_i) \) and corresponding confusion vector \( \mathbb{V}(\mathbb{A}_i) \) for each miRNA \( \mathbb{A}_i \in C \) using (1) and (5), respectively.
3. Calculate the relevance \( \gamma_{\mathbb{A}_i}(\mathbb{D}) \) of each miRNA \( \mathbb{A}_i \in C \) using (11).
4. Select the miRNA \( \mathbb{A}_i \) as the most relevant miRNA that has highest relevance value \( \gamma_{\mathbb{A}_i}(\mathbb{D}) \). In effect, \( \mathbb{A}_i \in S \) and \( C = C \setminus \mathbb{A}_i \).
5. Repeat the following two steps until \( C = \emptyset \) or the desired number of miRNAs is selected.
6. Repeat the following four steps for each of the remaining miRNAs of \( C \).

(a) Generate hypercuboid equivalence partition matrix \( \mathbb{H}([A_i, A_j]) \) using (12) between each selected miRNA \( A_i \in S \) and each miRNA \( A_j \in C \).
(b) Generate corresponding confusion vector \( \mathbb{V}([A_i, A_j]) \) for two miRNAs \( A_i \) and \( A_j \) using (5).
(c) Calculate the significance of each miRNA \( A_j \in C \) with respect to each of the already selected miRNAs of \( S \) using (14).
(d) Remove \( A_j \) from \( C \) if it has zero significance value with respect to any one of the selected miRNAs. In effect, \( C = C \setminus A_j \).
7. From the remaining miRNAs of \( C \), select miRNA \( A_j \) that maximizes the following condition:
\[
\omega \gamma_{A_i}(\mathbb{D}) + \frac{(1 - \omega)}{|S|} \sum_{A_i \in S \setminus \{A_j\}} \sigma_{\{A_i\} \setminus \{A_j\}}(\mathbb{D}, A_i).
\]

(18)

As a result of that, \( A_j \in S \) and \( C = C \setminus A_j \).
8. Stop.

**Computational complexity**

The proposed \( \mu \)HEM method has low computational complexity with respect to the number of miRNAs, samples, and classes. Prior to computing the relevance or significance of a miRNA, the hypercuboid equivalence partition matrix and confusion vector for each miRNA are to be generated first, which are carried out in Step 2 of the proposed algorithm. The computational complexity to generate a \((c \times n)\) hypercuboid equivalence partition matrix is \( O(cn) \), where \( c \) and \( n \) represent the number of classes and objects in the data set, respectively, while the generation of confusion vector has also \( O(cn) \) time complexity. In effect, the computation of the relevance of a miRNA has \( O(cn) \) time complexity. Hence, the total complexity to compute the relevance of \( m \) miRNAs, which is carried out in Step 3 of the proposed algorithm, is \( O(mcn) \). The selection of most relevant miRNA from the set of \( m \) miRNAs, which is carried out in Step 4, has a complexity \( O(m) \).

There is only one loop in Step 5 of the proposed miRNA selection method, which is executed \((d - 1)\) times, where \( d \) represents the number of selected miRNAs. The complexity to compute the significance of a candidate miRNA with respect to another miRNA has also the complexity \( O(cn) \). If \( m \) represents the cardinality of the already selected miRNA set, the total complexity to compute the significance of \((m - m)\) candidate miRNAs, which is carried out in Step 6, is \( O((m - m)cn) \). The selection of a miRNA from \((m - m)\) candidate miRNAs by maximizing relevance and significance, which is carried out in Step 7, has a complexity \( O(m - m) \). Hence, the total complexity to execute the loop \((d - 1)\) times is \( O((d - 1)((m - m) + (m - m)(cn)) = O((dcn(m - m))) \).

In effect, the selection of a set of \( d \) relevant and significant miRNAs from the whole set of \( m \) miRNAs using the proposed hypercuboid equivalence partition matrix based first order incremental search method has an overall computational complexity of \( O(mcn) + O(m) + O(dcn(m - m)) \) as \( c, m \ll m \).

**B.632+ error rate**

In order to minimize the variability and biasedness of derived result, the so-called B.632+ bootstrap approach [37] is used, which is defined as follows:

\[
B.632+ = (1 - \tilde{o})AE + \tilde{o}B1
\]

(19)

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:

\[
B1 = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{\sum_{k=1}^{M} I_{jk} Q_{jk}}{\sum_{k=1}^{M} I_{jk}} \right)
\]

(20)
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 \( i \)th bootstrap sample, then \( I_{j,k} = 1 \), otherwise \( 0 \). Similarly, if \( x_j \) is misclassified, \( Q_{j,k} = 1 \), otherwise \( 0 \). The weight parameter \( \omega \) is given by

\[
\omega = \frac{0.632}{1 - 0.368r}; \quad (21)
\]

where \( r = \frac{B1 - AE}{\gamma - AE} \);

and \( \gamma = \sum_{i=1}^{c} p_i (1 - q_i) \);

where \( c \) is the number of classes, \( p_i \) is the proportion of the samples from the \( i \)th class, and \( q_i \) is the proportion of them assigned to the \( i \)th class. Also, \( \gamma \) 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.

**Support vector machine**

In the current study, the support vector machine (SVM) [43] is used to evaluate the performance of the proposed \( \mu \)-HEM algorithm as well as several other feature selection algorithms. The SVM is a margin classifier that 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 nonlinear decision boundary. In the present work, linear kernels are used. The source code of the SVM has been downloaded from Library for Support Vector Machines (www.csie.ntu.edu.tw/~cjlin/libsvm/).

To compute different types of error rates obtained using the SVM, 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 were selected for the training set is used to generate test set and then the class label of the test sample is predicted using the SVM. For each data set, fifty top-ranked miRNAs are selected for the analysis.

In order to calculate the \( B.632+ \) error rate, apparent error (AE) is first calculated. This error is obtained when the same original data set is used to train and test a classifier. After that, the B1 error is computed from \( M \) bootstrap samples. Finally, the no-information error (\( \gamma \)) is calculated by randomly perturbing the class label of a given data set. The mutated data set is used for miRNA selection and the selected miRNA set is used to build the SVM. Then, the trained SVM is used to classify the original data set. The error generated by this procedure is known as \( \gamma \) rate. Finally, the B.632+ error rate is computed based on the AE, B1 error, and \( \gamma \) error using (19).

**Results and discussions**

The performance of the proposed hypercube equivalence partition matrix based miRNA selection (\( \mu \)-HEM) method is extensively studied and compared with that of some existing feature selection algorithms. The algorithms compared are mutual information based Info-Gain [44] and minimum redundancy-maximum relevance (mRMR) algorithm [45], method proposed by Golub et al. [46], rough set based maximum relevance-maximum significance (RSRMS) algorithm [9,28], boosting [47] and lasso [48]. The source code of the proposed \( \mu \)-HEM algorithm, written in C language, is available at www.isical.ac.in/~bibl/results/mihem/mihem.html. All the algorithms are run in Ubuntu 12.04 LTS having machine configuration Intel Core i7-2600 CPU @ 3.40GHz × 8, and 16 GB RAM.

**Performance analysis of \( \mu \)-HEM algorithm**

This section presents the performance of the proposed \( \mu \)-HEM algorithm on six miRNA data sets with respect to the \( B.632+ \) error rate of the SVM.

**Optimum value of weight parameter \( \omega \)**

The weight parameter \( \omega \) in (18) 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 \( \omega \) is one, only the relevance with the output class is considered for each miRNA selection. The presence of a \( \omega \) value lower than one 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 and insignificant miRNAs that may leave out useful complementary information. On the other hand, if \( \omega \) is zero, the miRNAs are selected based on their significance values only without considering the relevance of each miRNA. In effect, the selected miRNA set may contain a number of irrelevant miRNAs. Hence, the value of weight parameter \( \omega \) should be in between zero and one in order to obtain good results, that is, \( 0 < \omega < 1 \).

To find out the optimum value of \( \omega \) for each miRNA data set, the coefficient of variation (\( C_v \)) of average significance value is used. It is a measure of relative dispersion and defined as a quotient between standard deviation and mean value. Let the average significance value of the \( j \)th
selected miRNA set \( A_j \) with respect to the already selected miRNA set \( S_{j-1} \), for a given \( \omega \) value, be

\[
\Omega_j(\omega) = \frac{1}{|S_j|} \sum_{i \in A_j \in S_{j-1}} \sigma_i(\omega)(A_j, A_i)
\]

(24)

where \( D \) represents the set of class labels of the samples and \( S_j = S_{j-1} \cup \{ A_j \} \). If \( \mu(\omega) \) and \( s(\omega) \) represent the mean and standard deviation of the average significance values of \( d \) selected miRNAs for a particular value of \( \omega \), then the \( C_v \) index is defined as follows:

\[
C_v(\omega) = \frac{s(\omega)}{\mu(\omega)};
\]

(25)

where mean and standard deviation for \( d \) selected miRNAs are computed as follows:

\[
\mu(\omega) = \frac{1}{d} \sum_{j=1}^{d} \Omega_j(\omega);
\]

(26)

\[
s(\omega) = \frac{1}{\sqrt{d}} \left[ \sum_{j=1}^{d} \left( \frac{\mu(\omega) - \Omega_j(\omega)}{\sigma_j(\omega)} \right)^2 \right].
\]

(27)

The lower value of the \( C_v \) index, that is, the higher value of mean \( \mu \) and lower value of standard deviation \( s \), ensures that the average significance of the set of selected miRNAs is higher. A good miRNA selection method should make the value of \( C_v \) index as low as possible.

To find out the optimum value of \( \omega \), extensive experimentation is carried out on six miRNA expression data sets. The value of \( \omega \) is varied from 0.0 to 1.0. In the current study, \( d = 30 \) and \( d = 50 \) top-ranked miRNAs are selected for analysis. Figure 1 presents the variation of the \( C_v \) index obtained using the proposed \( \mu \)HEM algorithm for different values of \( \omega \) on six miRNA data sets. From the results reported in Figure 1, it is seen that as the value of weight parameter \( \omega \) increases, the \( C_v \) index decreases and attains its minimum value at a particular value of \( \omega = \omega^* \). After that the \( C_v \) index value increases with the increase in the value of \( \omega \). Hence, the optimum value of \( \omega \) for each data set is obtained using the following relation:

\[
\omega^* = \arg \min_{\omega} \{ C_v(\omega) \}.
\]

(28)

The optimum values of \( \omega \) obtained using (28) are 0.1 for GSE17681, GSE17846, GSE21036, GSE24709, and GSE28700, and 0.4 for GSE31408, irrespective of the number of selected miRNAs.

Figures 2 and 3 present the variation of the \( B.632+ \) error rate obtained using the proposed \( \mu \)HEM algorithm for different values of \( \omega \) on GSE17681, GSE17846, GSE21036, and GSE24709 data sets as examples considering \( d = 50 \). From the results reported in Figures 2 and 3, it is seen that the \( B.632+ \) error rate of the SVM decreases with the increase in the number of selected miRNAs, irrespective of the value of \( \omega \). Also, the error rate is lower for \( 0.0 < \omega < 0.5 \) than both \( \omega = 0.0 \) and 1.0. Similar results can also be seen for both GSE28700 and GSE31408 data sets.

Finally, Table 1 presents the minimal \( B.632+ \) error rate of the SVM for different values of weight parameter \( \omega \), along with the value of \( C_v \) index. For each miRNA data set, the minimum \( B.632+ \) error rate is written in italic, while the best \( C_v \) index is marked in bold. From the results reported in Table 1, it is seen that the proposed \( \mu \)HEM algorithm achieves its best performance at \( \omega = \omega^* \) in five cases out of total six miRNA data sets. Only for GSE28700 data set, the \( B.632+ \) error rate at \( \omega = \omega^* \) is higher than that of both \( \omega = 0.0 \) and 1.0. The lowest \( B.632+ \) error rate is achieved at \( \omega = 1.0 \) for this data set. All the results reported in Figures 1, 2, and 3, and Table 1 establish the importance of both relevance and significance criteria in the proposed \( \mu \)HEM method for selecting differentially expressed miRNAs from a microarray data.

![Figure 1 Variation of the \( C_v \) index for different values of weight parameter \( \omega \).](image-url)
Optimum number of selected miRNAs

According to Lu et al. [1], unlike with mRNAs, a modest number of miRNAs might be sufficient to classify human cancers. Also, the number of training samples is typically very small compared to the number of miRNAs. Hence, the use of large numbers of miRNAs in constructing classifier may degrade the prediction capability on test samples [10].

In order to find out the optimum number of selected miRNAs, extensive experimentation is carried out on six microarray data sets. Figure 4 depicts the relevance and average significance values of each of the selected miRNAs for six expression data sets. The results are presented for optimum values of $\omega$ considering 100 selected miRNAs. From the results reported in Figure 4, it can be seen that as the number of selected miRNAs increases, both relevance and significance values decrease. Also, the significance value remains constant after selecting forty to forty-five miRNAs, irrespective of the data sets used. Hence, in the current study, the selected number of miRNAs is set to $d = 50$.

Error rate and execution time

Figure 5 presents the variation of several error rates obtained using the proposed $\mu$HEM algorithm for different number of samples. The data sets in $x$-axis of Figure 5 are arranged in ascending order of the number of samples present in each data set, that is, the number of samples in GSE17681, GSE17846, GSE28700, GSE24709, GSE21036, and GSE31408 data are 36, 41, 44, 71, 141, and 148, respectively.

From all the results reported in Figure 5, it is seen that different error rates such as $AE$, $B1$, and $B.632+$ do not depend on the number of samples present in the data set,
rather, they depend on the distribution of the samples in different classes or categories. For example, although the number of samples in GSE17846 and GSE28700 data sets is almost equal, that is, 41 and 44, respectively, there is a significant difference in errors for these two data sets. The $B.632+$ errors for GSE17846 and GSE28700 data sets are 0.059 and 0.257, respectively. On the other hand, the $B.632+$ errors for GSE17846 data set with 41 samples and GSE31408 data set with 148 samples are 0.059 and 0.067, respectively.

Table 1 Performance of $\mu$HEM algorithm on six miRNA data sets for different values of $\omega$

| Value of $\omega$ | GSE17681 | GSE17846 | GSE21036 | GSE24709 | GSE28700 | GSE31408 |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| $B.632+$         | $C_r$     | $B.632+$  | $C_r$     | $B.632+$  | $C_r$     | $B.632+$  | $C_r$     | $B.632+$  | $C_r$     | $B.632+$  | $C_r$     |
| 0.0              | 0.0854    | 0.4951    | 0.0605    | 0.4275    | 0.0403    | 0.6528    | 0.1863    | 0.2312    | 0.2498    | 0.3388    | 0.0757    |
| 0.1              | 0.0842    | 0.4421    | 0.0590    | 0.4042    | 0.0388    | 0.5956    | 0.1803    | 0.2171    | 0.2566    | 0.2693    | 0.0753    |
| 0.2              | 0.0870    | 0.4502    | 0.0623    | 0.4094    | 0.0396    | 0.6124    | 0.1898    | 0.2213    | 0.2660    | 0.2752    | 0.0742    |
| 0.3              | 0.0851    | 0.4542    | 0.0644    | 0.4148    | 0.0410    | 0.6246    | 0.1878    | 0.2256    | 0.2572    | 0.2818    | 0.0732    |
| 0.4              | 0.0894    | 0.4611    | 0.0627    | 0.4206    | 0.0420    | 0.6319    | 0.1881    | 0.2312    | 0.2583    | 0.2889    | 0.0672    |
| 0.5              | 0.0882    | 0.4680    | 0.0640    | 0.4275    | 0.0394    | 0.6384    | 0.1970    | 0.2399    | 0.2587    | 0.2980    | 0.0690    |
| 0.6              | 0.0882    | 0.4951    | 0.0651    | 0.4319    | 0.0392    | 0.6447    | 0.1940    | 0.2429    | 0.2571    | 0.3079    | 0.0693    |
| 0.7              | 0.0893    | 0.5105    | 0.0637    | 0.4337    | 0.0402    | 0.6493    | 0.1951    | 0.2536    | 0.2632    | 0.3241    | 0.0683    |
| 0.8              | 0.0893    | 0.5202    | 0.0636    | 0.4366    | 0.0405    | 0.6528    | 0.1992    | 0.2564    | 0.2649    | 0.3388    | 0.0690    |
| 0.9              | 0.0893    | 0.5202    | 0.0636    | 0.4380    | 0.0398    | 0.6664    | 0.2002    | 0.2564    | 0.2650    | 0.3388    | 0.0697    |
| 1.0              | 0.0860    | 0.5958    | 0.0724    | 0.4575    | 0.0410    | 0.6801    | 0.2095    | 0.2950    | 0.2473    | 0.4191    | 0.0693    |

![Figure 4](http://www.biomedcentral.com/1471-2105/14/266)

Figure 4 Relevance and significance values of each of the selected miRNAs for different miRNA data sets.
optimum values of $\omega$ considering $d = 50$. Figures 7 and 8 represent various types of errors obtained by the proposed algorithm on GSE17681, GSE17846, GSE21036, and GSE24709 data sets as examples. From Figures 7 and 8, it is seen that different types of errors decrease as the number of selected miRNAs increases. Similar results are also found for both GSE28700 and GSE31408 data sets. For all six data sets, the $AE$ attains consistently lowest value, while $\gamma$ has highest value. On the other hand, the $B_1$ has smaller error rate than $\gamma$ but it is higher than the $AE$. Moreover, the $B_{6.32+}$ estimate has smaller error rate than the $B_1$ but higher than the $AE$.

Table 2 reports the minimum values of different errors, along with the number of miRNAs required to attain these values. From all the results reported in this table, it can be seen that the $B_{6.32+}$ estimator corrects the upward bias of $B_1$ and downward bias of $AE$. Also, it puts more weight on $B_1$ in situation where the amount of overfitting as measured by $(B_1 - AE)$ is relatively large. It thus is applicable in the present context where the prediction rule generated by the SVM may be overfitted.

**Comparative performance analysis**

This section compares the performance of the proposed $\mu$HEM algorithm with that of InfoGain [44], mRMR algorithm [45], method proposed by Golub et al. [46], RSMRMS algorithm [9], boosting [47], and lasso [48]. Table 3 and Figures 9, 10, 11, 12, 13, and 14 present different error rates obtained by various feature selection algorithms on six miRNA expression data sets.

**AE and $B_1$ error**

Table 3 compares the best performance of different feature selection algorithms based on the error rate of the SVM. From the results reported in Table 3, it is seen that the best $AE$ for each miRNA data set is same for most of the algorithms. Both proposed $\mu$HEM algorithm and mRMR method attain the best $AE$ value for all data sets, while the method proposed by Golub et al. and InfoGain achieve it for five data sets and boosting and RSMRMS method attain this value on two data sets. However, the $\mu$HEM achieves the best $AE$ value with lower number of selected miRNAs than that obtained by other methods on GSE17681, GSE17846, and GSE24709 data sets, while mRMR method attains it for GSE21036 and GSE28700 data sets and the method proposed by Golub et al. on GSE31408 data set. On the other hand, the boosting method attains lowest $B_1$ error rate in four cases out of total six data sets, while the $\mu$HEM method and lasso achieve it only for GSE21036 and GSE31408 data sets, respectively.

**Gap estimate**

However, according to Efron and Tibshirani [37], the bootstrap approach ($B_1$) overestimates the error. In this regard, the Gap function [49] is generally used to know whether the obtained $B_1$ error is smaller than that would be expected by chance, if the distribution of the class-membership label of the sample did not depend on its feature vector. The Gap function represents the difference between no-information error ($\gamma$) and bootstrap error ($B_1$), and is defined by

$$\text{Gap} = \gamma - B_1. \quad (29)$$

The larger value of Gap function indicates that the obtained or observed $B_1$ error is significantly lower than that of expected by chance. Figures 9, 10, and 11 depict the gap curves, which highlight the difference between $\gamma$ and $B_1$ errors obtained using different algorithms on six miRNA data sets. From the results reported in these figures, it can be found that the Gap estimate increases with the increase in the number of selected miRNAs,
irrespective of the algorithms and data sets used. Also, the Gap function always achieves significantly higher values for the proposed μHEM algorithm, while for both boosting and lasso, the gap estimate is very low. Table 3 compares the best values of the Gap function obtained using different algorithms. All the results reported here confirm that the proposed algorithm attains highest values of Gap function in five cases, while the method proposed by Golub et al. achieves it only for GSE31408 data set.

Table 2 Comparative analysis of different types of errors for μHEM algorithm

| Microarray data sets | Error | miRNAs | Error | miRNAs | Error | miRNAs | Error | miRNAs |
|----------------------|-------|--------|-------|--------|-------|--------|-------|--------|
| GSE17681             | 0.000 | 5      | 0.120 | 50     | 0.432 | 18     | 0.084 | 50     |
| GSE17846             | 0.000 | 2      | 0.087 | 49     | 0.469 | 5      | 0.059 | 49     |
| GSE21036             | 0.000 | 42     | 0.058 | 47     | 0.391 | 3      | 0.039 | 47     |
| GSE24709             | 0.000 | 20     | 0.234 | 49     | 0.462 | 22     | 0.180 | 49     |
| GSE28700             | 0.000 | 25     | 0.306 | 4      | 0.463 | 37     | 0.257 | 4      |
| GSE31408             | 0.000 | 44     | 0.098 | 2      | 0.383 | 10     | 0.067 | 50     |

Figure 7 Different error rates of the proposed algorithm on GSE17681 and GSE17846 data sets obtained using the SVM averaged over 50 random splits.

Figure 8 Different error rates of the proposed algorithm on GSE21036 and GSE24709 data sets obtained using the SVM averaged over 50 random splits.
Table 3 Comparative performance analysis of different algorithms

| Microarray data sets | Algorithms /Methods | Apparent error | B1 Error | Gap estimate | B.632+ Error |
|----------------------|---------------------|----------------|----------|--------------|--------------|
|                      | Error   | miRNAs | Error   | miRNAs | Error   | miRNAs | Error   | miRNAs |
| GSE17681             | Golub et al. | 0.000 | 19   | 0.194 | 32   | 0.258 | 32   | 0.146 | 32   |
|                      | Lasso    | 0.056 | 2    | 0.266 | 2    | 0.125 | 2    | 0.229 | 2    |
|                      | Boosting | 0.000 | 5    | 0.113 | 10   | 0.090 | 10   | 0.094 | 10   |
|                      | InfoGain | 0.000 | 6    | 0.154 | 21   | 0.290 | 21   | 0.111 | 21   |
|                      | mRMR     | 0.000 | 10   | 0.175 | 28   | 0.267 | 28   | 0.129 | 28   |
|                      | RSMRMS   | 0.000 | 8    | 0.142 | 24   | 0.299 | 24   | 0.102 | 24   |
|                      | μHEM     | 0.000 | 5    | 0.120 | 50   | 0.325 | 50   | 0.084 | 50   |
| GSE17846             | Golub et al. | 0.000 | 6    | 0.116 | 48   | 0.363 | 48   | 0.081 | 48   |
|                      | Lasso    | 0.024 | 3    | 0.102 | 3    | 0.241 | 3    | 0.079 | 3    |
|                      | Boosting | 0.000 | 4    | 0.037 | 9    | 0.170 | 9    | 0.025 | 9    |
|                      | InfoGain | 0.000 | 7    | 0.093 | 37   | 0.387 | 37   | 0.063 | 37   |
|                      | mRMR     | 0.000 | 3    | 0.101 | 48   | 0.379 | 48   | 0.069 | 48   |
|                      | RSMRMS   | 0.000 | 2    | 0.093 | 39   | 0.386 | 39   | 0.064 | 39   |
|                      | μHEM     | 0.000 | 2    | 0.087 | 49   | 0.392 | 49   | 0.059 | 49   |
| GSE21036             | Golub et al. | 0.000 | 35   | 0.069 | 48   | 0.368 | 48   | 0.047 | 48   |
|                      | Lasso    | 0.043 | 5    | 0.061 | 6    | 0.074 | 6    | 0.057 | 6    |
|                      | Boosting | 0.099 | 3    | 0.107 | 3    | 0.074 | 3    | 0.104 | 3    |
|                      | InfoGain | 0.000 | 39   | 0.073 | 50   | 0.372 | 44   | 0.049 | 50   |
|                      | mRMR     | 0.000 | 19   | 0.064 | 49   | 0.376 | 50   | 0.043 | 49   |
|                      | RSMRMS   | 0.050 | 5    | 0.089 | 5    | 0.328 | 5    | 0.075 | 5    |
|                      | μHEM     | 0.000 | 42   | 0.058 | 47   | 0.386 | 47   | 0.039 | 47   |
| GSE24709             | Boosting | 0.099 | 8    | 0.211 | 8    | 0.057 | 8    | 0.192 | 8    |
|                      | InfoGain | 0.000 | 26   | 0.257 | 45   | 0.218 | 46   | 0.203 | 45   |
|                      | mRMR     | 0.000 | 24   | 0.245 | 50   | 0.229 | 50   | 0.191 | 50   |
|                      | RSMRMS   | 0.141 | 11   | 0.402 | 11   | 0.123 | 2    | 0.366 | 11   |
|                      | μHEM     | 0.000 | 20   | 0.234 | 49   | 0.241 | 49   | 0.180 | 49   |
| GSE28700             | Boosting | 0.023 | 7    | 0.191 | 8    | 0.131 | 4    | 0.160 | 8    |
|                      | InfoGain | 0.000 | 35   | 0.309 | 8    | 0.159 | 8    | 0.271 | 21   |
|                      | mRMR     | 0.000 | 21   | 0.333 | 49   | 0.140 | 7    | 0.285 | 49   |
|                      | RSMRMS   | 0.023 | 34   | 0.331 | 19   | 0.140 | 15   | 0.285 | 19   |
|                      | μHEM     | 0.000 | 25   | 0.306 | 4    | 0.194 | 4    | 0.257 | 4    |
| GSE31408             | Golub et al. | 0.000 | 36   | 0.073 | 1    | 0.364 | 1    | 0.069 | 1    |
|                      | Lasso    | 0.061 | 3    | 0.072 | 4    | 0.184 | 1    | 0.068 | 4    |
|                      | Boosting | 0.081 | 2    | 0.087 | 2    | 0.085 | 1    | 0.085 | 2    |
|                      | InfoGain | 0.007 | 20   | 0.090 | 9    | 0.331 | 1    | 0.077 | 27   |
|                      | mRMR     | 0.000 | 37   | 0.094 | 6    | 0.331 | 1    | 0.074 | 6    |
|                      | RSMRMS   | 0.061 | 2    | 0.086 | 6    | 0.336 | 2    | 0.077 | 6    |
|                      | μHEM     | 0.000 | 44   | 0.098 | 2    | 0.354 | 50   | 0.067 | 50   |
Figure 9 Gap curve obtained using different methods on GSE17681 and GSE17846 data sets averaged over 50 random splits.

Figure 10 Gap curve obtained using different methods on GSE21036 and GSE24709 data sets averaged over 50 random splits.

Figure 11 Gap curve obtained using different methods on GSE28700 and GSE31408 data sets averaged over 50 random splits.
Figure 12 $B.632^+$ errors of the SVM obtained using different methods on GSE17681 and GSE17846 data sets averaged over 50 random splits.

Figure 13 $B.632^+$ errors of the SVM obtained using different methods on GSE21036 and GSE24709 data sets averaged over 50 random splits.

Figure 14 $B.632^+$ errors of the SVM obtained using different methods on GSE28700 and GSE31408 data sets averaged over 50 random splits.
**Biological significance analysis**

This section presents the biological significance of some miRNAs those are selected by the proposed \( \mu \)HEM algorithm for GSE21036 data set as an example. The manually curated database, termed as miR2Disease [50], is used here to biologically validate the results obtained by the \( \mu \)HEM algorithm. This database aims at providing a comprehensive resource of miRNA deregulation in various human diseases.

In GSE21036 data set, miRNA expression profiling has been done to understand the role of miRNAs that are responsible for the genesis and progression of prostate cancer [40]. The \( \mu \)HEM algorithm selects a set of differentially expressed miRNAs from each bootstrap sample of GSE21036 data set. A set of nine miRNAs, consisting of hsa-miR-145, hsa-miR-25, hsa-miR-153, hsa-miR-143, hsa-miR-19a, hsa-miR-96, hsa-miR-663, hsa-miR-20a, and hsa-miR-182, is identified from all bootstrap samples of GSE21036 data set. Among them, four miRNAs, namely, hsa-miR-19a, hsa-miR-20a, hsa-miR-663, and hsa-miR-182, are identified by the \( \mu \)HEM algorithm only, not by other feature selection algorithms.

One of the distinct characteristics of prostate cancer is over-expression of the ERG proto-oncogene. Several independent target prediction methods have indicated that the 3′ untranslated region of the ERG mRNA is a potential target of hsa-miR-145. The hsa-miR-145 is consistently down-regulated in prostate cancer. In [51], it has been shown that the ERG 3′ untranslated region is a regulative target of hsa-miR-145 in vitro. From this observation it is suggested that the miRNA hsa-miR-145 leads to progression of prostate cancer. The down regulation of hsa-miR-145 is also mentioned in [52,53].

In [54], it has been shown that the hsa-miR-20a is over expressed in prostate cancer. Moreover, Sylvestre et al. described an over expression of hsa-miR-20a in the human prostate cancer cell line PC3 using PCR [55]. Volinia et al. recorded an up-regulation of hsa-miR-20a in prostate cancer tissue using a microarray assay [56]. The identified function of hsa-miR-20a is the modulation of the translation of the E2F2 and E2F3 mRNAs via binding sites in their 3′-untranslated region [55], which supports the oncogenic behavior of hsa-miR-20a. The over expression of hsa-miR-20a reduces apoptosis in the prostate cancer cell line [55]. As suggested in [56] and miR2Disease, the hsa-miR-25 is also up-regulated in prostate cancer.

In [57,58], it is shown that hsa-miR-143 expression is clearly down-regulated during prostate cancer progression. ERK5 is known to promote cell growth and proliferation in response to growth factors and tyrosine kinase activation. Therefore, persistent decreased levels of hsa-miR-143 in cancer cells may be directly
involved in carcinogenesis through activation of the mitogen-activated protein kinase (MAPK) cascade via ERK5. Taken together these findings suggest that hsa-miR-143 could be a tumor suppressor and a potential novel diagnostic or prognostic marker in prostate cancer.

According to Hirata et al. [59], the hsa-miR-182 regulates FOXF2, RECK and MTSS1 genes and is therefore over expressed in prostate cancer. They have also shown experimentally that these three genes are potential targets of the hsa-miR-182 and play important role in progression of prostate cancer. Another miRNA, hsa-miR-96, is shown to be over expressed in prostate cancer as mentioned in [60].

Conclusion
The contribution of the paper is two fold, namely,

1. the development of the $\mu$HEM algorithm for miRNA selection, integrating the merits of rough sets and hypercuboid equivalence partition matrix; and
2. demonstrating the effectiveness of the proposed algorithm, along with a comparison with other algorithms, on several real life miRNA expression data sets.

The concept of hypercuboid equivalence partition matrix is found to be successful in selecting relevant and significant miRNAs of real valued microarray data sets. This formulation is geared towards maximizing the utility of rough sets and hypercuboid approach with respect to insilico identification of differentially expressed miRNAs. The results obtained on six miRNA data sets demonstrate that the proposed method can bring a remarkable improvement on miRNA selection problem, and therefore, it can be a promising alternative to existing models for 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 underlying etiology of a disease, providing a useful tool for exploratory analysis of miRNA data.

Availability and requirements
Project name: $\mu$HEM (Differentially expressed microRNA selection method)
Project home page: www.isical.ac.in/~bibl/results/mihem/mihem.html
Operating system: developed on Linux (Ubuntu 12.04 LTS)
Programming language: C

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
SP designed the current work. PM developed the concept of rough set based miRNA selection algorithm. SP implemented it and applied on different miRNA expression data sets. Both SP and PM analyzed the results and prepared the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
This work is partially supported by the Indian National Science Academy, New Delhi (grant no. SP/YS/68/2012). The work was done when one of the authors, S. Paul, was a Senior Research Fellow of Council of Scientific and Industrial Research, Government of India.

Published: 4 September 2013
Received: 18 March 2013 Accepted: 30 August 2013
References
1. Lu J, Getz G, Miska EA, Saavedra EA, Lamb J, Peck D, Cordon-Cardo C, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: MicroRNA expression profiles classify human cancers. Nat Lett 2005, 435(9):834–838.
2. Kim S, Ahn C, Hwang SJ, Yeo SJ, Yoon JH, Cho YS: Identification of differentially expressed microRNAs in human male breast cancer. BMC Bioinformatics 2010, 10:1–9.
3. Lehmann U, Streichert T, Otto B, Altab C, Hasemeier B, Christgen H, Schipper I, Hille U, Kreipe HH, Langer F: Identification of differentially expressed microRNAs in human male breast cancer. BMC Genomics 2010, 11:1–9.
4. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Mota NL, Teschendorff AE, Green AR, Ellis IO, Tavare S, Caldas C, Miska EA: MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 2007, 8:1–16.
5. Chen Y, Stallings RL: Differential patterns of microRNA expression in neuroblastoma are correlated with progression, differentiation, and apoptosis. Cancer Res 2007, 67:976–983.
6. Guo I, Miao Y, Xiao B, Huan R, Jiang Z, Meng D, Wang Y: Differential expression of microRNA species in human gastric cancer versus non-tumorous tissues. J Gastrointest Hepatol 2009, 24:652–657.
7. Schrauder MG, Stroch R, Schulz-Wendland R, Strissel PL, Kahlmann L, Loehberg CR, Lux MP, Judd SJ, Hartmann A, Hein A, Bayer CM, Bani MR, Richter S, Adamietz BR, Renkel E, Rauh C, Beckmann MW, Fasching PA: Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS ONE 2012, 7:1–9.
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, 5:1–12.
9. Paul S, Maiti P: Rough sets for Insilico identification of differentially expressed miRNAs. Int J Nanomedicine 2013, 8:1–12.
10. Ambrozio C, McClellan GJ: Selection bias in gene expression on the basis of microarray gene-expression data. Proc Natl Acad Sci, USA 2003, 99:6562–6566.
11. Iorio MV, Visone R, Levad GD, Donati V, Petrocca F, Casalini P, Taccioli C, Voltini S, Liu CG, Alder H, Calin GA, Menard S, Croce CM: MicroRNA signatures in human ovarian cancer. Cancer Res 2007, 67(18):8699–8707.
12. Li S, Chen X, Zhang H, Liang X, Yang Y, Yu C, Zen K, Li Y, Zhang CY: Differential expression of microRNAs in mouse liver under aberrant energy metabolic status. J Lipid Res 2009, 50:176–1765.
13. Nassier S, Arif N, Sridhar S, Haney L, Korn RL, Gotway MB, Weiss GI, Kim S: Identifying miRNA and imaging features associated with metastasis of lung cancer to the brain. In Proceedings of the 3rd IEEE International Conference on Bioinformatics and Biomedicine. Washington, 2009:246–251.
14. Ortega FJ, Moreno-Navarrete JM, Pardo G, Sabater M, Hummel M, Ferrer A, Rodriguez-Hermosa J, Ruiz B, Ricart W, Perez L, Real JMF: MicroRNA expression profile of human subcutaneous adipose and during adipocyte differentiation. PLoS ONE 2010, 5:1–12.
15. Pereira PM, Marques JP, Soares AR, Carreto L, Santos MAS: MicroRNA expression variability in human cervical tissues. PLoS ONE 2010, 5:1–12.
16. Raponi M, Dossey L, Jatkoe T, Wu X, Chen G, Fan H, Beer DG: MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. Cancer Res 2009, 69(14):5776–5783.
17. Azora S, Ranade AR, Tran NL, Nasser S, Srirad S, Korn RL, Ross JT, Dhruv H, Foss KM, Sibnaller Z, Ryken T, Gotway MB, Kim S, Weiss GJ: miRNA-328 is associated with Non-Small Cell Lung Cancer (NSCLC) brain metastasis and mediates NSCLC migration. Int J Cancer 2012, 131(9):2621–2631.

18. McIver AD, East P, Meene S, Maji P, Paul S, Schuettler N, Herionymus H, Gopalakrishnan C, Natarajan A, Xiao Y, Bansal V, Arora A, Thakur P, Efron B, Tibshirani R: Comparative miRNA expression profiles in individuals with latent and active tuberculosis. BMC Med 2013, 11:10-1–11.

19. Zhu M, Yi M, Kim CH, Deng C, Li Y, Medina D, Stephens RM, Green JE: miRNA expression profiles in individuals with latent and active tuberculosis. BMC Med 2013, 11:10-1–11.

20. Xue R, Xu J, Wunsch DC: MiRNA expression profile based cancer classification using default ARTMAP, Neural Netw 2009, 22:774–780.

21. Pawlak Z: Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer; 1991.

22. Maji P, Paul S: Rough-fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging. New Jersey: Wiley-IEEE Computer Society Press, 2012.

23. Fang J, Busse JWG: Mining of microRNA expression data: a rough set approach. In Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology. Berlin, Heidelberg: Springer; 2006:758–765.

24. Maji P, Paul S: Rough-fuzzy supervised attribute clustering algorithm and classification of microarray data. IEEE Tran Syst, Man, Cybern, Part B: Cybern 2011, 41:222–233.

25. Maji P, Paul S: Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. Int J Approximate Reasoning 2011, 52(3):408–426.

26. Maji P, Paul S: Rough-fuzzy clustering for grouping functionally similar genes from microarray data. IEEE/ACM Trans Comput Biol Bioinformatics 2013. doi:10.1109/TCBB.2012.103.

27. Paul S, Maji P: Robust RFCM algorithm for identification of co-expressed miRNAs. In Proceedings of the 6th IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, 2012:520–523.

28. Maji P, Paul S: Rough sets and support vector machine for selecting differentially expressed miRNAs. In Proceedings of the 6th IEEE International Conference on Bioinformatics and Biomedicine Workshops: NanoInformatics for Biomedicine. Philadelphia, 2012:864–871.

29. Slezak D: Rough sets and few-objects-many-attributes problem: the case study of analysis of gene expression data sets. In Proceedings of the Frontiers in the Convergence of Bioinference and Information Technologies. Cheju Island: IEEE Computer Society; 2007:233–240.

30. Slezak D, Wroblewski J: Roughification of numeric decision tables: the case study of gene expression data. In Proceedings of the 2nd International Conference on Rough Sets and Knowledge Technology. Berlin, Heidelberg: Springer; 2007:316–323.

31. Valdez JJ, Barton AJ: Relevant attribute discovery in high dimensional data: application to breast cancer gene expression. In Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology. Berlin: Springer; 2006:482–489.

32. Maji P, Paul S: Rough-fuzzy C-means algorithm: design and applications in coding and non-coding RNA expression data clustering. J Nanobiotechnol 2013, 11:124:1–2153–174.

33. Wei JM, Wang SQ, Yuan XJ: Ensemble rough hypercuboid approach for classifying cancers. IEEE Trans Knowl Data Eng 2010, 22(5):381–391.

34. Efron B, Tibshirani R: Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997, 92(438):548–560.

35. Keller A, Leidinger P, Wendschlag A, Scheffler M, Meese E, Wucherpfennig F, Huxer H, Borries A: miRNAs in lung cancer—studying complex fingerprints in patient’s blood cells by microarray experiments. BMC Cancer 2009, 9:353.

36. Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, Lenhof HP, Ruepkeht K, Meese E: Multiple sclerosis: MicroRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PLoS ONE 2009, 4(10):e744.

37. Taylor BS, Schultz N, Hieronymus H, Gopalakrishnan C, A disadvant AJ, Van Der Bus, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsuad S, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastman JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL: Integrative genomic profiling of human prostate cancer. Cancer Cell 2010, 18:1–22.

38. Tseng CW, Lin CC, Chen YN, Huang HC, Juan HF: Integrative network analysis reveals active microRNAs and their functions in gastric cancer. BMC Syst Biol 2011, 5:99.

39. Ralfkiaer U, Hagedorn PH, Bangsgaard N, Lovendorf MB, Ahler CB, Svensson L, Kopp KL, Vennegaard ML, Lauenborg B, Zibert JR, Krejgaard T, Bonefeld CM, Sokilde R, Gjerdrum LM, Labuda T, Mathiessen AM, Grønbaek K, Wasik MA, Sokolowska-Wojdylo M, Quella-Roussel C, Gniadecki R, Ralfkiaer E, Geisler C, Litman T, Wottrmann A, Glue C, Ropke MA, Skyv L, Odum N: Diagnostic microRNA profiling in cutaneous T-cell lymphoma (CTCL). Blood 2011, 118(2):5891–5900.

40. Vapnik V: The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1999.

41. Quinlan JR: C4.5: Programs for Machine Learning. CA: Morgan Kaufmann, 1993.

42. Deng C, Peng H: Minimum redundancy feature selection from Microarray gene expression data. J Bioinformatics Comput Biol 2005, 3(2):185–205.

43. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286:531–537.

44. Buelmann P, Yu B: Boosting with the L2 loss: regression and classification. J Am Stat Assoc 2003, 98:324–339.

45. Tibshirani R: Regression shrinkage and selection via the lasso. J R Stat Soc B 1996, 58:267–288.

46. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: ‘Gene Shaving’ as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 2000, 1(2):1–21.

47. Jiang Q, Wang Y, Hao Y, Juan L, Ming Z, Zhang X, Li M, Wang G, Liu Y: miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009, 37:D98–D104.

48. Hart M, Wach S, Natke E, Szczyzobiala J, Menon R, Taubert H, Hartmann A, Stoevhr R, Wieland W, Gissi E, Wullich B: The proto-oncogene ERG is a target of microRNA miR-145 in prostate cancer. FEBS J 2013, 280(9):2105–2116.

49. Ozen M, Creighton CJ, Ozdemir M, Ittmann M: Widespread deregulation of microRNA expression in human prostate cancer. Oncogene 2003, 21:788–793.

50. Wang L, Tang H, Thaynithy V, Subramanian S, Obeng AL, Cunningham JM, Cerhan JR, Steer CJ, Thibodeau SH: Gene networks and microRNAs implicated in aggressive prostate cancer. Cancer Res 2009, 69:4940–9497.

51. Pesta M, Klecka J, Kulda V, Topolan C, Hora M, Eret V, Ludvikova M, Babjuk M, Novak K, Stold J, Holubec L: Importance of miR-20a expression in prostate cancer tissue. Anticancer Res 2010, 30(9):3579–3583.

52. Sylvestre Y, De Guire V, Querido E, Mukhopadhyay UK, Bourdeau V, Major F, Ferbeyre G, Chartrand P: An E2F/miR-20a autoregulatory feedback loop. J Biol Chem 2007, 282(4):2135–2143.

53. Vellinis S, Calin GA, Liu CG, Ambros M, Cammro D, Petracca F, Visone R, Iorio M, Roido S, Ferracin M, Pruet R, Yahara H, Lanza G, Scarpa A, Vecchione A, Negrimi M, Harris CC, Croce CM: A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Nat Acad Sci, USA 2006, 103(7):2257–2261.

54. Claes C, Fritz V, Henriquet C, Apparailly F, Fernandez PL, Iborra F, Avancé C, Villalba M, Culine S, Fajas L: miR-143 interferes with ERK5 signaling, and abrogates prostate cancer progression in mice. PLoS ONE 2009, 4(10):e7594.
58. Porkka KP, Pfeiffer MJ, Waltering KK, Vessella RL, Tammela TL, Visakorpi T: MicroRNA expression profiling in prostate cancer. *Cancer Res* 2007, 67(13):6130–6135.

59. Hirata H, Ueno K, Shahryari V, Deng G, Tanaka Y, Tabatabai ZL, Hinoda Y, Dahiya R: MicroRNA-182-5p promotes cell invasion and proliferation by down regulating FOXF2, RECK and MTSS1 genes in human prostate cancer. *PLoS ONE* 2013, 8(1):e55502.

60. Schaefer A, Jung M, Mollenkopf HJ, Wagner I, Stephan C, Jentzmik F, Miller K, Lein M, Kristiansen G, Jung K: Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *Int J Cancer* 2010, 126(5):1166–1176.

doi:10.1186/1471-2105-14-266

Cite this article as: Paul and Maji: μHEM for identification of differentially expressed miRNAs using hypercuboid equivalence partition matrix. *BMC Bioinformatics* 2013 14:266.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit