Fundamenta Informaticae 127 (2013) 289–305
DOI 10.3233/FI-2013-910
IOS Press

A Weighted Threshold for Detection of Cancerous miRNA Expressions

Jayanta Kumar Pal*, **Shubhra Sankar Ray**†, **Sankar K. Pal**†
Center for Soft Computing Research
Indian Statistical Institute
203 B. T. Road, Kolkata- 700108, India
jkp_it08@isical.ac.in; shubhra@isical.ac.in; sankar@isical.ac.in

Abstract. MicroRNAs (miRNA) are one kind of non-coding RNA which play many important roles in eukaryotic cell. Investigations on miRNAs show that miRNAs are involved in cancer development in animal body. In this article, a threshold based method to check the condition (normal or cancer) of miRNAs of a given sample/patient, using weighted average distance between the normal and cancer miRNA expressions, is proposed. For each miRNA, the city block distance between two representatives, corresponding to scaled normal and cancer expressions, is obtained. The average of all such distances for different miRNAs is weighted by a factor, to generate the threshold. The weight factor, which is cancer dependent, is determined through an exhaustive search by maximizing the F score during training. In a part of the investigation, a ranking algorithm for cancer specific miRNAs is also discussed. The performance of the proposed method is evaluated in terms of Matthews Correlation Coefficient (MCC) and by plotting points $(1 - Specificity\ vs.\ Sensitivity)$ in Receiver Operating Characteristic (ROC) space, besides the F score. Its efficiency is demonstrated on breast, colorectal, melanoma lung, prostate and renal cancer data sets and it is observed to be superior to some of the existing classifiers in terms of the said indices.

Keywords: miRNA expression analysis, cancer detection, pattern recognition, bioinformatics

*Address for correspondence: Center for Soft Comp. Res., Indian Statistical Institute, 203 B. T. Road, Kolkata- 700108, India
†Also works: Center for Soft Computing Research, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

1. Introduction

Different macromolecules such as DNA, RNA, protein play several important roles in various developmental stages in animal body [11, 13]. It is observed that one of the major functions of protein is cell signaling, by which cell proliferation (the growth and reproduction of cells), apoptosis (the process of programmed cell death) etc., are controlled. The coding RNAs (e.g., messenger RNA) are directly involved in the protein translation process and a non coding RNA, called micro RNA (miRNA), controls the messenger RNAs (mRNA) in protein translation process. MiRNAs can act as both, onco-miRNA (miRNAs which promotes the development of tumors) and suppressor-miRNA (miRNAs which suppress the development of tumors) [16, 17] by inhibiting protein translation process. The onco-miRNAs show over-expression and the suppressor-miRNAs show under-expression in cancer. MiRNA and its expression deregulation in different tissue conditions (normal and disease) is an important research area over a decade. The deregulation of miRNA expressions refers to the change of expression value from the normal expression value. Measuring the fold change is one of the ways to identify the deregulated miRNAs, where the fold change of a miRNA is determined by taking the ratio between the deregulated and the normal expression value of that miRNA. Abnormality in expressions of miRNAs can be observed even in the case of very little sign of cancer and miRNA expressions show better performance than mRNA expressions to separate normal and cancer tissues [9]. MiRNAs can be found at tissue locations and also in blood [7, 14, 16], which helps to detect the cancerous miRNA expressions from the blood sample. These types of characteristic features make miRNAs very important in cancer research.

As miRNAs show deregulated expressions in cancer condition, a particular cancer can be predicted from the expression values of the responsible miRNAs for that cancer. For the normal condition of a patient, the miRNAs should show normal expressions and in cancer condition, the miRNAs (i.e., responsible miRNAs for the cancer) should show deregulated expressions.

MiRNAs act as an important indicator of the cancerous state of a tissue and early detection of cancer is one way to increase the survival rate and time of cancer patients. The first investigation on miRNA and its role on cancer was done in 2002 [3]. In [9] it is shown that miRNA gets globally downregulated in the cancer patients and miRNAs are much more informative than mRNAs for cancer detection, even in the case of poorly differentiated cancers (i.e., the cancerous tissues which does not show significant symptoms of cancer). From the investigation in [1], it is observed that 352 human miRNAs are used to identify responsible miRNAs for the colorectal cancer and 37 miRNAs are identified as deregulated miRNAs in colorectal cancer. Initially those miRNAs are selected which show median normalized signal intensity greater than 100. The miRNAs having average fold change (positive or negative), between cancer and normal expressions, greater than 1.5 and with p-value less than 0.05 are finally selected as the deregulated miRNAs. Similar type of investigation with melanoma cancer is observed in [7], where, two criteria are checked for selecting the miRNAs. One is, whether a miRNA shows 2-fold up or down-regulation in the expression of cancer cells as compared to the normal cells and the other is whether combined median intensity for all normal and all cancer expressions separately exceeds 100 for that miRNA. In this way 51 out of 866 miRNAs are identified as deregulated miRNAs in the cancer condition. One important characteristic of this investigation is that the miRNAs are collected from blood, not from the cancerous tissue.

In [2], expressions of 309 unique human miRNAs are initially generated. Then, unsupervised hierarchical clustering (using Pearson correlation and average linkage) is applied on some selected miRNA expressions which separates five breast cancer subtypes according to the different clinicopathological characteristics, like tumor size, lymph node status etc.

Emphasis is given to the ranking of miRNAs in [10]. In [8], miRNA expressions are generated using biochemical methods and a clear separation between normal and cancer expressions is obtained using hierarchical clustering with spearman rank correlation as the similarity measure.

From the existing literature survey it is observed that most of the investigations are based on clustering techniques(i.e., unsupervised). In this investigation we propose a weighted threshold (WT) based method (supervised) to detect the condition of a miRNA of a given patient. For each miRNA, the city block distance between two representatives of that miRNA, belonging to normal and cancer classes, are obtained and the process is repeated for all the miRNAs. The average of all such distances is calculated and again scaled by the standard deviation of those distances. Finally, this scaled average distance is weighted by a factor (w), where the value of w is dependent on the type of cancer, to generate the WT. The weight factor is determined through an exhaustive search by maximizing the F score during training. For a particular cancer, WT, thus generated, reflects the actual distance between normal and cancer miRNA classes in order to make them separable as much as possible. The categorization performance of WT is compared with those of SVM, kNN (k=1, 2, 3 and 4) and the two class classifiers [12], and the method is found to be superior to them in terms of F score and Matthews Correlation Coefficient (MCC).

The rest of the article is organized as follows. Section 2 describes the miRNA data set generation and the data sets used in the proposed investigation. The details about the proposed method is discussed in Section 3. The experimental results are reported in Section 4, and finally Section 5 concludes the investigation.

2. MiRNA Data Sets

MiRNA data set contains the expression values of different miRNAs at different conditions (normal or cancer) corresponding to the different patients. In this section, first we discuss about the transcription process of miRNA in the cell, and then we discuss the procedures for obtaining miRNA expressions. Finally, we describe the data sets used for testing the proposed methodology.

2.1. Generation of miRNA in the Cell

Initially miRNAs are transcribed as precursor miRNA (pre-miRNA) of length $\sim$1000 nt in the nucleolus. Then it is cleaved by RNase endonuclease-III enzyme Drosha and its partner DGCR8/Pasha to generate precursor miRNA (pre-miRNA) of length $\sim$70 nt and transported to the cytoplasm through the pores of the nuclear membrane with the help of RanGTP and exprotin-5. Pre-miRNAs are further processed by Dicer enzyme and generates $\sim$22 nt mature miRNA duplex, containing a guide strand and a passenger strand. From this duplex, passenger strand degrades and the guide strand generates simplex mature miRNA. The simplex miRNA creates bond with RNA-induced silencing complex (RISC) and binds with mRNA at 3'-untranslated region (3'UTR) of the mRNA to stop the translation process (the process of generating protein from mRNA).

2.2. Generation of miRNA Data Sets

Three major procedures are available for obtaining the miRNA expressions. These are (i) miRNA expression profiling by cloning and sequencing, (ii) microarray analysis and (iii) microbead expression analysis [4]. One of the most successful technologies for obtaining miRNA expressions is xMAP, which belongs to the microbead expression analysis methodology. In xMAP method, expressions of 100 different miRNAs can be obtained at a time. Every miRNA is identified by a specific color (fluorescent dye) code and the color intensity value measured by the scanner, is stored as the expression value of that miRNA.

2.3. miRNA Subset Selection

Data preprocessing, involving ranking of miRNAs, is necessary for miRNAs as all of them are not deregulated significantly for a particular type of cancer. In this regard, first the miRNAs are ranked and then a set of miRNAs is selected according to the fold change with p value. As mentioned in Section 1 the fold change is defined as the ratio between the deregulated expression and the normal expression and the deregulation of miRNA expression is defined as the change of the miRNA expression from the normal expression. In [10], miRNAs are ranked on the basis of deregulation in multiple cancer types. Hence, the method can ignore an important miRNA responsible for a particular cancer only, and the set of miRNAs is dominated by those miRNAs which are responsible for multiple cancer types. Therefore, in this investigation for every data set miRNAs are ranked by considering the related cancer type only. Moreover, in [10], fold change is calculated over the paired samples (i.e., miRNA expressions are generated form the normal and cancer tissue of the same patient) but in many data sets unpaired samples are also present, where, in classification phase, one has to classify an unknown expression by comparing with unpaired normal and cancer samples. Hence, in this investigation fold change between the unpaired samples are also considered for the training puppose.

Let, x_i^k and y_j^k be the ith ($i = 1, 2, 3, ..., N$) normal and jth ($j = 1, 2, 3, ..., M$) cancer expressions of the kth ($k = 1, 2, 3, ..., L$) miRNA, respectively. The steps for the ranking algorithm are as follows.

S1) Calculate the fold change between normal and cancer expressions of the kth miRNA as

$$F_{ij}^k = \frac{y_j^k}{x_i^k}, \tag{1}$$

to detect the top ranked upregulated (over expressed) miRNAs. For the detection of top ranked downregulated (under expressed) miRNAs, calculate the fold change as

$$F_{ij}^k = \frac{x_i^k}{y_j^k}. \tag{2}$$

S2) For a particular value of i and j, sort all the miRNAs in descending order using F_{ij}^k and store the index values (i.e., rank of miRNAs for particular values of i and j) in R_{ij}^k. Repeat the step for all i and j.

S3) The rank consistency score (RCoS) [10] of kth miRNA corresponding to ith normal and jth cancer patient is defined as

$$S_{ij}^{k} = R_{ij}^{k}/L \quad \forall\, i, j \tag{3}$$

S4) For each k (i.e., for every miRNA), sort the values of S_{ij}^{k} in ascending order and select the first m numbers of S_{ij}^{k}, where, $m = \lceil \frac{N \times M}{2} \rceil$.

S5) For each k, determine the maximum S_{ij}^{k} among the selected m numbers of S_{ij}^{k}, as

$$S_{m}^{k} = max_{m}(S_{ij}^{k}) \tag{4}$$

S6) Compute the p-value of S_{m}^{k} as

$$p_{m}^{k} = \sum_{l=m}^{r} \binom{r}{l} a^{l} (1-a)^{(r-l)} \tag{5}$$

where, $a = S_{m}^{k}$ and $r = N \times M$.

S7) Remove all miRNAs with p-value greater than or equal to 0.005 (i.e., $p_{m}^{k} \geq 0.005$).

S8) Select a portion (say, $P\%$) of remaining miRNAs starting from the top of the list.

In this investigation we used six different types of data sets, viz., breast [2], colorectal [1], melanoma [7], lung [6], prostate [15] and renal cancer [5], for testing the performance of the proposed method. The summary of the processed data sets is represented in Table 1.

Table 1. Summary of the used data sets

Cancer Type	*Total No. of human miRNAs*	*No. of selected miRNAs*	*No. of Normal Patients*	*No. of Cancer Patients*
Breast cancer	309	9	5	93
Colorectal cancer	352	10	8	58
Melanoma cancer	866	9	22	35
Lung cancer	866	22	19	17
Prostate cancer	12033	10	12	12
Renal cancer	12033	18	12	12

3. Proposed Approach

The goal of this investigation is to detect the condition (normal or cancer) of the miRNAs of a given patient. In this regard we propose a weighted threshold (WT) based method where, for a particular miRNA, we scale all of its expression values (i.e., both the normal and cancer) by the overall standard deviation computed over all its expression values, and two representatives corresponding to normal and cancer classes of that miRNA are determined by calculating the mean values of the scaled normal and cancer expression values, respectively. The city block distance between the two representatives is then calculated and the process is repeated for all the miRNAs. The average of all the distances is scaled by the standard deviation of those distances and the scaled average distance is weighted by a factor, to generate the WT. The weight is determined through exhaustive search by maximizing the F score (see Eq. 13) in the training process. In the testing process, for a particular miRNA we scaled its expression value by the standard deviation of that miRNA (calculated at the time of training) and calculated the distances (city block) from both the class representatives. If the city block distance of the scaled unknown expression from the normal class representative is closer to the WT value then the expression is considered as cancerous and it is considered as normal for the opposite condition. The process is repeated for all the miRNAs in the test sample. In this investigation, leave-one-out cross validation procedure is used for training and testing, where, at a particular instance one sample is kept for testing purpose and all other samples are used for training. The process is repeated for all the samples, one by one, and the average result over all the samples is considered as the performance of the method.

Let, N, M and L be the total numbers of the normal patients, cancer patients and miRNAs, respectively, in a data set. The ith ($i = 1, 2, ..., N$) normal expression (i.e., the expression of the ith normal patient) of the kth ($k = 1, 2, ..., L$) miRNA is represented as x_i^k and the jth ($j = 1, 2, ..., M$) cancerous expression (i.e., the expression of the jth cancer patient) of the kth miRNA is represented as y_j^k . As we are using leave-one-out cross-validation, the numbers of training samples in the normal and the cancer patients are $N - 1$ and M, respectively, when the test sample is selected from the normal patients. Similarly, there are N and $M - 1$ numbers of training samples, respectively, if the test sample is selected from the cancer patients.

The steps for the proposed method are as follows:

S1) If the test sample is chosen from the normal patients, and x_p^k is the expression of the kth miRNA of that patient, calculate the class representatives of the kth miRNA for the normal and cancer classes as

$$r_n^k = \frac{1}{\sigma^k(N-1)} \sum_{i=1, i\neq p}^{N} x_i^k \tag{6}$$

and

$$r_c^k = \frac{1}{\sigma^k \times M} \sum_{j=1}^{M} y_j^k, \tag{7}$$

respectively. Where, $i,\ j,\ k,\ x_i^k\ and\ y_j^k$ are the same variables as mentioned earlier and σ^k represents the standard deviation of all the expression values (i.e., normal and cancer) of the kth miRNA in training samples. Note that, p can be any i ($i = 1, 2, ..., N$).

Calculate the representatives of the kth miRNA for the normal and the cancer classes, if test sample is chosen from the set of the cancer patients and y_q^k is the expression of the kth miRNA of that patient, as

$$r_n^k = \frac{1}{\sigma^k \times N} \sum_{i=1}^{N} x_i^k \tag{8}$$

and

$$r_c^k = \frac{1}{\sigma^k(M-1)} \sum_{j=1, j\neq q}^{M} y_j^k, \tag{9}$$

respectively. where, $i,\ j,\ k,\ x_i^k,\ y_j^k\ and\ \sigma^k$ are the same variables as mentioned earlier. Note that, q can be any j ($j = 1, 2, ..., M$).

S2) Calculate the city block distance between r_n^k and r_c^k as

$$d^k = |r_n^k - r_c^k| \tag{10}$$

S3) Repeat Steps S1 to S2 for all k ($k = 1, 2, 3, ..., L$).

S4) Calculate the scaled average distance as

$$D = \frac{1}{\sigma \times L} \sum_{k=1}^{L} d^k \tag{11}$$

where, σ represents the standard deviation of all the calculated city block distances between two class representatives (i.e., d^k for all k).

S5) The weighted threshold (WT) is generated as

$$WT = D \times w \tag{12}$$

where, w is the weight factor which depends on the type of cancers. Its value is determined through an exhaustive search (the initial value of w is 1 and increase the value of w in steps of 0.25) by maximizing the F score using the training samples. In other words, WT represents the actual distance between normal and cancer miRNA classes in order to make them separable as much as possible, where the optimum value of w is cancer dependent. Its value may be determine by maximizing the F score, during training, which is defined as

$$F = \frac{\textit{2} \times \textit{Sensitivity} \times \textit{Specificity}}{\textit{Sensitivity+Specificity}} \tag{13}$$

where, the sensitivity (Sn) is defined as

$$Sn = \frac{\textit{true positives (TP)}}{\textit{true positives (TP) + false negatives (FN)}} \tag{14}$$

and, the specificity (Sc) is defined as

$$Sc = \frac{\textit{true negatives (TN)}}{\textit{true negatives (TN) + false positives (FP)}}. \tag{15}$$

Here, the true positive refers to the number of correctly detected cancer miRNA expressions and false negative refers to the number of undetected cancer miRNA expressions. True negative implies the number of correctly detected normal miRNA expressions and false positive implies the wrongly detected cancer miRNA expressions (i.e., detected as cancer expressions, but actually they are normal expressions).

S6) In the testing phase, the goal is to detect the condition of the miRNAs of the test sample. For the kth miRNA of the test sample, calculate the scaled expression of the miRNA as

$$u'^k = \frac{u^k}{\sigma^k} \tag{16}$$

where, u^k represents the kth miRNA expression of the test sample.

S7) Now, the distances of u'^k from both the class representative values (r_n^k and r_c^k) are calculated as

$$d_n^k = |r_n^k - u'^k| \tag{17}$$

$$d_c^k = |r_c^k - u'^k| \tag{18}$$

S8) If u^k is the expression of a normal miRNA then the distance of u'^k from r_c^k will be much closer to WT as it represents a threshold for expression difference between the normal and cancer miRNAs (see Step S4) and if u^k is a deregulated (i.e., cancerous) miRNA expression then the distance of r_n^k from u'^k will be much closer to WT for the same reason. So, for considering the kth miRNA of the test sample as normal, the testing condition is given as

$$|WT - d_n^k| > |WT - d_c^k| \tag{19}$$

and for considering as cancer, the testing condition is given as

$$|WT - d_n^k| < |WT - d_c^k| \tag{20}$$

S9) Repeat Steps, S5 to S8 for all k (i.e., for all the miRNAs in the test sample), where, $k = 1, 2, ..., L$.

S10) Repeat Steps S1 to S9, for all the patients considering as test sample one by one (i.e., varying p_1 from 1 to N and q_2 from 1 to M).

S11) Evaluate the performance of the this method in terms of Mathews Correlation Coefficient (MCC), by plotting '$1 - specificity\ vs\ sensitivity$' in receiver operating characteristic (ROC) space and F score (see Eq. 13). The MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{21}$$

where, the TP, TN, FP and FN are similar to those defined in equations 14 and 15.

The value of MCC lies between -1 to +1. while, MCC value less than zero implies that the prediction capability is worse than random prediction, a value greater than zero indicates that the prediction capability is better than random prediction. In the ROC space, any point on the straight line, passing through the coordinates (0, 0) and (1, 1) (see Figs. 5(a)-5(c)), indicates that the prediction performance is the same as that of random prediction. On the other hand, coordinate (0, 1) implies a perfect prediction, the coordinate (1, 0) indicates a totally wrong prediction. So, any point above the line, passing through the coordinates (0, 0) and (1, 1), indicates a prediction better than random prediction.

4. Experimental Results

The performance of the *weighted threshold* (WT) is evaluated on six data sets viz., breast, colorectal, melanoma, lung, prostate and renal cancer (Table 1). As mentioned in the Section 3, a weight factor (w) is used to generate the WT for each data set in the training process. In this investigation, D (see Eq. 11) represents scaled average distance between normal and cancer miRNAs and w is introduced for increasing the distance between these two types of miRNAs to make them separable as much as possible. Hence the initial value of w is chosen as 1 and it is increased in step of 0.25 to maximize the F score during training. The variation of F with w for different data sets is shown in the Figs. 1(a), 1(b), 1(c), 1(d), 1(e) and 1(f). From the Figs. in 1(a), 1(b), 1(d), 1(e) and 1(f) it is observed that initially the F score is increasing with the increasing value of w and after a certain value of w (say, w_O) the F score is remaining the same and w_O is seen to lie usually between 1.5 to 2 for the data sets considered here.

Initially, for $w = 1$, some normal miRNA expressions are classified accurately using Eq. 19, as d_c^k (see Eq. 18) is more likely to be closer to WT than d_n^k (see Eq. 17). For the normal miRNAs which satisfy $d_n^k < d_c^k$ and do not satisfy Eq. 19, if the value of w is incremented WT increases, $|WT - d_n^k|$ increases and $|WT - d_c^k|$ decreases. Hence the specificity increases. Note that, those normal miRNAs which do not satisfy the condition $d_n^k < d_c^k$ (i.e., normal expression is closer to the cancer class representative than the normal class representative), are not correctly predicted at any value of WT and therefore the specificity becomes stable after a particular value of w. For a similar reason the sensitivity (for the cancer patient) initially increases and become stable after a value of w. As both the sensitivity and specificity, initially increases with w and become stable after a particular value of w, the F score also behaves in a similar manner. From the Fig. 1(c) for melanoma cancer data, it is observed that the F score is not changing with w. For this data set there are no normal miRNAs with $d_n^k < d_c^k$ and also not satisfying Eq. 19 and cancer miRNAs with $d_n^k > d_c^k$ and also not satisfying equation 20. So, in this data set the F score is not changing with w.

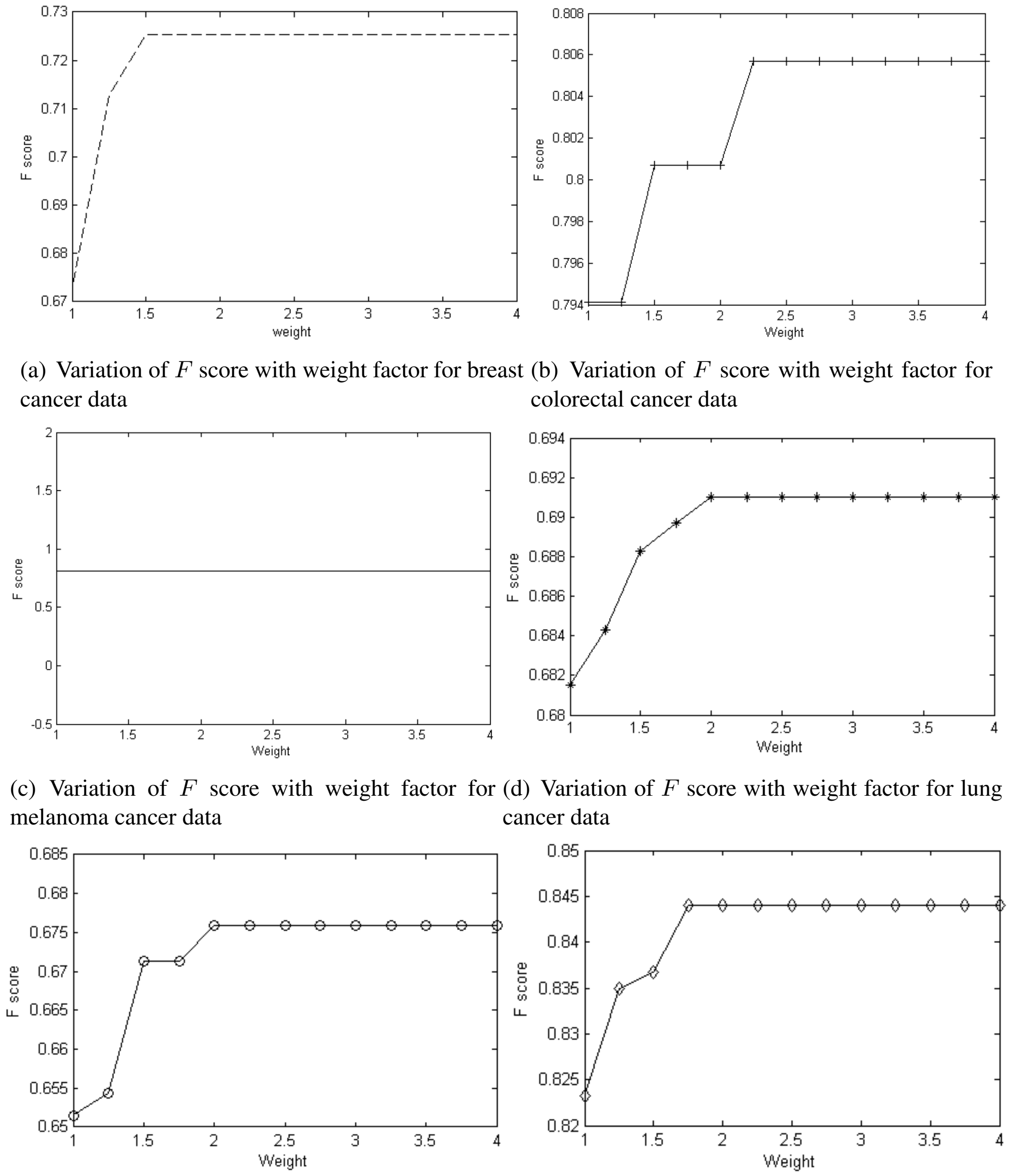
Let us now compare our WT based method with some existing classifiers, such as 'average expression based method' [12],'intraclass distance based method' [12], kNN and SVM in terms of F score. The difference between the proposed WT based method and those in [12] is that the former uses overall standard deviation and the same threshold irrespective of miRNAs instead of individual class wise standard deviation and different threshold, respectively, as used in the latter cases. We present the results of comparison in terms of F score along with the sensitivity and specificity in Table 2. It is observed that the WT based method performs better than all other classifiers in terms of F score for all the data sets. For example, using WT based method the F score varies from 0.6759 to 0.8440 for different data sets, whereas the second highest F score value obtained from different algorithms varies from 0.6444 to 0.8356 for the other classifiers.

Figs. 2(a) and 2(b) show the superior performance of WT as compared to other classifiers in terms of MCC value. It is observed that MCC value for WT ranges between 0.3519 to 0.6931 for different data sets, and the second best MCC value achieved by the other algorithms ranges from 0.3357 to 0.6718 for different data sets.

As mentioned in Section 1, a subset of miRNAs is already reported as the responsible miRNAs for breast, colorectal and melanoma cancer, in [2], [1] and [7], respectively. In [2], out of 309 miRNAs, 38 miRNAs are pointed out as differentially expressed in the normal and the cancerous breast samples. Similarly, in [1] and [7], 37 out of 352 and 51 out of 866 miRNAs are identified as differentially expressed between the normal and the cancer samples in colorectal and melanoma cancer, respectively.

Table 2. Comparison of sensitivity specificity and F scores

$Methods$	$Performance$ $Measure$	$Different\ data\ sets$					
		$Breast$ $cancer$	$Colorectal$ $cancer$	$Melanoma$ $cancer$	$Lung$ $cancer$	$Prostate$ $cancer$	$Renal$ $cancer$
WT	Sensitivity	0.7658	0.7293	0.8413	0.7143	0.6759	0.8088
	Specificity	0.6889	0.9000	0.7825	0.6692	0.6759	0.8824
	Fscore	0.7253	0.8057	0.8108	0.6910	0.6759	0.8440
Average expression based method	Sensitivity	0.7000	0.8586	0.7997	0.6504	0.5370	0.8480
	Specificity	0.6667	0.7125	0.7735	0.7044	0.8056	0.8235
	Fscore	0.6858	0.7788	0.7864	0.6763	0.6444	0.8356
Intraclass distance based method	Sensitivity	0.7037	0.8621	0.7997	0.6511	0.5448	0.8382
	Specificity	0.6667	0.7250	0.7632	0.7009	0.7948	0.8039
	Fscore	0.6847	0.7876	0.7810	0.6751	0.6465	0.8207
SVM	Sensitivity	0.9988	0.8862	0.8794	0.7727	0.6667	0.5098
	Specificity	0.0667	0.1100	0.4798	0.5957	0.6296	0.5931
	Fscore	0.1250	0.1957	0.6208	0.6728	0.6476	0.5483
kNN (k=1)	Sensitivity	0.9594	0.9103	0.7756	0.5187	0.6296	0.7892
	Specificity	0.2889	0.4000	0.7333	0.7919	0.5833	0.7794
	Fscore	0.4441	0.5558	0.7539	0.6268	0.6056	0.7843
kNN (k=2)	Sensitivity	0.9701	0.9000	0.7881	0.5053	0.6019	0.7794
	Specificity	0.2667	0.3625	0.7502	0.7823	0.5741	0.7843
	Fscore	0.4183	0.5168	0.7687	0.6140	0.5876	0.7819
kNN (k=3)	Sensitivity	0.9869	0.9293	0.7556	0.5642	0.6667	0.8333
	Specificity	0.1111	0.3250	0.8131	0.7799	0.6574	0.8333
	Fscore	0.1997	0.4816	0.7833	0.6547	0.6620	0.8333
kNN (k=4)	Sensitivity	0.9904	0.9259	0.7429	0.5722	0.6667	0.7941
	Specificity	0.1111	0.3625	0.8131	0.7751	0.6389	0.8382
	Fscore	0.1998	0.5210	0.7764	0.6584	0.6525	0.8156



(a) Variation of F score with weight factor for breast cancer data

(b) Variation of F score with weight factor for colorectal cancer data

(c) Variation of F score with weight factor for melanoma cancer data

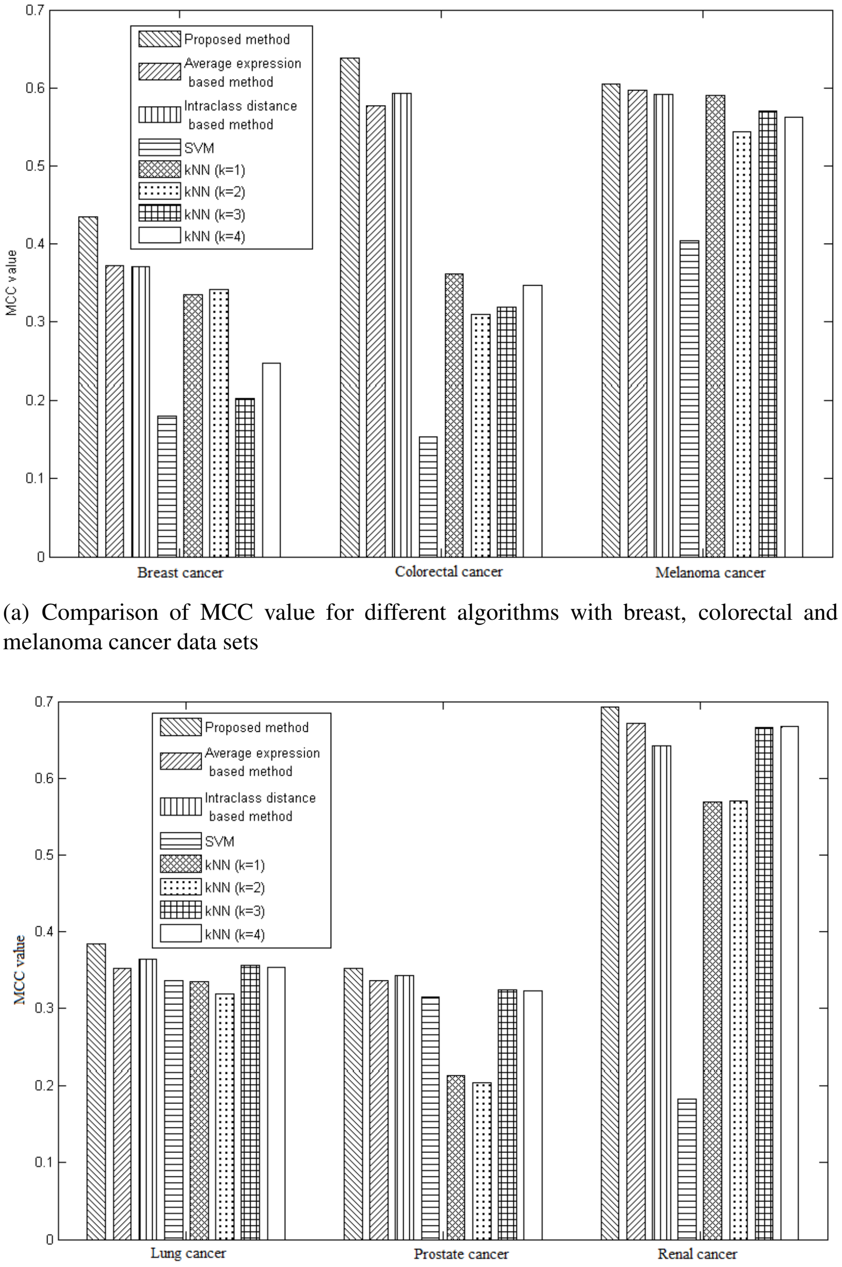
(d) Variation of F score with weight factor for lung cancer data

(e) Variation of F score with weight factor for prostate cancer data

(f) Variation of F score with weight factor for renal cancer data

Figure 1. Variation of F score corresponding to different weight factors

In our current investigation these miRNAs are also considered separately, to check the performance of the proposed method in terms of F score, MCC value and by plotting points $(1-Specificity\ vs\ Sensitivity)$ in ROC space. While the F score varies from 0.6357 to 0.8356, MCC varies from 0.2729 to 0.6043 for different data sets (i.e., breast, colorectal and melanoma cancer data sets). It is observed from the Figs. 3 and 4, that the proposed method performs better, even with these data subsets, than other methods



(a) Comparison of MCC value for different algorithms with breast, colorectal and melanoma cancer data sets

(b) Comparison of MCC value for different algorithms with lung, prostate and renal cancer data sets

Figure 2. Comparison of MCC values obtained from different subsets of miRNA corresponding to different methods and data sets

in terms of the said indices. It is also observed that the F score obtained by the various algorithms corresponding to the melanoma cancer data set is comparable with each other, except for SVM. From

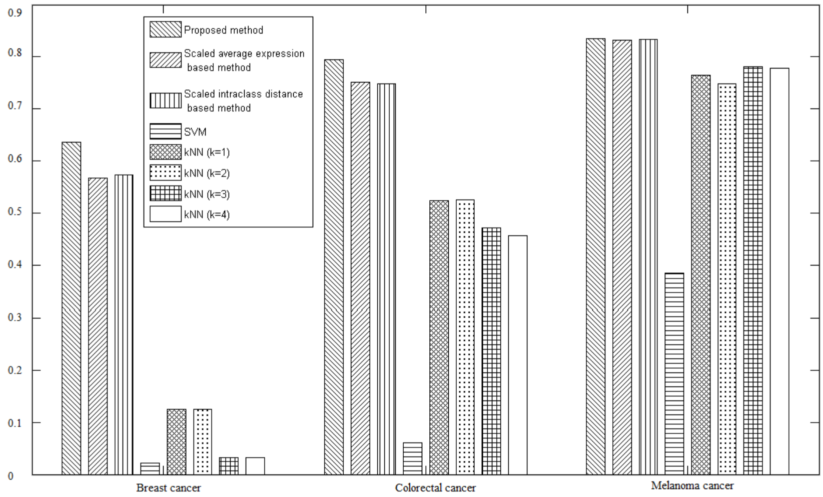


Figure 3. Comparison of F score obtained from different algorithms with different data subsets reported in the corresponding investigation

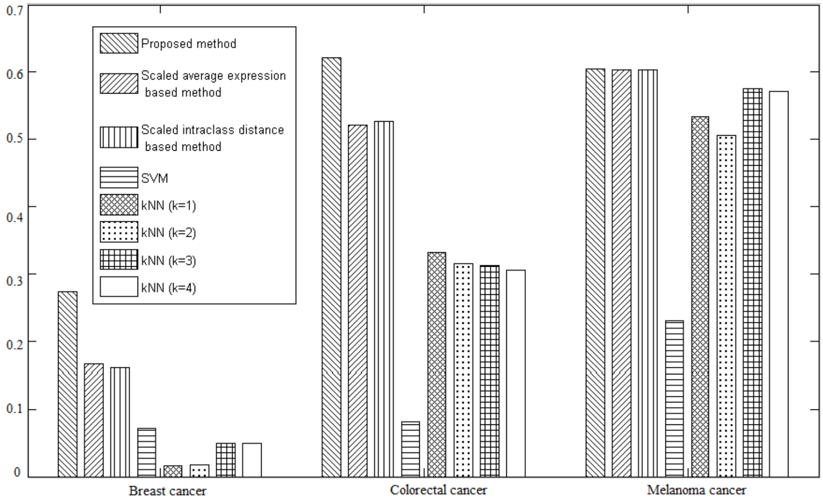
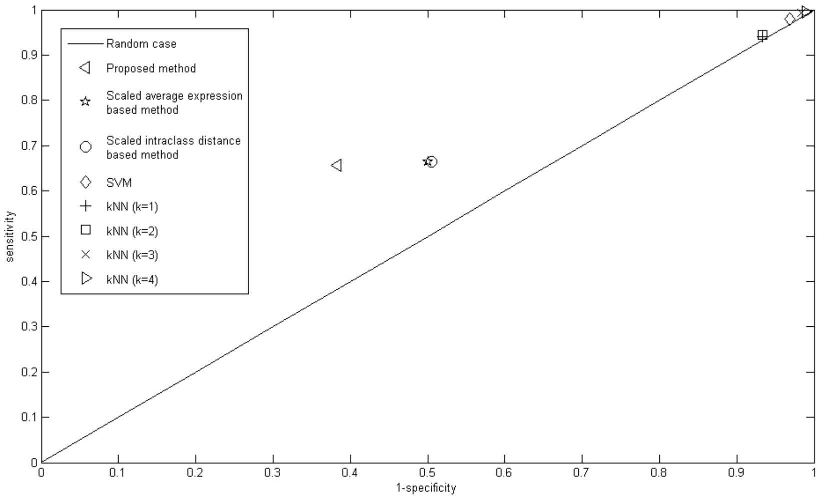
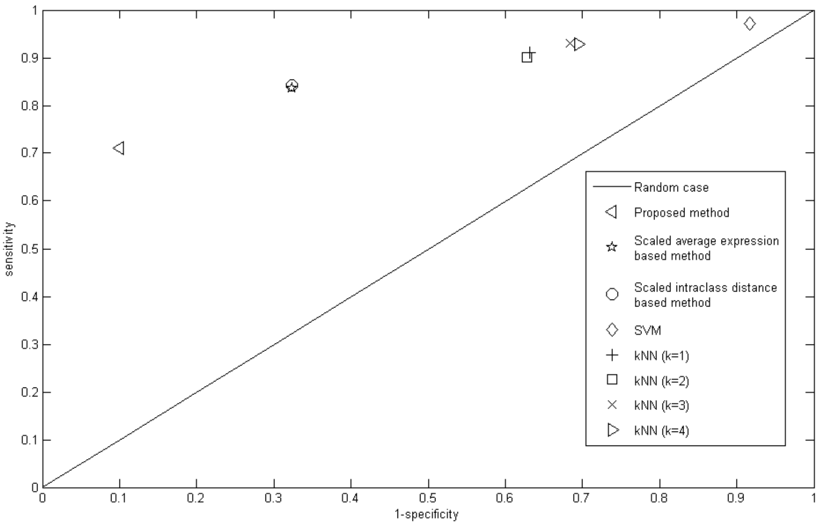


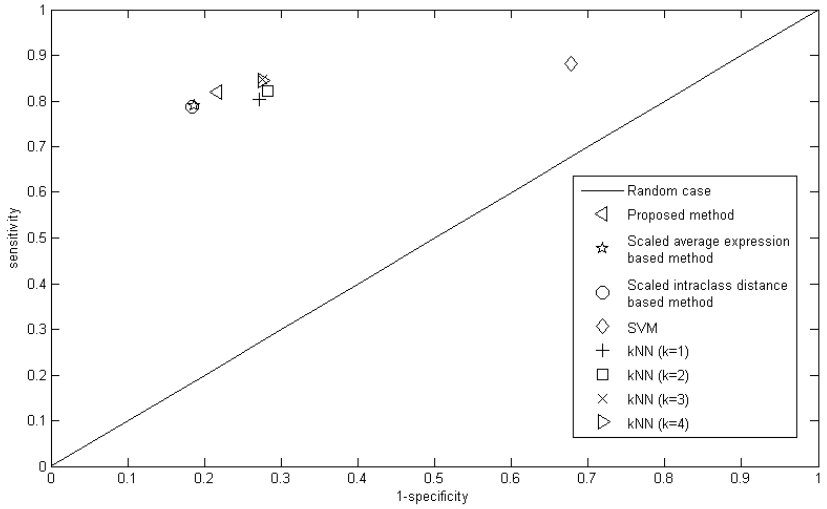
Figure 4. Comparison of MCC value for different algorithms with different data subsets reported in the corresponding investigation



(a) Comparison of the proposed method with 'average expression based method', 'intraclass distance based method', SVM and kNN in ROC space for breast cancer data



(b) Comparison of the proposed method with 'average expression based method', 'intraclass distance based method', SVM and kNN in ROC space for colorectal cancer data



(c) Comparison of the proposed method with ‘average expression based method’, ‘intraclass distance based method’, SVM and kNN in ROC space for melanoma cancer data

Figure 5. Comparison of the proposed method with different methods for different data subsets obtained from the related investigations, in ROC space

the points in ROC space (see Figs. 5(a), 5(b) and 5(c)), it is observed that for the breast and colorectal cancer data sets the specificity achieved by the proposed method is higher than other methods used for comparison. For the melanoma cancer data set the sensitivity obtained by the proposed method is higher than the average expression based method and intraclass distance based method, and the specificity achieved by the method is higher than the SVM and kNN.

5. Conclusion

In this investigation, a weighted threshold based method for cancerous miRNA detection is presented. The threshold is obtained using a weight factor (w) and the scaled average distance between the normal and cancerous miRNA expressions. It also considers the expression variation among the patients and the variation of distances (i.e., the distance between normal and cancer expressions) between different miRNAs. The performance of the proposed method is evaluated in terms of F score, MCC value and by plotting points ($1 - Specificity\ vs.\ Sensitivity$) in ROC space by using six types of data sets (viz., breast, colorectal, melanoma, lung, prostate and renal cancer data sets). The optimum value of the weight (w) for these data sets is seen to lie between 1.5 to 2. While, the F score value of the proposed method varied from 0.6759 to 0.8440, the MCC value varied from 0.3519 to 0.6931 for different data sets. The weighted threshold based method is found to be superior than the related methods in [12], SVM and kNN (k=1, 2, 3, 4) in terms of F score and MCC value for the data sets, mentioned above. From the table 2, it is observed that the proposed method gives both sensitivity and specificity values more than 0.65. The performance of the method is also tested in a part of the investigation on larger number of miRNAs, reported [1, 2, 7] as responsible miRNAs in the corresponding investigations, for breast, colorectal and melanoma cancer data sets, and superior results than other methods are achieved for all the data sets. The experimental results on different data sets confirmed the potential value of the WT framework for the detection of cancerous miRNAs using their expression values.

Acknowledgment

The authors are thankful to the Department of Science and Technology, Government of India, for establishing the Center for Soft Computing Research at Indian Statistical Institute, Kolkata. S. K. Pal acknowledges the J. C. Bose fellowship of the Government of India.

References

[1] Arndt, G. M., Dossey, L., Cullen, L. M., Lai, A., Druker, R., Eisbacher, M., Zhang, C., Tran, N., Fan, H., Retzlaff, K., Bittner, A., Raponi, M.: Characterization of global microRNA expression reveals oncogenic potential of miR-145 in metastatic colorectal cancer, *BMC Cancer*, **9**(374), 2009, 1–17.

[2] Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S., Dunning, M. J., Barbosa-Morais, N. L., Teschendorff, A. E., Green, A. R., Ellis, I. O., Tavar, S., Caldas, C., Miska, E. A.: MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype, *Genome Biology*, **8**, 2007, R214.1–R214.16.

[3] Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F., Croce, C. M.: Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia, *Proceedings of the National Academy of Sciences*, **99**(24), 2002, 15524–15529.

[4] Einat, P.: *Methodologies for High-Throughput Expression Profiling of MicroRNAs*, springer, New Jersey, 2006, 139–157.

[5] Jung, M., Mollenkopf, H.-J., Grimm, C., Wagner, I., Albrecht, M., Waller, T., Pilarsky, C., Johannsen, M., Stephan, C., Lehrach, H., Nietfeld, W., Rudel, T., Jung, K., Kristiansen, G.: MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy, *Journal of Cellular and Molecular Medicine*, **13**(9b), 2009, 3918-3928.

[6] Keller, A., Leidinger, P., Borries, A., Wendschlag, A., Wucherpfennig, F., Scheffler, M., Huwer, H., Lenhof, H.-P., Meese, E.: miRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments, *BMC Cancer*, **9**(353), 2009, 1-10.

[7] Leidinger, P., Keller, A., Borries, A., Reichrath, J., Rass, K., Jager, S. U., Lenhof, H. P., Meese, E.: High-throughput miRNA profiling of human melanoma blood samples, *BMC Cancer*, **10**(262), 2010, 1–11.

[8] Lodes, M. J., Caraballo, M., Suciu, D., Munro, S., Kumar, A., Anderson, B.: Detection of Cancer with Serum miRNAs on an Oligonucleotide Microarray, *PLoS ONE*, **4**(7), 2009, e6229.

[9] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., R.Horvitz, H., Golub, T. R.: MicroRNA expression profiles classify human cancers, *Nature*, **435**(7043), 2005, 834–838.

[10] Navon, R., Wang, H., Steinfeld, I., Tsalenko, A., Ben-Dor, A., Yakhini, Z.: Novel Rank-Based Statistical Methods Reveal MicroRNAs with Differential Expression in Multiple Cancer Types, *PLoS ONE*, **4**, 2009, e8003.

[11] Ray, S. S., Halder, S., Kaypee, S., Bhattacharyya, D.: HD-RNAS: an automated hierarchical database of RNA structures, *Frontiers in Genetics*, **3**(59), 2012, 1–10.

[12] Ray, S. S., Pal, J. K., Pal, S. K.: Computational Approaches for Identifying Cancer miRNA Expressions, *Gene Expression*, **15**(5-6), 243–253.

[13] Ray, S. S., Pal, S. K.: RNA Secondary Structure Prediction using Soft Computing, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2012 (accepted).

[14] Resnick, K. E., Alder, H., Hagan, J. P., Richardson, D. L., Croce, C. M., Cohn, D. E.: The detection of differentially expressed microRNAs from the serum of ovarian cancer patients using a novel real-time PCR platform, *Gynecologic Oncology*, **112**, 2009, 55–59.

[15] Schaefer, A., Jung, M., Mollenkopf, H.-J., Wagner, I., Stephan, C., Jentzmik, F., Miller, K., Lein, M., Kristiansen, G., Jung, K.: Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma, *International Journal of Cancer*, **126**(5), 2010, 1166-1176.

[16] Schrauder, M. G., R. Strick, R. D. S.-W., Strissel, P. L., Kahmann, L., Loehberg, C. R., Lux, M. P., Jud, S. M., A. Hartmann, A. H., Bayer, C. M., Bani, M. R., Richter, S., Adamietz, B. R., Wenkel, E., Rauh, C., Beckmann, M. W., Fasching, P. A.: Circulating Micro-RNAs as Potential Blood-Based Markers for Early Stage Breast Cancer Detection, *PLoS ONE*, **7**(1), 2012, e29770.

[17] Wang, Q., Wang, S., Wang, H., Li, P., Ma, Z.: MicroRNAs: novel biomarkers for lung cancer diagnosis, prediction and treatment, *Experimental Biology and Medicine*, **237**, 2012, 227–235.