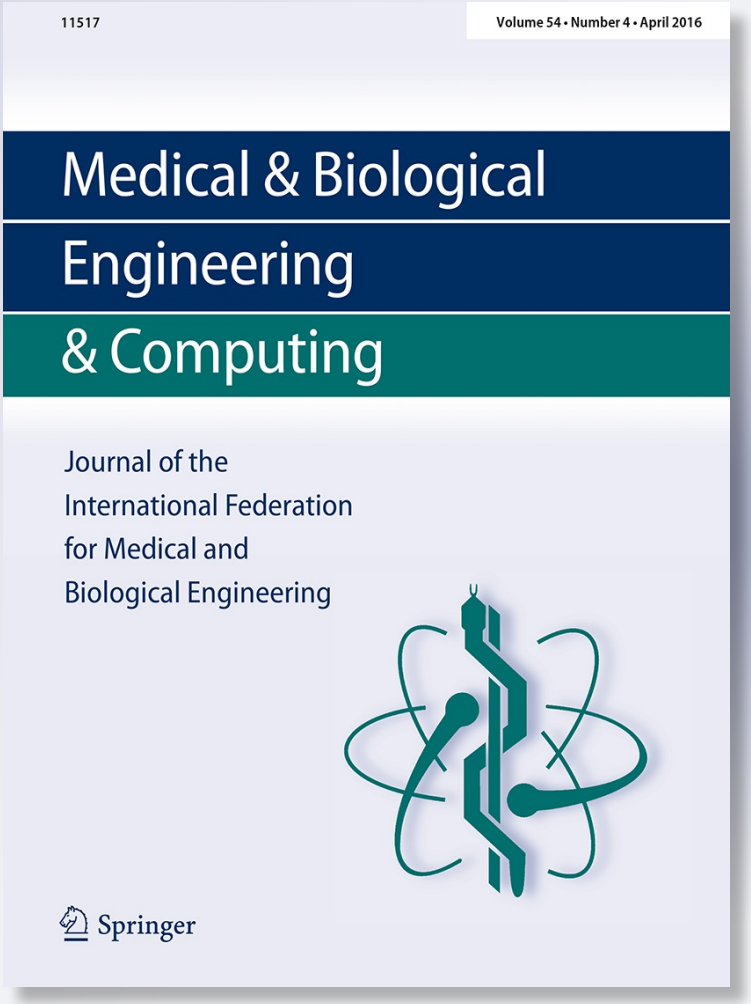
Identifying relevant group of miRNAs in cancer using fuzzy mutual information

Jayanta Kumar Pal, Shubhra Sankar Ray & Sankar K. Pal

Medical & Biological Engineering & Computing

ISSN 0140-0118
Volume 54
Number 4

Med Biol Eng Comput (2016) 54:701-710
DOI 10.1007/s11517-015-1360-1



Springer

Med Biol Eng Comput (2016) 54:701–710
DOI 10.1007/s11517-015-1360-1

ORIGINAL ARTICLE

Identifying relevant group of miRNAs in cancer using fuzzy mutual information

Jayanta Kumar Pal[1] · **Shubhra Sankar Ray**[2] · **Sankar K. Pal**[2]

Received: 4 February 2015 / Accepted: 21 July 2015 / Published online: 12 August 2015

Abstract MicroRNAs (miRNAs) act as a major biomarker of cancer. All miRNAs in human body are not equally important for cancer identification. We propose a methodology, called FMIMS, which automatically selects the most relevant miRNAs for a particular type of cancer. In FMIMS, miRNAs are initially grouped by using a SVM-based algorithm; then the group with highest relevance is determined and the miRNAs in that group are finally ranked for selection according to their redundancy. Fuzzy mutual information is used in computing the relevance of a group and the redundancy of miRNAs within it. Superiority of the most relevant group to all others, in deciding normal or cancer, is demonstrated on breast, renal, colorectal, lung, melanoma and prostate data. The merit of FMIMS as compared to several existing methods is established. While 12 out of 15 selected miRNAs by FMIMS corroborate with those of biological investigations, three of them viz., "hsa-miR-519," "hsa-miR-431" and "hsa-miR-320c" are possible novel predictions for renal cancer, lung cancer and melanoma, respectively. The selected miRNAs are found to be involved in disease-specific pathways by targeting various genes. The method is also able to detect the responsible miRNAs even at the primary stage of cancer. The related code is available at http://www.jayanta.droppages.com/FMIMS.html.

Keywords miRNA · Cancer · Bioinformatics · Fuzzy information measure · Soft computing

✉ Jayanta Kumar Pal
jkp_it08@yahoo.com

Shubhra Sankar Ray
shubhra@isical.ac.in

Sankar K. Pal
sankar@isical.ac.in

[1] Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India

[2] Center for Soft Computing Research, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

1 Introduction

Early detection of cancer [12, 16, 30] and its treatment [32] before metastasis can increase the survival rate and time of the cancer patients. Various investigations [2, 3] in this domain identified microRNAs (miRNAs) as important indicators of cancers in human body. MiRNAs are non-coding RNAs [34], and they work on messenger RNAs (mRNA) to inhibit protein translation by degrading the mRNAs [28]. All the miRNAs present in the body are not responsible for cancers, and the role of various miRNAs is diverse for different types of cancers [1, 2, 11]. Presence of any irrelevant miRNA may decrease the classification accuracy and increase both the biochemical and computational costs. Therefore, selection of most informative miRNAs is important for identifying the condition of a sample/patient.

In the existing investigations [14, 21], emphasis is given on the ranking and selection of miRNAs. Navon et al. [21] proposed a ranking method on the basis of the fold change between paired samples (i.e., both the normal and cancer tissues were collected from the same patient). In real life, it may happen that one has to classify an unknown expression of a miRNA, where the paired sample is not available. In that case, the method proposed by Navon et al. [21] can be reframed by considering the difference of miRNA expressions between unpaired samples as well. In the same investigation [21], those miRNAs were given importance which were globally deregulated in eight types of cancers. However,

there are some miRNAs which are very important only for a particular type of cancer [4, 17, 23, 24, 27]. Hence, identifying those miRNAs is more useful in cancer detection than dealing with the group of globally deregulated miRNAs for multiple cancers. Leidinger et al. [14] applied three hypothesis test-based algorithms, and total 213 miRNAs were found to be responsible for melanoma. Finally, fold change of each miRNAs was checked, and 51 miRNAs were selected as the deregulated miRNAs. Here, further contribution can be made in terms of the removal of redundant miRNAs.

Besides miRNA selection methods, existing gene selection algorithms can be useful for miRNA selection. Hence, it will be prudent to have some idea about these algorithms also. Guyon et al. [8] used a SVM classifier-based recursive feature elimination technique (SVMRFE) for gene selection. However, the problem of removing the redundant genes was not addressed. Peng et al. [26] described a gene ranking method based on maximum relevance and minimum redundancy (MRMR) using mutual information. A combined approach of SVMRFE and MRMR was reported in [20]. The method is based on the trade-off between the rankings obtained by SVMRFE and MRMR. The methodology to calculate mutual information between two fuzzy sets (fuzzy mutual information) is proposed by Maji et al. [18]. Here, the fuzzy mutual information (FMI) is used to rank the genes according to the maximum relevance and minimum redundancy.

The objective of the present investigation is to detect the relevant miRNAs to predict cancerous expressions in an unknown patient. The method involves generation of different groups of miRNAs using SVM, followed by selection of the most relevant group using FMI and removal of redundancy therein, if required.

The rest of the article is organized as follows. A brief description about FMI and the details of the proposed methodology are described in Sect. 2. The experimental results along with the biological relevance of selected miRNAs are provided in Sect. 3. Section 4 provides a brief discussion on this investigation. Section 5 concludes this article.

2 Materials and methods

In this section, first we briefly describe the concept of FMI [18] (the details are available in Section 1 of supplementary material at http://www.isical.ac.in/~shubhra/FMIMSsupplementary.pdf) which is used in the proposed methodology for calculating the relevance and redundancy. Next the proposed methodology is explained. For our study, we collected the miRNA expressions data from Gene Expression Omnibus (GEO), an international public repository. These data sets are reported in the investigations [1, 2, 10, 11, 14] and [31] by taking care of the proper ethical issues and then submitted in the GEO.

2.1 Fuzzy mutual information

Consider A as a fuzzy attribute set in a finite set U, d as the number of fuzzy equivalence classes and t as the total number of objects in U. Now, the fuzzy equivalent partition matrix (M_A) is denoted as

$$M_A = \begin{pmatrix} \mu_{11}^A & \mu_{12}^A & \cdots & \mu_{1t}^A \\ \mu_{21}^A & \mu_{22}^A & \cdots & \mu_{2t}^A \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{d1}^A & \mu_{d2}^A & \cdots & \mu_{dt}^A \end{pmatrix} \tag{1}$$

where M_A is a $d \times t$ matrix, $\sum_{u=1}^{d} \mu_{uv}^A = 1 \forall v$, and $\mu_{uv}^A \in [0, 1]$ represents the membership of the vth object in the uth fuzzy equivalence class F_u. Each row of the matrix M_A represents a fuzzy equivalence class in U, and each column represents an object in U. The fuzzy relative frequency is represented by λ_{F_u}, and it is defined as

$$\lambda_{F_u} = \frac{1}{t} \sum_{v=1}^{t} \mu_{uv}^A. \tag{2}$$

The entropy of the fuzzy attribute set A is defined as [18]

$$H(A) = -\sum_{u=1}^{d} \left[\frac{1}{t} \sum_{v=1}^{t} \mu_{uv}^A \right] \log \left[\frac{1}{t} \sum_{v=1}^{t} \mu_{uv}^A \right]. \tag{3}$$

The joint entropy $H(P, Q)$ for two fuzzy sets P and Q is defined as [18]

$$H(P,Q) = -\sum_{a=1}^{p} \sum_{b=1}^{q} \left[\frac{1}{t} \sum_{c=1}^{t} (\mu_{ac}^P \cap \mu_{bc}^Q) \right] \times \log \left[\frac{1}{t} \sum_{c=1}^{t} (\mu_{ac}^P \cap \mu_{bc}^Q) \right]. \tag{4}$$

The mutual information (FMI) between the two fuzzy sets P and Q is represented as

$$I(P,Q) = H(P) + H(Q) - H(P,Q). \tag{5}$$

Using Eqs. 3 and 4 in Eq. 5, we get

$$\begin{aligned} I(P,Q) = &-\sum_{a=1}^{p} \left[\frac{1}{t} \sum_{c=1}^{t} \mu_{ac}^P \right] \log \left[\frac{1}{t} \sum_{c=1}^{t} \mu_{ac}^P \right] \\ &-\sum_{b=1}^{q} \left[\frac{1}{t} \sum_{c=1}^{t} \mu_{bc}^Q \right] \log \left[\frac{1}{t} \sum_{c=1}^{t} \mu_{bc}^Q \right] \\ &+\sum_{a=1}^{p} \sum_{b=1}^{q} \left[\frac{1}{t} \sum_{c=1}^{t} (\mu_{ac}^P \cap \mu_{bc}^Q) \right] \\ &\times \log \left[\frac{1}{t} \sum_{c=1}^{t} (\mu_{ac}^P \cap \mu_{bc}^Q) \right]. \end{aligned} \tag{6}$$

2.2 Proposed method

As stated earlier, the present investigation deals with the problem of miRNA selection for cancerous miRNA classification. A new method, called FMI for miRNA selection (FMIMS), is described here. The block diagram is shown in Section 2 of the supplementary material.

2.2.1 Grouping

The methodology works by grouping miRNAs whose normal and cancer class representatives (see Eqs. 14, 15) are separable by the same class boundary. The main steps for this method are:

S1. Calculate the distance (d_{ij}^k) between ith normal and jth cancer sample of kth miRNA as

$$d_{ij}^k = |x_i^k - y_j^k|; \forall i, j \tag{7}$$

where x_i^k and y_j^k are the ith normal and jth cancer expressions of kth miRNA, respectively, $1 \le i \le N, 1 \le j \le M$, $1 \le k \le L$ and N, M and L are total number of normal samples, cancer samples and miRNAs, respectively.

S2. For kth miRNA, determine the scaled average interclass distance (z^k) between normal and cancer class as

$$z^k = \frac{1}{\sigma^k MN} \sum_{i=1}^{N} \sum_{j=1}^{M} d_{ij}^k \tag{8}$$

where σ^k represents the standard deviation of d_{ij}^k for all i and j.

S3. Compute the distance between two normal expressions of kth miRNA as

$$w_{i_1 i_2}^k = |x_{i_1}^k - x_{i_2}^k|; \forall i_1, i_2 \tag{9}$$

where $i_1 = 1, 2, \ldots, N-1$ and $i_2 = i_1 + 1, i_1 + 2, \ldots, N$

S4. For kth miRNA, compute the scaled average intraclass distance (h^k) of normal class as

$$h_n^k = \frac{1}{\sigma_1^k [(N-1) + (N-2) + \cdots + 1]} \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} (w_{i_1 i_2}^k) \tag{10}$$

where standard deviation of $w_{i_1 i_2}^k$, over all i_1 and i_2, is represented by σ_1^k.

S5. Similarly, calculate the scaled intraclass distance (h_c^k) of cancer class corresponding to the kth miRNA as

$$h_c^k = \frac{1}{\sigma_2^k [(M-1) + (M-2) + \cdots + 1]} \sum_{j_1=1}^{M-1} \sum_{j_2=j_1+1}^{M} (l_{j_1 j_2}^k). \tag{11}$$

where $l_{j_1 j_2}^k$ is the distance between two cancer expressions and is represented as

$$l_{j_1 j_2}^k = |y_{j_1}^k - y_{j_2}^k|; \forall j_1, j_2. \tag{12}$$

Here, $j_1 = 1, 2, \ldots, M-1$, $j_2 = j_1 + 1, j_1 + 2, \ldots, M$ and σ_2^k represents the standard deviation of $l_{j_1 j_2}^k$ computed over all j_1 and j_2.

S6. Calculate the class separability index (α^k) of the kth miRNA as

$$\alpha^k = \frac{z^k}{h_n^k + h_c^k} \tag{13}$$

S7. Repeat Steps S1 to S6 for all k ($1 \le k \le L$) and sort all the miRNAs in descending order according to the value of α^k. The miRNA with highest α^k value is the top-ranked one among all those in the data set.

S8. Determine the representative for the kth miRNA of each class (say, r_n^k for normal and r_c^k for cancer) as

$$r_n^k = \frac{1}{\sigma_n^k N} \sum_{i=1}^{N} x_i^k \quad \text{and} \tag{14}$$

$$r_c^k = \frac{1}{\sigma_c^k M} \sum_{j=1}^{M} y_j^k, \tag{15}$$

where σ_n^k and σ_c^k represent standard deviations of the normal and cancer expression values, respectively.

S9. Repeat Step S8 for all miRNAs.

S10. Select the top-ranked miRNA (initial group point) among ungrouped miRNAs and train SVM (with linear kernel) using r_n^k and r_c^k of the top-ranked one.

S11. From the remaining miRNAs, find those r_n^k and r_c^k values which are correctly classified by the trained SVM. Assign those miRNAs to the group of the top-ranked miRNA.

S12. Repeat Steps S10 and S11 for the remaining miRNAs until all miRNAs are assigned to a group.

2.2.2 Selection of most relevant group

Let G number of groups be generated by the grouping technique (see Sect. 2.2.1) and the kth miRNA belongs to the gth group where $1 \le g \le G$. The steps for selecting the most relevant group are as follows:

S1. Compute the relevance of the kth ($1 \le k \le L$) miRNA by calculating $I(R^k, D^k)$ (see Eq. 6), where R^k is the set of membership values (see Section 5 in supplementary

Table 1 *F* score values achieved for all the miRNAs and the miRNAs in the top two groups corresponding to different data sets using SVM and *k*-NN classifiers

Data set	Total samples/patients	Input miRNAs (no. and *F* score)			miRNAs in most relevant group (no. and *F* score)			miRNAs in second relevant group (no. and *F* score)		
		No.	SVM	*k*-NN	No.	SVM	*k*-NN	No.	SVM	*k*-NN
Breast	98	309	0.60	0.18	2	0.86	0.66	5	0.68	0.29
Renal	24	12,033	0.56	0.56	2	0.83	0.83	2	0.81	0.76
Colorectal	66	352	0.61	0.52	2	0.74	0.70	2	0.70	0.31
Lung	36	866	0.52	0.45	3	0.79	0.79	2	0.68	0.58
Melanoma	57	866	0.61	0.31	4	0.76	0.72	121	0.69	0.65
Prostate	24	12,033	0.52	0.58	2	0.72	0.76	2	0.58	0.66

material) of all patients belonging to kth miRNA and D^k is the set of membership values of the class label (i.e., normal or cancer) of the same patients.

S2. Repeat Step S1 for all the miRNAs in the gth group and calculate the average relevance of the miRNAs in that group.
S3. Repeat Steps S1–S2 for all the groups.
S4. Select the group with highest average relevance value of the miRNAs.

2.2.3 *Removal of redundant miRNAs*

The removal of redundant miRNAs in a group is optional, but it helps in reducing both the biochemical and computational costs. The steps for removing the redundant miRNAs are as follows:

S1. Calculate the redundancy $I(R^k, R^{k'})$ of kth miRNA with respect to k'th ($k \neq k'$) miRNA in the most relevant group using Eq. 6.
S2. Repeat Step S1 for all values of k' and calculate the average redundancy value of the kth miRNA. Here, $1 \leq k' \leq L$ and L is the number of miRNAs in the most relevant group.
S3. Repeat Steps S1–S2 for all k to determine the average redundancy of each miRNA in the selected group.
S4. Rank the miRNAs according to redundancy value.
S5. Select miRNAs with low redundancy according to user's need.

3 Results

Six data sets, viz. breast [2], renal [10], colorectal [1], lung [11], melanoma [14] and prostate [31], are used in this investigation to demonstrate the effectiveness of the FMIMS (the details of these data sets are available in Section 3 of supplementary file at http://www.isical.ac.in/~shubhra/FMIMSsupplementary.pdf). The classification performance of the selected miRNAs is computed using SVM and *k*-NN separately in terms of sensitivity, specificity, *F* score [27] and accuracy (see Section 4 in supplementary material for these measures). Linear kernel is used for the SVM classifier. The value of '*k*' in the *k*-NN classifier is varied from 1 to K where $K = \min(N, M)$, and N and M are the total number of normal and cancer samples, respectively. The best *F* score among all those obtained with different values of '*k*' is reported. After the selection of the group of miRNAs, its performance for cancer classification is evaluated by leave-one-out cross-validation method. The procedure for calculating the membership values is described in Section 5 of the supplementary material.

3.1 Performance evaluation

Here we evaluate the performance of FMIMS in terms of *F* score. Table 1 shows the *F* scores related to various groups of miRNAs. It can be observed from the table that the *F* score value is considerably improved for the miRNAs in the most relevant group as compared to the total set of miRNAs. The number of miRNAs in this group is also considerably less than that of the total miRNAs. *F* score of the most relevant group varies from 0.72 to 0.86 and 0.66 to 0.83 for SVM and *k*-NN classifiers, respectively. These values are also found to be higher than those of the other groups for all the data sets.

In Table 2 we report the selected miRNAs (i.e., the miRNAs in the most relevant group) by FMIMS. Biological relevance of these miRNAs and their involvement in disease specific pathways are discussed in Sect. 4.

The efficacy of the redundancy removal method is demonstrated on some groups with relatively large number of miRNAs. The results are reported in Table 3. As seen, the *F* score values even with the 50% miRNAs are found to be almost the same with those obtained with all the miRNAs in that group. Note that the redundancy removal method is not applied on the most relevant group as it contains only very few miRNAs (see Table 1).

Table 2 Selected miRNAs for different data sets using FMIMS

	Data set					
	Breast	Renal	Colorectal	Lung	Melanoma	Prostate
miRNA	hsa-mir-30a hsa-miR-10a	hsa-miR-15a **hsa-miR-519e**	hsa-miR-103 hsa-miR-125a	**hsa-miR-431** hsa-miR-200b hsa-miR-22	**hsa-miR-320c** hsa-miR-943 hsa-miR-199a hsa-miR-127-3p	hsa-miR-505 hsa-mir-181b

The novel predictions for various data sets are marked in bold font

Table 3 Classification performance after redundancy removal on some groups corresponding to different data sets

Data set	Group no.	Before redundancy removal			After redundancy removal		
		No. of miRNAs	*F* score		No. of miRNAs	*F* score	
			SVM	*k*-NN		SVM	*k*-NN
Breast	2	5	0.68	0.23	3	0.66	0.21
Renal	3	28	0.71	0.69	14	0.67	0.60
Colorectal	4	40	0.61	0.33	20	0.60	0.32
Lung	4	11	0.57	0.68	6	0.52	0.61
Melanoma	2	121	0.65	0.64	61	0.64	0.60
Prostate	8	4	0.53	0.52	2	0.50	0.47

The *F* score values are found to be almost the same even after the removal of redundant miRNAs

3.2 Comparison with other approaches

In this section, we compare the performance of our method with several well-known methods. The methods considered are SVMRFE [8], MRMR [26], SVMRFE with MRMR [20] and the method by Maji et al. [18] using FMI (FRSIM).

In Table 4, the performance of the miRNAs selected by FMIMS is compared with those achieved by the other methods in terms of sensitivity and specificity, *F* score and accuracy. For the sake of fair comparison, the number of top-ranked miRNAs obtained by different methods is kept the same as obtained by FMIMS. The number of miRNAs is 2, 2, 2, 3, 4 and 2 for breast, renal, colorectal, lung, melanoma and prostate data, respectively. It is observed from the table that FMIMS performs the best in terms of sensitivity, specificity, *F* score and accuracy in most of the cases. The cases where other methods perform better than FMIMS can be easily seen by observing the results in bold font corresponding to the rows for different methods in the table. In summary, out of 48 (=6 datasets × 2 classifiers × 4 performance measures) comparisons, only in seven cases the FMIMS performs inferior.

We also compared the performance of FMIMS with related algorithms by varying the number of selected miRNAs. The experimental results for breast cancer using *k*-NN are shown in Fig. 1. Similar curves are also obtained from the other data sets and shown in Section 6 of the supplementary file.

3.3 Effectiveness of fuzzy mutual information

Let us now compare the effectiveness of FMI over some other fuzzy information measures and a non-fuzzy version of mutual information in selecting the most relevant group of miRNAs. Other fuzzy information measures considered are fuzzy *V* information measure (FVI) and fuzzy χ^2 information measure (FCI) [18]. In case of non-fuzzy version, mutual information is computed with Parzen window method (PMI).

The comparisons among the miRNAs selected by FMI, FVI, FCI and PMI in terms of *F* scores corresponding to SVM and *k*-NN classifiers are shown in Fig. 2a, b, respectively. It is evident from the figures that the best *F* scores for different data sets are obtained using FMI.

4 Discussion

In this investigation, we have emphasized on the selection of important miRNAs. In this section, we discuss the significance of the miRNAs, selected by the proposed FMIMS (Table 2), to the related cancers. For example, in breast cancer, "hsa-mir-30a" and "hsa-miR-10a" are found as the important ones. Our computational findings are similar to those investigations by Ouzounova et al. [23], Cheng et al. [5] and Zhang et al. [37] where they have pointed out that miRNA "hsa-mir-30a" targets the metadherin (MTDH), FOXD1 and AVEN genes and plays

Table 4 Comparison of classification performance among different miRNA/gene selection methods in terms of sensitivity, specificity, *F* score and accuracy using SVM and *k*-NN classifiers

Method	Measures	Breast		Renal		Colorectal		Lung		Melanoma		Prostate	
		SVM	*k*-NN	SVM	*k*-NN	SVM	*k*-NN	SVM	*k*-NN	SVM	*k*-NN	SVM	*k*-NN
FMIMS	Sensitivity	**0.83**	**0.95**	**0.87**	**0.87**	0.65	0.64	**0.79**	**0.79**	**0.75**	**0.71**	**0.75**	**0.92**
	Specificity	**0.90**	0.51	**0.80**	**0.80**	**0.87**	**0.78**	**0.79**	**0.79**	0.78	0.73	**0.72**	0.62
	F score	**0.86**	**0.66**	**0.83**	**0.83**	**0.74**	**0.70**	**0.80**	**0.79**	**0.76**	**0.72**	**0.72**	0.76
	Accuracy	**86.67**	**73.46**	**83.33**	**83.33**	**74.68**	**70.45**	**79.00**	**79.00**	**76.55**	**72.50**	**73.11**	**77.08**
MRMR	Sensitivity	0.74	0.71	0.83	0.74	0.56	0.77	0.59	0.46	0.50	0.39	**0.75**	0.79
	Specificity	0.60	0.48	0.62	0.77	0.50	0.56	0.57	0.79	**0.86**	**0.84**	0.71	**0.75**
	F score	0.66	0.57	0.71	0.75	0.53	0.65	0.58	0.49	0.63	0.53	**0.72**	**0.77**
	Accuracy	66.83	59.38	72.92	75.59	53.02	66.55	59.10	48.87	67.46	61.63	72.91	**77.08**
SVMRFE	Sensitivity	0.24	0.90	0.58	0.37	0.52	**0.84**	0.51	0.41	0.66	0.57	0.33	0.54
	Specificity	0.38	0.12	0.37	0.37	0.56	0.19	0.67	0.58	0.70	0.48	0.58	0.50
	F score	0.29	0.21	0.46	0.37	0.54	0.31	0.58	0.46	0.68	0.53	0.42	0.52
	Accuracy	30.88	51.57	47.91	37.00	53.90	51.18	58.82	50.18	68.56	53.09	45.83	52.08
SVMRFE with MRMR	Sensitivity	0.53	0.62	0.56	0.54	0.69	0.32	0.36	0.59	0.54	0.56	0.45	0.55
	Specificity	0.60	**0.68**	0.75	0.75	0.69	0.30	0.43	0.68	0.52	0.70	0.42	0.50
	F score	0.56	0.65	0.64	0.62	0.69	0.31	0.39	0.63	0.53	0.62	0.43	0.52
	Accuracy	56.45	65.20	65.41	64.58	68.86	31.11	39.56	63.57	53.26	63.20	43.70	52.62
FRSIM	Sensitivity	0.60	0.90	0.63	0.72	**0.78**	0.67	0.60	0.66	0.74	0.70	0.74	0.74
	Specificity	0.80	0.38	0.79	0.79	0.71	0.62	0.52	0.55	0.67	0.54	0.71	0.72
	F score	0.69	0.53	0.70	0.74	**0.74**	0.64	0.56	0.60	0.71	0.61	**0.72**	0.72
	Accuracy	70.10	53.44	71.45	75.50	74.36	64.50	56.10	60.52	71.82	62.00	72.40	72.40

For a data set, the best result achieved by any method is marked by bold font

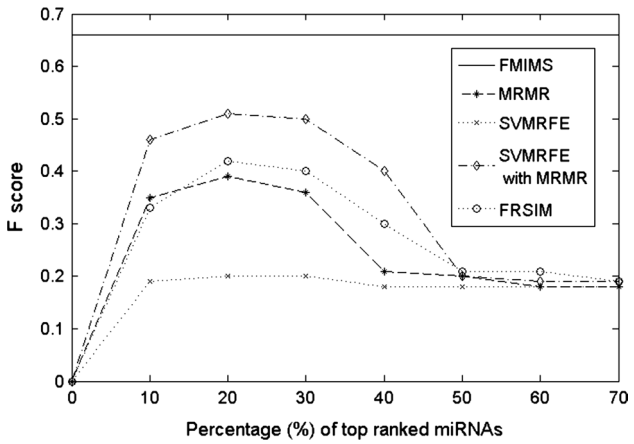


Fig. 1 Comparison of classification performance in terms of *F* score for different methods using various percentage of miRNAs in breast cancer data set. The percentage of top miRNAs is varied from 10 to 70, in steps of 10, for different methods. The performance of FMIMS is shown by a *straight line* ($y = 0.66$) parallel to the x axis as it is constant for a particular data set where the miRNAs are automatically selected through the most relevant group

a major role in the growth of breast tissues. The involvement of the miRNA "hsa-miR-10a" in amplifying the genes related to breast cancer is reported by Lund et al. [17]. The miRNA "hsa-miR-15a" is selected as one of the responsible miRNAs for renal cancer, and our findings are in agreement with the investigation by Wulfken et al. [35]. The most relevant group for colorectal cancer data contains the miRNAs "hsa-miR-103" and "hsa-miR-125a." Interestingly, the investigation by Chen et al. [4] revealed that "hsa-miR-103" regulates the expressions of DAPK and KLF4 genes and acts as a metastasis suppressor in colorectal cancer, and upregulation of "hsa-miR-125a" is found to be responsible by Nishida et al. [22] for the same disease. Similarly for lung cancer, upregulation of "hsa-miR-22" is identified as one of the reasons by Keller et al. [11]. The miRNA "hsa-miR-200b" plays an important role in lung cancer chemotherapy [29]. Upregulation of "hsa-miR-199a" and "hsa-miR-943" [7, 19] and downregulation of "hsa-miR-127-3p" [36] are found responsible for melanoma. The investigations by Feng et al. [6] and He et al. [9] reported that upregulation of "hsa-miR-505" and "hsa-mir-181b," respectively, causes prostate cancer by promoting cell proliferation in prostate gland, and our miRNA selection methodology also identified them as the relevant ones.

FMIMS also predicted some miRNAs which target cancer-related genes. For example in the database [13] published by Laganá et al, the miRNA "hsa-miR-519e" is reported as one of the responsible miRNAs for kidney

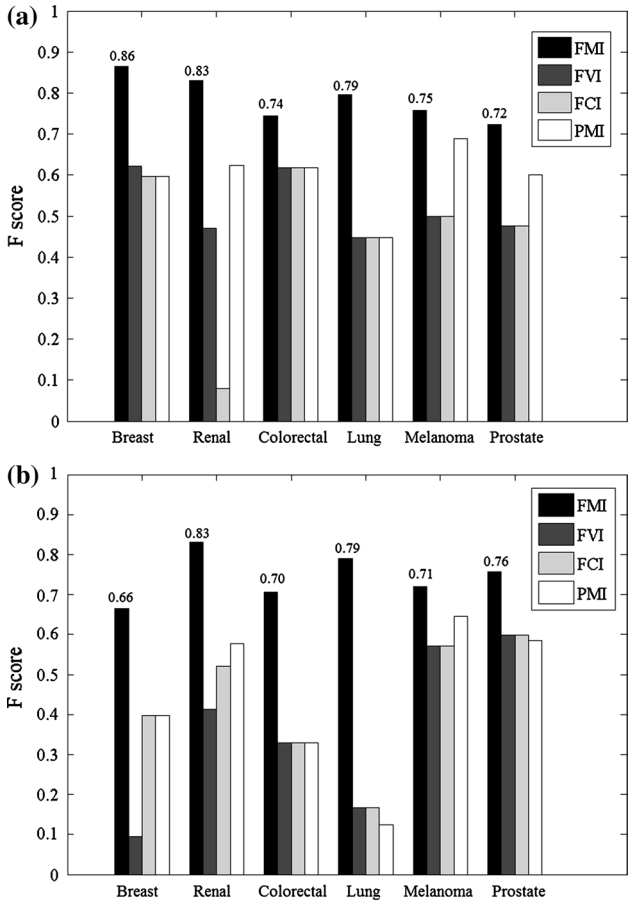


Fig. 2 Comparison between FMI, FVI, FCI and PMI in terms of *F* score using the selected miRNAs for different data sets. **a**, **b** shows the performance of these information measures using SVM and *k*-NN, respectively, **a** performance with SVM classifier, **b** performance with *k*-NN classifier

cancer by regulating the ABCG2 gene. Our investigation selects the same miRNA for renal cancer, the most common type of kidney cancer that originates in the lining of the proximal convoluted tubule. Similarly, in the same database, "hsa-miR-431" and "hsa-miR-320c" are mentioned as responsible for targeting 12 and 4 genes for lung cancer and melanoma, respectively. Interestingly, the same miRNAs are also identified as the most relevant ones for the same cancers by FMIMS. Therefore, the miRNAs "hsa-miR-519e," "hsa-miR-431" and "hsa-miR-320c" may be viewed as novel computational predictions by FMIMS.

Among the six data sets used in our study, while the breast cancer data [2] deals only with early stage of cancer, the colorectal cancer data [1] consist of "Duke A" (early stage), "Duke B," "Duke C" and "Duke D" stages of cancer. The results of breast cancer data are already discussed before. Further, in a part of the investigation, we performed experiments only with Duke A stage of colorectal cancer and found three miRNAs "has-miR-130b," "has-miR-422b" and "has-miR-501" as important for early stage detection. The classification accuracies achieved by these miRNAs are 83.75 and 82.50 % corresponding to SVM and *k*-NN classifier. As our method is also capable of identifying the relevant miRNAs from these data sets (viz. breast and colorectal cancer), the algorithm is suitable for handling the cancers before metastasis.

The miRNAs identified by the proposed FMIMS are also evaluated for their involvement in disease-specific pathways by using the pathway analysis tools, namely DIANA [25, 33] and Starbase [15]. Given a miRNA, these tools can identify the pathways through the target genes. The tools provide a merged p value by combining all the p values of the target genes for a specific miRNA through Fisher's combined probability method [33]. Out of 15 miRNAs identified by FMIMS, 12 miRNAs are found to be involved in disease-specific pathways. Among these 12, 8 miRNAs are identified with p value <0.01 for targeting genes and the remaining 4 are identified with p value ranging from 0.06 to 0.09. For example, the miRNA "hsa-miR-22" is involved in lung cancer pathway where the target genes have the merged p value of 1.21×10^{-5}. In a similar way, the remaining miRNAs are also observed to be related with specific disease pathways with different p values, as mentioned in Table 5.

In our methodology, certain criteria are followed to achieve more accurate selection of miRNA than those in [8, 20, 21, 26] and [18]. For example, unlike the investigation by Navon et al. [21], we used samples without considering any pair (pair: normal and cancer samples from the same patient) between them and considered a particular cancer for selecting miRNAs rather than selecting a group of globally deregulated miRNAs. Improvement is also made over the related investigations [18, 20, 26] by automatic selection of the most relevant miRNAs and optional removal of redundant miRNAs. Optional removal of redundancy helps the user to keep all the relevant miRNAs, if necessary. In our methodology, all the non-cancerous patients (i.e., normal samples and benign tumors) are considered to form one class (normal) and all the cancer samples (i.e., different stages of cancers) as another class (cancer). Therefore, the miRNAs selected by our method will also be able to separate malignant tumors (as cancer class) from benign tumors (as normal ones), while classifying these two classes.

As our target is to select a group of miRNAs which provides maximum class separability between normal and cancer expressions, initially the miRNAs are divided into several groups and then the most relevant group is selected in terms of class separability by using the FMI theory. So, the most relevant group should contain the optimum set of miRNAs for classification. However, there are two issues concerning the grouping:

Table 5 Pathway analysis of the miRNAs selected by FMIMS using DIANA and Starbase

Pathway	miRNA	No. of target genes	p Value related to gene targeting
Breast	hsa-mir-30a	3	6.00×10^{-3}
	hsa-miR-10a	9	8.97×10^{-5}
Renal	hsa-miR-15a	10	1.50×10^{-3}
	hsa-miR-519e	1	9.00×10^{-2}
Colorectal	hsa-miR-125a	17	3.41×10^{-5}
Lung	hsa-miR-431	1	6.00×10^{-2}
	hsa-miR-22	10	1.21×10^{-5}
	hsa-miR-200b	12	4.54×10^{-5}
Melanoma	hsa-miR-320c	11	1.70×10^{-5}
	hsa-miR-943	1	7.00×10^{-2}
Prostate	hsa-miR-505	1	7.00×10^{-2}
	hsa-mir-181b	2	5.30×10^{-4}

The selected miRNAs, number of related target genes and corresponding p values are shown in different columns

1. The patients have labels (i.e., normal and cancer), but miRNAs do not. Therefore, grouping of miRNAs stands out to be an unsupervised task, where miRNAs are considered as patterns and patients are the features. Further, a conventional unsupervised (clustering) method considers normal and cancer expressions within a single profile vector in order to compute the similarity between two miRNAs. Hence it would be unable to capture the existing class difference information between normal and cancer which is required to classify an unknown miRNA expression finally.
2. Patients are of two types (normal and cancer) and their number in each type is different. Therefore, computing the similarity/distance between normal and cancerous classes is not feasible.

   To address the first issue, one may need to use a supervised classifier which can take into account the class difference information during training and then group the unlabeled miRNAs accordingly. As the class difference information is preserved within the resulting groups, the miRNAs in the most relevant group is therefore expected to be most appropriate for classification of an unknown miRNA. Since SVM is efficient for two-class classification, it is used in our investigation. To address the second issue, we have considered point representatives for each class, computed from all the expression values of that class. Moreover, the normal and cancerous expressions of a miRNA may overlap with each other. Similarly, expressions of two miRNAs may overlap. This is handled using FMI measures which takes care of the former overlapping in computing the relevance of a miRNA, while the latter one in computing the redundancy between two miRNAs. Thus, the novelty of our method lies in demonstrating a way to handle unpaired samples using SVM and FMI for finding informative miRNAs associated with a cancer.

5 Conclusions

In this investigation, a method for selection of cancerous miRNA is explained. It consists of three steps, grouping of miRNAs, selection of the most relevant group and removal of redundancy from the selected group. The number of groups is automatically identified by the grouping algorithm. The miRNAs in the most relevant group are considered as the best group in terms of the separation between their normal and cancer expressions for a particular type of cancer. They also provide better classification accuracy as compared to the miRNAs in other groups. The most relevant miRNAs are ranked according to their redundancy which is used as a selection criterion, if the number of miRNAs in the relevant group is greater than the user-defined one.

The comparative study demonstrates the superiority of the FMI measure in terms of selection of the most relevant group. It is also seen that the FMIMS performs better than some other methods for most of the data sets. All the selected miRNAs by FMIMS are found to be relevant according to related biological investigations or database. It is evident from the breast and colorectal data sets, where expressions for early stage of cancers are also available, that our algorithm can identify the relevant miRNAs even before metastasis. There are various stages of cancer where FMIMS may be useful to diagnose them. In particular, it is very challenging and important to detect cancer before it metastasize. The experimental findings on multiple data sets and their relevance to those of biological experiments reveal the significance of this investigation.

Acknowledgments S. K. Pal acknowledges the J. C. Bose fellowship of the Government of India and the INAE Chair Professorship.

References

1. Arndt GM et al (2009) Characterization of global microRNA expression reveals oncogenic potential of mir-145 in metastatic colorectal cancer. BMC Cancer 9:374–390
2. Blenkiron C et al (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 8:R214.1–R214.16
3. Calin GA et al (2002) Frequent deletions and down-regulation of micro-RNA genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci 99:15524–15529
4. Chen HY et al (2012) miR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4. J Cancer Res 72:3631–3641

5. Cheng CW et al (2012) MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. Breast Cancer Res Treat 134:1081–1093
6. Feng J et al (2013) Screening biomarkers of prostate cancer by integrating microRNA and mRNA microarrays. Genet Test Mol Biomark 17:807–813
7. Greenberg E et al (2011) Regulation of cancer aggressive features in melanoma cells by microRNAs. Plos One 6:e18936
8. Guyon I et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422
9. He L et al (2013) MicroRNA-181b expression in prostate cancer tissues and its influence on the biological behavior of the prostate cancer cell line pc-3. Genet Mol Res 12:1012–1021
10. Jung M et al (2009) MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy. J Cell Mol Med 13:3918–3928
11. Keller A et al (2009) miRNAs in lung cancer-studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer 9:353–362
12. Kusy M, Obrzut B, Kluska J (2013) Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients. Med Biol Eng Comput 51:1357–1365
13. Laganá A et al (2009) miro: a miRNA knowledge base. Database Oxf J 2009:1–7
14. Leidinger P et al (2010) High-throughput miRNA profiling of human melanoma blood samples. BMC Cancer 10:262–272
15. Li J-H et al (2014) starbase v2.0: decoding mirna-cerna, mirna-ncrna and proteinrna interaction networks from large-scale clip-seq data. Nucleic Acids Res 14:D92–D97
16. Lopez F et al (2012) Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. Med Biol Eng Comput 50:981–990
17. Lund AH (2010) mir-10 in development and cancer. Cell Death Differ 17:209–214
18. Maji P, Pal SK (2010) Fuzzy-rough sets for information measures and selection of relevant genes from microarray data. IEEE Trans Syst Man Cybern B Cybern 40:741–752
19. Mueller DW, Rehli M, Bosserhoff AK (2009) miRNA expression profiling in melanocytes and melanoma cell lines reveals miRNAs associated with formation and progression of malignant melanoma. J Investig Dermatol 129:1740–1751
20. Mundra PA, Rajapakse JC (2010) SVM-RFE with MRMR filter for gene selection. IEEE Trans Nanobiosci 9:31–37
21. Navon R et al (2009) Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. PLoS One 4:e8003
22. Nishida N et al (2012) Microarray analysis of colorectal cancer stromal tissue reveals upregulation of two oncogenic miRNA clusters. Clin Cancer Res 18:3054–3070
23. Ouzounova M et al (2013) MicroRNA mir-30 family regulates non-attachment growth of breast cancer cells. BMC Genomics 14:139
24. Pal JK, Ray SS, Pal SK (2013) A weighted threshold for detection of cancerous miRNA expressions. Fundam Inf 127:289–305
25. Papadopoulos GL et al (2009) Diana-mirpath: integrating human and mouse micrornas in pathways. Bioinformatics 15:1991–1993
26. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy, IEEE Trans. Pattern Anal Mach Intell 27:1226–1238
27. Ray SS, Pal JK, Pal SK (2013) Computational approaches for identifying cancer miRNA expressions. Gene Expr 15:243–253
28. Ray SS, Maiti S (2015) Noncoding RNAs and their annotation using metagenomics algorithms. Wiley Interdiscip Rev Data Min Knowl Discov 5:1–20
29. Rui W et al (2011) Identification of microRNA profiles in docetaxel-resistant human non-small cell lung carcinoma cells (spc-a1). J Cell Mol Med 14:206–214
30. Salim M et al (2013) Measurement of bioelectric and acoustic profile of breast tissue using hybrid magnetoacoustic method for cancer detection. Med Biol Eng Comput 51:459–466
31. Schaefer A et al (2010) Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. Int J Cancer 126:1166–1176
32. Su D, Ma R, Zhu L (2011) Numerical study of nanofluid infusion in deformable tissues for hyperthermia cancer treatments. Med Biol Eng Comput 49:1233–1240
33. Vlachos IS et al (2012) Diana mirpath v.2.0: investigating the combinatorial effect of micrornas in pathways. Nucleic Acids Res (Web server issue), pp W498–W504
34. Wang Z et al (2007) Unravelling the world of cis-regulatory elements. Med Biol Eng Comput 45:709–718
35. Wulfken LM et al (2011) MicroRNAs in renal cell carcinoma: diagnostic implications of serum mir-1233 levels. Cell Death Differ 6:e25787
36. Zehavi L et al (2012) Silencing of a large microRNA cluster on human chromosome 14q32 in melanoma: biological effects of mir-376a and mir-376c on insulin growth factor 1 receptor. Mol Cancer 11:44–58
37. Zhang N et al (2013) MicroRNA-30a suppresses breast tumor growth and metastasis by targeting metadherin. Oncogenet 3119–3128

Jayanta Kumar Pal received B.Tech. in Information Technology and M.Tech. in Computer Science and Engineering in 2008 and 2010, respectively, from the West Bengal University of Technology, India. He is currently working as a Senior Research Fellow in the Center for Soft Computing Research, Indian Statistical Institute, Kolkata. His research interests include soft computing and bioinformatics.

Shubhra Sankar Ray is an Associate Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, and associated also with its Center for Soft Computing Research. He received M.Sc. in Electronic Science and M.Tech. in Radio Physics and Electronics in 2000 and 2002, respectively, from University of Calcutta, and Ph.D. (Engg.) in 2008 from Jadavpur University, Kolkata. He worked as a Post-Doc Fellow at Saha Institute of Nuclear Physics, Calcutta, during 2008–2009. His current research activities are in bioinformatics, neural networks, genetic algorithms and soft computing. Three of his publications are listed as curated paper in Saccharomyces Genome

Database, Stanford University, USA. He is a recipient of the Microsoft Young Faculty Award in 2010.

Sankar K. Pal received the Ph.D. degrees from Calcutta University and Imperial College, London. He joined the Indian Statistical Institute in 1975 as a CSIR senior research fellow where he became a full professor in 1987, a distinguished scientist in 1998 and the Director in 2005. He is a J.C. Bose fellow of the Government of India and INAE Chair Professor. He founded the Machine Intelligence Unit and the Center for Soft Computing Research at the Institute in Calcutta which are enjoying international recognition. He worked at UC Berkeley and UMD, College Park, the NASA JSC, Houston, Texas, and the US Naval Research Lab, Washington, DC. He has been a distinguished visitor of the IEEE Computer Society since 1987 and held several visiting positions in Italy, Poland, Hong Kong, and Australian Universities. He is a Fellow of the IEEE, TWAS, IAPR, IFSA, and all four National Academies for science/engineering in India. He is a coauthor of 17 books and more than 400 research publications in the areas of pattern recognition and machine learning, image processing, data mining, web intelligence, soft computing, bioinformatics and cognitive machines. He is/was on the editorial boards of 20 journals including IEEE Transactions. He received several national and international awards including the most coveted S.S. Bhatnagar Prize and Padma Shri in India, and NASA Tech Brief Awards in USA and Khwarizmi International Award from Iran.