Computers in Biology and Medicine 89 (2017) 540–548

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Fuzzy mutual information based grouping and new fitness function for PSO in selection of miRNAs in cancer

Jayanta Kumar Pal [a,*], Shubhra Sankar Ray [b], Sankar K. Pal [b]

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

ARTICLE INFO

Index terms:
miRNA expression
Gene expression
Cancer
Fuzzy mutual information
Soft computing
Pattern recognition

ABSTRACT

MicroRNAs (miRNA) are one of the important regulators of cell division and also responsible for cancer development. Among the discovered miRNAs, not all are important for cancer detection. In this regard a fuzzy mutual information (FMI) based grouping and miRNA selection method (FMIGS) is developed to identify the miRNAs responsible for a particular cancer. First, the miRNAs are ranked and divided into several groups. Then the most important group is selected among the generated groups. Both the steps viz., ranking of miRNAs and selection of the most relevant group of miRNAs, are performed using FMI. Here the number of groups is automatically determined by the grouping method. After the selection process, redundant miRNAs are removed from the selected set of miRNAs as per user's necessity. In a part of the investigation we proposed a FMI based particle swarm optimization (PSO) method for selecting relevant miRNAs, where FMI is used as a fitness function to determine the fitness of the particles. The effectiveness of FMIGS and FMI based PSO is tested on five data sets and their efficiency in selecting relevant miRNAs are demonstrated. The superior performance of FMIGS to some existing methods are established and the biological significance of the selected miRNAs is observed by the findings of the biological investigation and publicly available pathway analysis tools. The source code related to our investigation is available at http://www.jayanta.droppages.com/FMIGS.html.

1. Introduction

MicroRNAs (miRNA) are small RNAs of length ~ 22 nucleotides and play a major role in cancer development. MiRNAs inhibit the protein translation process by degrading the messenger RNAs (mRNA) [1–3]. As protein plays an important role in cell signaling to perform cell proliferation and apoptosis (programmed cell death), miRNAs posses as one of the important biomarkers for cancer. In cancer, the affected cells grow uncontrollably and suppress the nearby organs; thereby causing damage to those organs resulting in death of a patient. Cancer can be detected by using miRNA expressions extracted from blood samples [4,5] instead of using any tissue sample. This makes the sample collection easier and less stressful for both the patient and doctor. Moreover, poorly differentiable cancers can also be detected by studying miRNA expressions [6]. These types of interesting characteristics attract researchers to perform various investigations on miRNA expressions in cancer. However, more than one thousand miRNAs are discovered in human body and only a portion of them are responsible for cancer. So, there is a need to select those responsible miRNAs which will help to obtain higher classification accuracy while classifying the normal and cancer expressions of an unknown patient. In other words, the selection of relevant miRNAs increases the classification accuracy and decreases the time and biochemical cost required for the miRNA expression generation at the time of diagnosing a patient.

Identification of important miRNAs for cancer detection is performed in few investigations [7,8]. Although the number of investigations are less in miRNA selection, there are several gene selection algorithms such as SVMRFE [9], MRMR [10], SVMRFE with MRMR [11] etc., which can be used for miRNA ranking/selection. From the viewpoint of the methodology, the miRNA/gene ranking or selection algorithms can be divided into several categories such as fold change based method [7], classifier based method [9], information theory based method [10,12] etc. Combination of classifier based method and information theory based method are also used in different investigations [8,11]. The algorithms regarding the relevant miRNA identification can be designed in two ways, such as (i) ranking of miRNAs as per their individual relevance to a cancer and (ii) selection of a group of miRNAs automatically. While the former one helps in visualizing the relevance of individual miRNAs corresponding to

* Corresponding author.
E-mail addresses: jkp_it08@yahoo.com (J.K. Pal), shubhra@isical.ac.in (S.S. Ray), sankar@isical.ac.in (S.K. Pal).

http://dx.doi.org/10.1016/j.compbiomed.2017.08.013
Received 30 June 2016; Received in revised form 17 July 2017; Accepted 11 August 2017

a cancer, the latter one provides a group of relevant miRNA without any supervision of the user.

In this investigation fuzzy mutual information (FMI) [12] based grouping and selection of miRNAs for cancer (FMIGS) is proposed to provide a group of relevant miRNAs instead of a ranked list of miRNAs. The method is focused on automated selection of relevant miRNAs because the automatic selection facilities to find the optimal group of miRNAs without any users supervision. First, miRNAs are ranked using fuzzy mutual information and then the ranked miRNAs are divided into several groups using SVM. As our ultimate goal is to find the group of miRNAs having expression values with maximum class (normal and cancer) separability, SVM is used to preserve the class difference information between normal and cancer classes during the formation of groups. After the formation of different groups the most relevant one among them is selected. There is also an optional step to remove redundant miRNAs from the selected group for reducing computational burden. In our investigation FMI is used for handling the overlapping expressions of normal and cancer patients of each miRNA in the ranking and in finding the relevant group. During group formation SVM is used to identify those miRNAs which have the same class boundaries between the representatives (see Section 3.2) of normal and cancer expressions. In a part of our investigation we developed a FMI based particle swarm optimization (PSO) method for selecting relevant miRNAs. In this method FMI is used as a fitness function to determine the fitness of the particles and thereby helping to select the optimal set of miRNAs for classification. Therefore, the novelty of our investigation primarily lies in using fuzzy mutual information in the first stage of the method (i.e., ranking miRNAs) where, FMI based ranking helps in starting a group with better initial point compared to that obtained with a crisp method. The novelty also lies in the use of FMI in PSO to determine the fitness of the particles.

The rest of the article is organized as follows. In Section 2 some of the existing investigations, useful for miRNA selection, are discussed. The details of the proposed investigation is described in Section 3. A brief description of the used data sets and the experimental results on those data sets are demonstrated in Section 4. Finally Section 5 concludes this investigation.

2. Related investigations

In this section we discussed some existing investigations, useful for miRNA selection, along with their main utilities and major limitations. In Ref. [7] a method based on fold change is developed to find the commonly deregulated miRNAs in different cancers. However the method is not suitable for identifying the cancer specific miRNAs. In Ref. [13] an unsupervised feature selection method based on feature similarity is developed. The method reduces the number of features by minimizing the redundancy, however it does not consider the relevance of the features. A gene selection method based on SVM [9] is developed by Guyon et al. but the method does not address the problem of redundancy removal. A Laplacian Score based feature selection method is available in Ref. [14], which performs on the basis of the relevance of individual features and does not consider the redundancy between different features. Zhao et al. developed a graph based feature selection technique [15] which unifies supervised and unsupervised feature selection techniques. Like the two methods in Refs. [9] and [14], the method by Zhao et al. also does not address the redundancy removal problem. An efficient method for gene selection is available in Ref. [10] which works on the basis of maximizing relevance and minimizing redundancy of the genes. However the method does not consider the overlapping of different classes. A gene ranking method based on the tradeoff between the methods in Ref. [9] & [10] is available in Ref. [11]. This method also ignores the class overlapping information. In Ref. [16], three methods are developed for gene ranking where the class overlapping is taken into the account. Although the investigation in Ref. [16] considers the class overlapping but degree of overlapping of the individual elements (i.e., expression values) is not taken into the account. A gene ranking method using fuzzy mutual information is developed in Ref. [12], where, a gene is ranked on the basis of the maximum relevance and minimum redundancy. The investigation addresses the problem of overlapping classes by considering degree of overlapping of the gene expressions but the investigation does not automatically select optimum group of miRNAs. A multi-cluster feature selection method for obtaining optimal group of features is available in Ref. [17] but, like some other investigations discussed earlier, the investigation does not address the issue of redundancy removal. In Ref. [8] a method for automatic selection of miRNAs is introduced but the method does not consider the class overlapping during the formation of different groups of the miRNAs. In this investigation we addressed all the mentioned limitations for selecting a group miRNAs for classifying normal and cancerous miRNA expressions and are described in the following section.

3. Proposed method

In the proposed investigation, fuzzy mutual information based grouping and selection of miRNAs for cancer (FMIGS), our first target is to divide the miRNAs into some groups and then determine the relevance of the miRNAs within the groups. Finally the most relevant group would be selected and redundancies will be removed from it. While the relevance can be determined by computing mutual information between the expressions and their corresponding labels, the redundancy can be obtained by determining mutual information between two miRNAs. However the normal and cancer classes may overlap with each other. The degree of overlapping of the class elements (i.e., miRNA expressions) can be determined by using fuzzy set theoretic approach. Therefore, fuzzy mutual information (FMI) is used in this investigation for determining the relevance and redundancy of the miRNAs.

Our method has four steps. In the first step, FMI is used to rank the miRNAs according to their individual relevance to a cancer. In the second step, the top ranked miRNA is used as the initial group point. Then SVM is used to create the first group from the ranked miRNAs. After creation of the first group, the top miRNAs among the remaining ones (if any) are used as the initial group point and another group is formed in a similar way. The process continues until all the miRNAs are included in some group. In the third step, the relevance of each miRNAs in a group is computed by using FMI and the average of them are calculated. The group with the highest average FMI value, among all the groups, is selected as the most relevant group. The fourth and the last step deals with the removal of redundancy from the most relevant group, which is an optional step. To perform this step, FMI values of a miRNA with respect to all other miRNAs (taking two miRNAs at a time) are computed and the average of those values is considered as the redundancy corresponding to that particular miRNA. The technique is applied to all the miRNAs in the most relevant group and a portion of miRNAs with high redundancy is removed as per the user's choice. Note that, the classification accuracy depends on the relevance of the selected miRNAs, but there can be redundant ones which can further be removed to reduce the biochemical and computational cost. The block diagram of FMIGS is presented in Fig. 1 and the detail steps of the method are described as follows.

3.1. Fuzzy mutual information based ranking

The steps for performing the FMI based grouping are provided below.

(i) Consider the normal and cancer expressions of a miRNA as two classes.
(ii) Compute the fuzzy membership values of each expression for both the classes. Here we used the same membership function used in Ref. [8]. When the distance of a miRNA expression decreases with respect to a class center, the membership function provides a non-decreasing membership value of that expression.

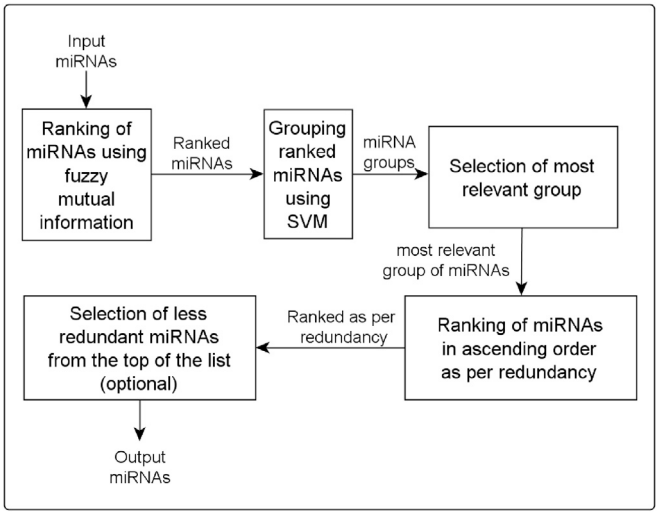


Fig. 1. Selection of miRNAs using proposed FMIGS.

(iii) Compute the FMI of a miRNA using its expression values and the class labels. Note that the membership of a class label, say normal, in normal class is 1 and 0 in cancer class.
(iv) Repeat Steps (i)-(iii) for all the miRNAs in a data set.
(v) Rank all the miRNAs is descending order, according to the value of FMI.

3.2. Grouping

In our investigation we divided the whole set of miRNAs into some groups. As our ultimate goal is to classify normal and cancer patients, the class difference information between the two types of expressions (normal and cancer) must be preserved. Application of any conventional clustering method, where a miRNA is considered as a single vector with all its expressions, does not serve the purpose. In order to utilize the class difference information we used a two class classifier (SVM) based grouping method. The detail steps for grouping the miRNAs are described below.

Consider x_i & y_j are the expressions of ith normal & jth cancer patients of a miRNA, respectively, where $1 \le i \le N$ & $1 \le j \le M$ and N & M correspond to the total number of normal & cancer patients. For a miRNA r_n & r_c are the class representatives corresponding to the normal & cancer classes and σ_n & σ_c are standard deviations of the normal & cancer expressions, respectively.

(i) Compute the representative of each class (i.e., normal and cancer), of a miRNA, as

$$r_n = \frac{1}{\sigma_n \times N} \sum_{i=1}^{N} x_i \ and \tag{1}$$

$$r_c = \frac{1}{\sigma_c \times M} \sum_{j=1}^{M} y_j \tag{2}$$

Note that, we need two patterns (normal and cancer) which can be used for training the SVM. However, for most of the data sets the number of normal and cancer patients are unequal. Training a SVM using two different pattern vectors with unequal dimension (i.e., different numbers of normal & cancer patients) is not possible. Therefore, Eqs. (1) and (2) are introduced which can handle two classes (i.e., normal and cancer) by taking their representatives. The class labels of r_n (Eq. (1)) and r_c (Eq. (2)) are normal & cancer, respectively, as they are the corresponding representatives of normal and cancer classes.

(ii) Repeat Step (i) for all the miRNAs.
(iii) Select the top ranked miRNA from the ranked list (see Step (v), Section 3.1) as the initial group point and train SVM (with linear kernel) using r_n and r_c of the top ranked one.
(iv) 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 initial group point. Here, r_n^k and c_n^k are the corresponding r_n and r_c value of the kth miRNA in a data set, where k is a positive integer.
(v) From the un-grouped miRNAs (if any) select the top ranked one among them and train SVM with the selected miRNA. Repeat Step (iv) to form the next group. The process will continue until all the miRNAs are included in any group.

In our method, first we ranked the miRNAs (see Section 3.1) according to their individual relevance to a cancer. In grouping procedure the top miRNA in the ranked list is used as the initial group point to form the first group. From the second groups the top miRNAs among the ungrouped ones are used as the initial group point. After the grouping process the most relevant group (see Section 3.3) is selected and is further used for classification. As the formation of groups depends on the top ranked miRNA, perfection in ranking ultimately results in the selection of a miRNA group with higher relevance. The normal and cancer expressions of a miRNA can overlap with each other. So, the use of fuzzy set theoretic approach helps in utilizing the class overlapping information during ranking and thereby obtaining highly relevant miRNA group at the end of the selection process.

Two important characteristics of the grouping method are (i) The number of groups are automatically determined by the method, and (ii) Normal & cancer expressions of the miRNAs belonging to the same group are better separable, than those of the other miRNAs, with respect to a particular class boundary. The miRNAs in the same group may not have the same behaviour but they are similar in terms of the class separability of their expressions (normal and cancer) with respect to a particular class boundary.

3.3. Selection of the most relevant group

In this section we will discuss the selection process of the most relevant group.

(i) Determine the relevance of a miRNA by computing FMI between its expression values and class labels. Repeat the process for all the miRNAs in a group.
(ii) Compute the average of all the relevance values corresponding to the miRNAs in a group.
(iii) Repeat Steps (i)-(iii) for all the groups.
(iv) Select the group with the highest average relevance value.

Note that the most relevant group may not contain the best among all the miRNA as per its individual performance, rather the group provides the best performing miRNAs when they are in a group.

3.4. Redundancy removal

The removal of redundant miRNAs is optional as one may want to investigate all the relevant miRNAs for their role in cancer. In other words, redundancy removal step is useful for those data sets where the most relevant group contains numerous miRNAs. In that situation redundancy removal may help the user to further reduce the number of miRNAs. The steps for this process are as follows.

(i) Compute the FMI values of a miRNA with respect to all other miRNAs (considering two miRNAs at a time) in the most relevant group.

(ii) Compute the average and store the value as the redundancy of that miRNA.
(iii) Repeat Steps (i)-(iii) for all miRNAs within a group to determine the redundancy of each miRNA in the selected group.
(iv) Rank the miRNAs in ascending order according to the redundancy value.
(v) Select a percentage of less redundant miRNAs from the top of the list, generated in Step (iv).

3.5. MiRNA selection using fuzzy mutual information based particle swarm optimization

In a part of our investigation we proposed a fuzzy mutual information (FMI) based particle swarm optimization (PSO) method for selecting the relevant miRNAs. PSO is a well known swarm intelligence method which consists a collection of particles where a particle moves around a search space guided by its own best location (pbest) and the best location achieved by any particle into that swarm (gbest). A particle updates its position with a velocity and the fitness of that particle in its new position is checked by a fitness function. The locations pbest and gbest are updated into a new location according to the fitness value (i.e., increase in fitness value) of that particle. In our investigation we used FMI in order to determine the fitness of a particle, where each particle consists of a fixed number of miRNAs. The steps for determining fitness of each particle is provided below.

(i) For each miRNA, belonging to a particle, compute FMI between its expressions and corresponding class labels.
(ii) Compute the average FMI of all the miRNAs within the aforementioned particle. The computed average is considered as the fitness of a particle.
(iii) Repeat the process for all the particles.
(iv) The higher average FMI value corresponding to a particle indicates better fitness of that particle.

4. Experimental results

We used five data sets viz, colorectal [18], lung [19], pancreas [20] cancers, nasopharyngeal carcinoma [21] and melanoma [4] for evaluating our method. A summary of the used data sets is provided in Table 1. It shows the total number of patients (normal and cancer) and miRNAs corresponding to different data sets. For example, nasopharyngeal carcinoma data set consists of 887 miRNAs and 19 normal & 31 cancer patients. All these data sets are publicly available and can be downloaded from the gene expression omnibus (GEO) by using the accession number mentioned in the corresponding articles. During the generation of the data sets authors of the corresponding articles maintained all the ethical issues and uploaded to GEO.

The classification performance achieved by the selected miRNAs by FMIGS is tested using sensitivity, specificity, F score and accuracy. The sensitivity, specificity and F score are determined by the same way as in the investigation by Pal et al. [8] and the accuracy is computed as

$$Accuracy = \frac{Total\ no.\ of\ correct\ classification}{Total\ no.\ of\ patients} \quad (3)$$

In this study we used SVM (with linear kernel) and Naive Bayes classifiers for performance evaluation. We used leave-one-out cross validation technique to evaluate the performance of our methods in Sections 4.1 and 4.2. Comparisons of our method (i.e., FMIGS) with some other techniques using the same cross validation procedure is shown in Section 4.3. Moreover five-fold cross validation technique is also used to compare our method with the other methods in Section 4.4.

4.1. Performance evaluation

The performance of the miRNAs in most relevant group is discussed in this Section from various aspects. First we checked whether the most relevant group contains the top miRNA obtained by the ranking method (Section 3.1). Using the ranking method we obtained hsa-miR-422a, hsa-let-7e, hsa-miR-130b, hsa-miR-548q and hsa-let-7d corresponding to colorectal, lung, pancreas cancers, nasopharyngeal carcinoma and melanoma as the top ranked miRNAs. It is observed none of these top ranked miRNAs belong to the most relevant group (See Table 2) of the corresponding data sets. This proves the importance or necessity of creating & evaluating different groups and selecting the most relevant one among them.

Table 3 shows the F scores achieved by the miRNAs selected by the proposed method as compared to those obtained by all the miRNAs, corresponding to different data sets. From Table 3 it is observed that the classification performance in terms of F score is considerably improved for the selected miRNAs using FMIGS. For example, using SVM classifier,

Table 3
Comparison of F scores achieved by all the miRNAs and the selected miRNAs using SVM and Naive Bayes classifiers.

Cancer Type	Total Samples/ Patients	Total miRNAs (no. and F score)			Selected miRNAs (no. and F score)		
		No.	SVM	Naive Bayes	No.	SVM	Naive Bayes
Colorectal	66	352	0.61	0.66	2	0.83	0.90
Lung	36	866	0.52	0.53	3	0.79	0.80
Pancreas	158	847	0.62	0.68	3	0.85	0.92
Nasopharyngeal	50	887	0.31	0.31	2	0.65	0.67
Melanoma	57	866	0.61	0.62	2	0.82	0.85

Table 1
Summary of the data sets.

Cancer Type	Total No. of miRNAs	No. of Normal Patients	No. of Cancer Patients
Colorectal	352	8	58
Lung	866	19	17
Pancreas	847	22	136
Nasopharyngeal	887	19	31
Melanoma	864	22	35

Table 2
Selected miRNAs for different data sets.

	Cancer Type				
	Colorectal	Lung	Pancreas	Nasopharyngeal	Melanoma
miRNA	hsa-miR-224	hsa-let-7f hsa-miR-25	hsa-miR-23a hsa-miR-30a	hsa-miR-181a	hsa-miR-30d
	hsa-miR-15a	hsa-let-7c	hsa-miR-199b-3p	hsa-miR-92a	hsa-miR-601

Table 4
Classification performance of the selected miRNAs for various data sets using SVM and Naive Bayes classifiers.

Classifier	Performance	Colorectal	Lung	Pancreas	Nasopharyngeal	Melanoma
SVM	Sensitivity	0.77	0.79	0.79	0.58	0.81
	Specificity	0.90	0.79	0.92	0.74	0.82
	Accuracy	0.83	0.79	0.86	0.66	0.82
Naive Bayes	Sensitivity	0.90	0.82	0.96	0.61	0.87
	Specificity	0.90	0.78	0.89	0.73	0.84
	Accuracy	0.90	0.81	0.93	0.67	0.86

Table 5
F score values before and after the redundancy removal from a selected group of miRNAs.

Cancer Type	Group No.	Before Redundancy Removal			After Redundancy Removal		
		No. of miRNAs	F score		No. of miRNAs	F score	
			SVM	Naive Bayes		SVM	Naive Bayes
Colorectal	3	36	0.62	0.74	29	0.61	0.74
Lung	3	7	0.70	0.65	6	0.69	0.64
Pancreas	2	10	0.80	0.91	8	0.78	0.90
Nasopharyngeal	2	26	0.57	0.65	21	0.57	0.64
Melanoma	3	8	0.74	0.75	6	0.74	0.74

the *F* score increased from 0.62 to 0.85 after selection of 3 miRNAs from 847 miRNAs in pancreas cancer data set. Using Naive Bayes classifier the F score increased from 0.68 to 0.92 for the same 3 miRNAs. The performance of FMIGS in terms of sensitivity, specificity and accuracy are reported in Table 4. From the table it can be observed that using SVM classifier the corresponding values of sensitivity, specificity and accuracy vary from 0.58 to 0.81, 0.74 to 0.92 and 0.66 to 0.86, depending on different data sets. Using Naive Bayes classifier these measures vary from 0.61 to 0.96, 0.73 to 0.90 and 0.67 to 0.93, for various data sets.

As the number of miRNAs are very few (see Table 3) in the most relevant group for the data sets used in this investigation, the performance of the redundancy removal technique is tested on some different groups having larger number of miRNAs. The results after the removal of redundant miRNAs are reported in Table 5. It is seen that after removing 20% of the redundant miRNAs from a set of selected miRNAs the corresponding *F* scores are almost the same as compared to those obtained using all the selected miRNAs. For example, the *F* score decreased from 0.65 to 0.64 (with Naive Bayes classifier) when 20% of redundant miRNAs are removed for nasopharyngeal carcinoma data set. Similar results are observed for other cases also.

4.2. Results using FMI based PSO

Here we evaluate the performance achieved in terms of sensitivity, specificity, accuracy and F Score by the selected miRNAs using FMI based PSO and the results are reported in Table 6. The number of selected miRNAs in FMI based PSO is kept the same with those of the selected miRNAs in FMIGS. For example, in pancreas cancer data three miRNAs are selected by using FMI based PSO method as the same number of miRNAs are automatically selected by the FMIGS. It is observed that using SVM classifier the sensitivity, specificity, *F* score and accuracy vary from 0.62 to 0.92, 0.78 to 0.84, 0.70 to 0.86 and 0.71 to 0.89 corresponding to various data sets. Similar results are also observed by Naive Bayes classifier.

4.3. Comparison with other methods

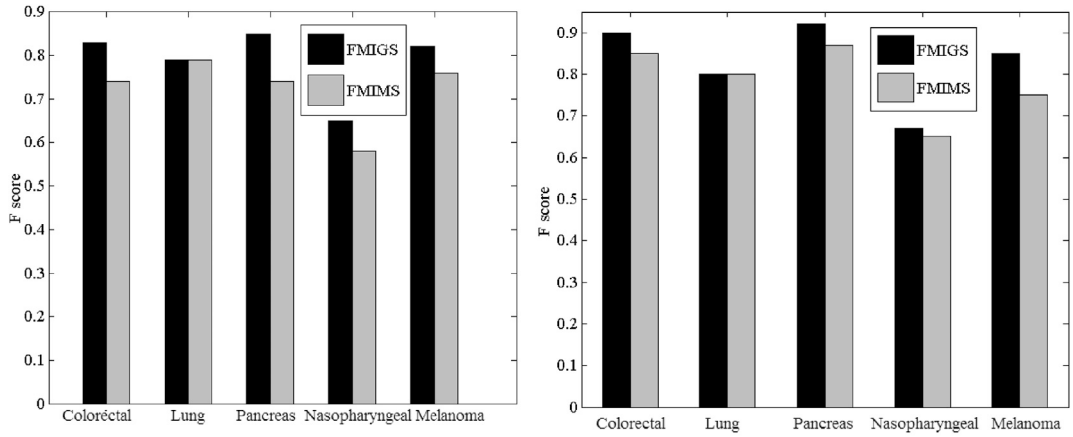
In this section the proposed FMIGS is compared with some other methods to verify its effectiveness. The methods used for comparisons are identifying relevant group of miRNAs in cancer using fuzzy mutual information (FMIMS) [8], SVM classifier based recursive feature elimination technique (SVMRFE) [9], feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy (MRMR) [10], SVMFRE with MRMR [9] and a ranking method by using FMI (FRSIM) [12].

While in FMIMS [8] miRNAs were grouped by an SVM based algorithm where the grouping method was designed by using some crisp measures, in the proposed FMIGS fuzzy mutual information is used to construct the groups of miRNAs in order to consider the overlapping between the expressions of two different patients (normal and cancer). First comparison is made between FMIGS and FMIMS as these two algorithms select the most relevant group of miRNAs automatically. Then the comparison is made with the other ranking algorithms. The results regarding the former comparison using SVM and Naive Bayes classifiers are presented in Fig. 2(a) and (b), respectively, and the latter ones are shown in Fig. 3(a) and (b) corresponding to the same classifiers.

From Fig. 2(a) and (b) it is observed that the FMIGS outperforms FMIMS for all the data sets except for the lung cancer data set, where the *F* scores achieved by the selected miRNAs for both the algorithms are the same. It is also observed, while for various data sets the *F* score corresponding to the proposed method varies from 0.65 to 0.85 it varies from 0.58 to 0.79, for FMIMS, using SVM as the classifier. Using Naive base classifier FMIGS achieves *F* scores from 0.67 to 0.92 and FMIMS provides *F* score values from 0.65 to 0.87 corresponding to different data sets.

Table 6
Classification performance of the selected miRNAs using FMI based PSO.

Classifier	Performance	Colorectal	Lung	Pancreas	Nasopharyngeal	Melanoma
SVM	Sensitivity	0.90	0.76	0.92	0.62	0.90
	Specificity	0.81	0.82	0.84	0.78	0.84
	F score	0.86	0.79	0.88	0.70	0.86
	Accuracy	0.86	0.79	0.89	0.71	0.87
Naive Bayes	Sensitivity	0.97	0.86	0.93	0.61	0.95
	Specificity	0.75	0.77	0.93	0.86	0.79
	F score	0.85	0.81	93.48	0.71	87.62
	Accuracy	0.86	0.79	0.89	0.71	0.87

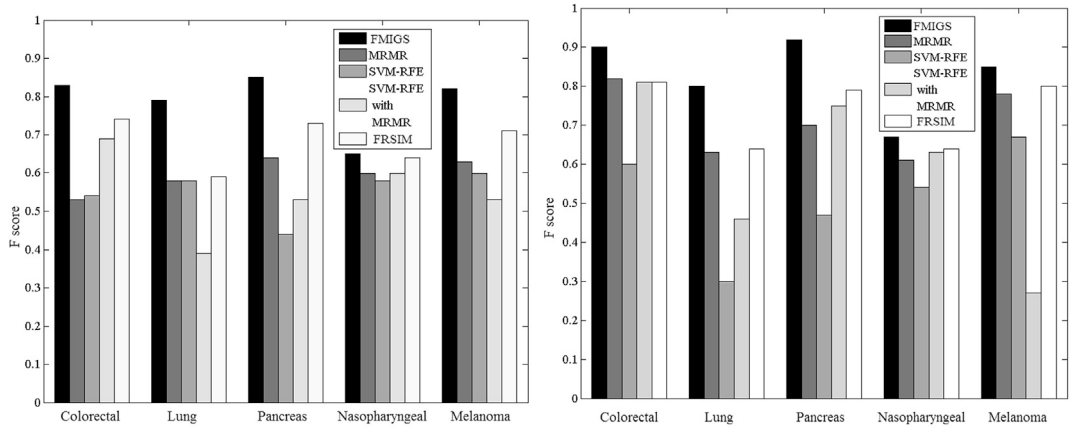


(a) Comparison using SVM classifier (b) Comparison using Naive Bayes classifier

Fig. 2. Comparison of classification performances in terms of *F* score using the selected miRNAs by FMIGS and FMIMS corresponding to different data sets. Leave-one-out cross validation is use to evaluate the performances.

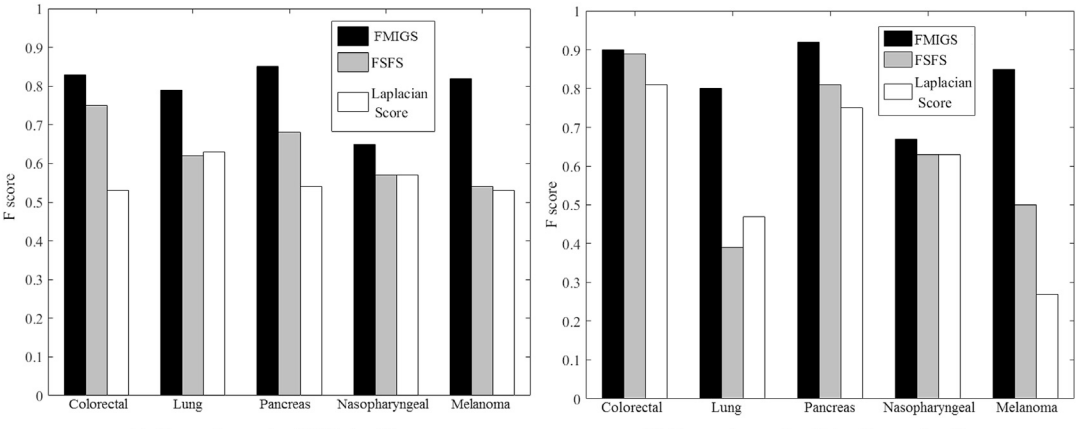
From Fig. 3(a) and (b) it is observed that FMIGS performs the best in terms of *F* score for all the data sets and classifiers as compared to those of the other ranking methods. For example, while using colorectal cancer data set FMIGS achieves the best *F* score of 0.83 with SVM classifier, the second best performance is achieved by FRSIM having *F* score 0.74. Similar results are observed for other data sets using both the classifiers.

Apart from the comparisons with some miRNA/gene selection methods (in Figs. 2 and 3) we also compared our method with two classical feature selection methods viz., unsupervised feature selection using feature similarity (FSFS) [13] and Laplacian score for feature selection [14]. The comparison is made in terms of *F* score and the results are shown in Fig. 4(a) and (b). It is observed that the FMIGS outperforms both the methods for all cases. For example, while the *F* scores vary from 0.65 to 0.85 for FMIGS, the same score corresponding to FSFS and



(a) Comparison using SVM classifier (b) Comparison using Naive Bayes classifier

Fig. 3. Comparison of classification performances in terms of *F* score using the selected miRNAs by FMIGS, MRMR, SVMRFE, SVMRFE with MRMR, FRSIM corresponding to different data sets. Leave-one-out cross validation use to evaluate the performances.



(a) Comparison using SVM classifier (b) Comparison using Naive Bayes classifier

Fig. 4. Comparison with FSFS and Laplacian score based feature selection using leave-one-out cross validation.

Laplacian score based feature selection methods vary from 0.54 to 0.75 and 0.53 to 0.63 using SVM as classifier.

In Figs. 3 and 4 the same number of miRNAs, as selected by FMIGS, are used for determining the performance of the other methods. For example, in pancreas cancer data top three miRNAs are used for all other ranking methods as the same number of miRNAs are automatically selected by the proposed FMIGS. The number of selected miRNAs are kept the same for all the methods in the mentioned figures due to the sake of fair comparison. Note that for any biological studies where more miRNAs are required than that of the miRNAs in most relevant group, one need to choose the next relevant group/s as per the necessity.

4.4. Other experiments

In a part of our investigation we tested the performance of the proposed FMIGS by five-fold cross validation technique using SVM classifier and the corresponding results are reported in Table 7. It is observed from the table that sensitivity, specificity, *F* score and accuracy of our method range between 0.71 & 0.82, 0.68 & 0.88, 0.72 & 0.83, and 0.72 & 0.83, respectively, depending on various data sets.

We also compared the performance of FMIGS, with the methods used for comparison in Section 4.3, by five-fold cross validation technique using SVM. The results related to these comparisons are reported in Fig. 5(a), (b) and 6 where the first two figures show the comparison of the proposed method with some miRNA/gene selection methods, the last one shows the comparison with some classical feature selection methods. In Figs. 5(b) and 6, the numbers of selected miRNAs by the methods, other than the proposed FMIGS, are kept the same with those of FMIGS. This is done for fair comparison purpose as those other methods do not select the number of optimum miRNAs automatically. It is observed that our method performs the best in all cases using five-fold cross validation technique. While the *F* score varies from 0.72 to 0.83 for the proposed FMIGS, the score varies between 0.57 & 0.73 for FMIMS, 0.41 & 0.63 for MRMR, 0.32 & 0.62 for SVM-RFE, 0.51 & 0.59 for SVM-RFE with MRMR, 0.56 & 0.77 for FRSIM, 0.42 & 0.66 for FSFS, and 0.41 & 0.62 for Laplacian score based feature selection.

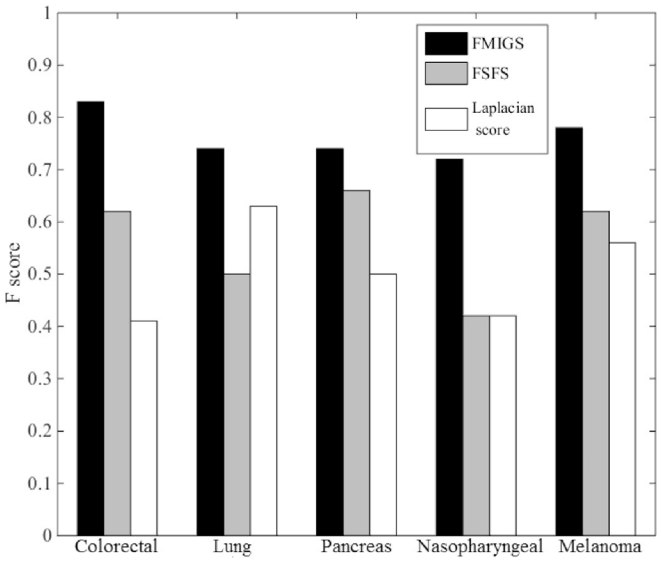


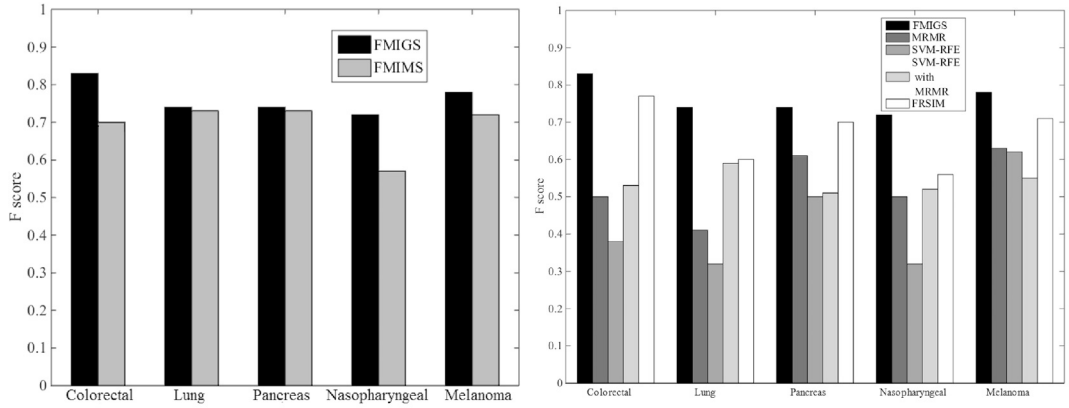
Fig. 6. Comparison with FSFS and Laplacian score based feature selection using five-fold cross validation technique.

4.5. Biological relevance

The involvement of the selected miRNAs by FMIGS (see Table 2) in various cancers are verified using publicly available pathway analysis tools DIANA [22] and Starbase [23]. These tools can identify the target genes of a miRNA causing the development of cancers. Table 8 shows the miRNAs selected as the important ones for different cancers with their target genes and the corresponding p-values for targeting those genes. From the table it can be observed that 11 out of 12 selected miRNAs, are found to be responsible for various cancers according to the aforesaid pathway analysis tools. For example, in pancreas cancer hsa-miR-23a interacts with 14 genes and controls the development of this disease. Similar interactions can be observed for other data sets also. The miRNA hsa-miR-199b-3p which is not identified as responsible miRNA for

Table 7
Classification performance using five-fold cross validation.

Classifier	Performance	Colorectal	Lung	Pancreas	Nasopharyngeal	Melanoma
SVM	Sensitivity	0.79	0.76	0.82	0.71	0.81
	Specificity	0.88	0.71	0.68	0.74	0.74
	F score	0.83	0.74	0.74	0.72	0.78
	Accuracy	0.83	0.74	0.74	0.72	0.78



(a) Comparison between FMIGS and FMIMS

(b) Comparison between FMIGS, MRMR, SVMRFE, SVMRFE with MRMR and FRSIM

Fig. 5. Comparison of classification performances in terms of *F* score using the selected miRNAs of various miRNA/gene selection algorithms using five-fold cross validation technique.

Table 8
Pathway Analysis of Selected miRNAs using DIANA/Starbase.

Cancer Type	miRNA	Target genes	p value related to gene targeting
Colorectal	hsa-miR-15a	BRAF, SMAD2, SMAD3 BCL2, BIRC5, APPL1 KRAS, TP53, AKT2 MAPK9, JUN, CCND1 CTNNB1, MSH6, AXIN2 MAPK8, AKT1, MYC MSH2, RAC1, AKT3 PIK3CA, MAP2K1 CCND1, RALGDS	0.0003
	hsa-miR-224	ARAF, CCND1, GSK3B MSH2, PIK3CD PIK3R3, RAC1 SMAD4, TGFB3 TGFBR1, MSH6	0.003
Lung	hsa-let-7f	CCND1	0.001
	hsa-miR-25	BRAF, CDK4, NRAS PIK3CB, KRAS, CDK6 TP53, CCND1, E2F3, SOS1	0.02
	hsa-let-7c	AKT3, APAF1, BCL2L1 CCND1, CDK4, CDK6 CDKN1B, CHUK, COL4A1 COL4A2, COL4A5, COL4A6 E2F2, LAMA1, LAMC1 MYC, PTK2, RB1 TP53, TRAF1, XIAP	0.0005
Pancreas	hsa-miR-23a	AUH, BNIP3L, CCNG1 CCNT2, LRIG1, PDE4B PLAGL1, PTEN, SATB1 SLC1A1, SPOCK1, TIA1 TRIB1, ZNF292	0.009
	hsa-miR-30a	TGFBR1, NFKB1, CDK4 SMAD2, BRCA2, EGFR CDK6, TP53, IKBKB PIK3CD, CCND1, MAPK8 AKT1, RB1, RAC1 BCL2L1, AKT3, PIK3CA STAT1, VEGFA, MAPK1 TGFBR2, JAK1	0.0003
Nasopharyngeal	hsa-miR-181a	ADAMTS1, ADAMTS5, BCLAF1 CPEB4, DDX3Y, DYRK2,ETV6 HSP90B1, ILF3, LIN28B LRP12, NR1D2, PLAG1 TBL1XR1, TNFRSF11B, TRIM2	1.40×10^{-05}
	hsa-miR-92a	APPL1, ATRX, CPEB4 DDX3Y, DYRK2, FBXW7 FNDC3B, FOXP2, NFIA SESN3, SLC25A36 UBE2W, WASL, TBL1XR1	0.0003
Melanoma	hsa-miR-30d	APP, JAK1 CDKN1A	0.020
	hsa-miR-601	ADAM9, ARID4B, BAZ2B BNIP3L, CEP350, CHD1 CHMP2B, CUL2, EEA1 FAP, GALNT1, GOLGA4 ITGA6, MMD, MYH10 MYO5A, NFIB, NRIP1 PPARGC1A, PTPN13, RAI14 SLC38A2, SOCS3, SON STAG2, USP48, XPO1 ZCCHC2, ZNF148, ZNF644	1.02×10^{-06}

pancreas cancer by DIANA or Starbase is found as responsible for the same cancer by Ma et al. [24] which corroborates our investigation.

5. Conclusion

A fuzzy mutual information (FMI) based grouping and miRNA selection method (FMIGS) is developed in this investigation. At first miRNAs are ranked using FMI. Then several groups of miRNAs are created by using a SVM framework. Then the selection of most relevant group among all groups is selected. Finally the removal of redundant miRNAs within it is performed. The redundancy removal technique is optional in our method to facilitate the user to keep all the miRNAs or to remove a percentage from the most relevant group of miRNAs as per the necessity. A FMI based particle swarm optimization (PSO) is also proposed where a FMI based fitness function is used to determine the fitness of the particles.

The classification performance in terms of sensitivity, specificity, F score and accuracy of the selected miRNAs using FMIGS vary from 0.58 to 0.81, 0.74 to 0.92, 0.65 to 0.85 and 0.66 to 0.86, respectively using SVM classifier with the leave-one out cross validation technique. Using Naive Bayes classifier and the same cross validation technique these four measures for various data sets vary between 0.61 & 0.96, 0.73 & 0.90, 0.67 & 0.92 and 0.67 & 0.93. We also evaluated our method by five-fold cross validation technique and obtained similar results in terms of sensitivity $(0.71 - 0.82)$, specificity $(0.68 - 0.88)$, F score $(0.72 - 0.83)$ and accuracy $(0.72 - 0.83)$. The results regarding the redundancy removal technique show the efficacy in terms of F scores when 20% of the redundant miRNAs are removed from the selected miRNAs. The automatic identification of the most relevant group of miRNAs helps the user to easily focus on a subset rather than all miRNAs of a patient. The comparison results with some other methods show the superiority of FMIGS to those methods using both the cross validation methods used in this investigation. Moreover, all the miRNAs selected by our method are found to be relevant according to pathway analysis tools or biological investigation. Using FMI based PSO method the achieved sensitivity, specificity, F score and accuracy vary from 0.62 to 0.92, 0.78 to 0.84, 0.70 to 0.86 and 0.71 to 0.89 corresponding to different data sets using SVM classifier. Similar results are also obtained by Naive Bayes classifier. These results along with the biological relevance of the selected miRNAs show the importance of the investigation.

Acknowledgements

S. K. Pal acknowledges the J. C. Bose fellowship and the Raja Ramanna fellowship of the Govt. of India.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compbiomed.2017.08.013.

References

[1] S.S. Ray, J.K. Pal, S.K. Pal, Computational approaches for identifying cancer miRNA expressions, Gene Expr. 15 (nos. 5–6) (2013) 243–253.
[2] J.K. Pal, S.S. Ray, S.K. Pal, A weighted threshold for detection of cancerous miRNA expressions, Fundam. Inf. 127 (nos. 1–4,) (2013) 289–305.
[3] E. Guruceaga, V. Segura, Functional interpretation of microRNA-mRNA association in biological systems using R, Comput. Biol. Med. 44 (2014) 124–131.
[4] P. Leidinger, A. Keller, A. Borries, J. Reichrath, K. Rass, S.U. Jager, H.P. Lenhof, E. Meese, High-throughput miRNA profiling of human melanoma blood samples, BMC Cancer 10 (no. 1) (2010) 1–11.
[5] M.G. Schrauder, R. Strick, R.D. Schulz-Wendtland, P.L. Strissel, L. Kahmann, C.R. Loehberg, M.P. Lux, S.M. Jud, A.H.A. Hartmann, C.M. Bayer, M.R. Bani, S. Richter, B.R. Adamietz, E. Wenkel, C. Rauh, M.W. Beckmann, P.A. Fasching, Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection, PLoS One 7 (no. 1) (2012), e29770.
[6] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H. R.Horvitz, T.R. Golub, MicroRNA expression profiles classify human cancers, Nature 435 (no. 9) (2005) 834–838.
[7] R. Navon, H. Wang, I. Steinfeld, A. Tsalenko, A. Ben-Dor, Z. Yakhini, Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types, PLoS One 4 (no. 11) (2009), e8003.
[8] J.K. Pal, S.S. Ray, S.K. Pal, Identifying relevant group of miRNAs in cancer using fuzzy mutual information, Med. Biol. Eng. Comput. 54 (no. 4) (2016) 701–710.
[9] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (no. 1–3) (2002) 389–422.

[10] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy, IEEE Trans. Pattern Analy. Mach. Intell. 27 (no. 8) (2005) 1226–1238.
[11] P.A. Mundra, J.C. Rajapakse, SVM-RFE with MRMR filter for gene selection, IEEE Trans. Nanobiosci. 9 (no. 1) (2010) 31–37.
[12] P. Maji, S.K. Pal, Fuzzy-rough sets for information measures and selection of relevant genes from microarray data, IEEE Trans. Syst. Man Cybernetics-Part B Cybern. 40 (no. 3) (2010) 741–752.
[13] P. Mitra, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Analy. Mach. Intell. 24 (no. 3) (2012) 301–312.
[14] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, 2005, pp. 507–514.
[15] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 1151–1157.
[16] A. Sharma, S. Imoto, S. Miyano, A between-class overlapping filter-based method for transcriptome data analysis, J. Bioinforma. Comput. Biol. 10 (no. 5) (2012), 1 250 010:1–1 250 010:20.
[17] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 333–342.
[18] G.M. Arndt, L. Dossey, L.M. Cullen, A. Lai, R. Druker, M. Eisbacher, C. Zhang, N. Tran, H. Fan, K. Retzlaff, A. Bittner, M. Raponi, Characterization of global microRNA expression reveals oncogenic potential of mir-145 in metastatic colorectal cancer, BMC Cancer 9 (no. 1) (2009) 1–17.
[19] A. Keller, P. Leidinger, A. Borries, A. Wendschlag, F. Wucherpfennig, M. Scheffler, H. Huwer, H.-P. Lenhof, E. Meese, miRNAs in lung cancer - studying complex fingerprints in patient's blood cells by microarray experiments, BMC Cancer 9 (no. 1) (2009) 1–10.
[20] A.S. Bauer, A. Keller, E. Costello, W. Greenhalf, M. Bier, A. Borries, M. Beier, J. Neoptolemos, M. Bchler, J. Werner, N. Giese, J.D. Hoheisel, Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue, PLoS One 7 (no. 4) (2012), e34151.
[21] X.-H. Zheng, C. Cui, H.-L. Ruan, W.-Q. Xue, S.-D. Zhang, Y.-Z. Hu, X.-X. Zhou, W.-H. Jia, Plasma microRNA profiling in nasopharyngeal carcinoma patients reveals mir-548q and mir-483-5p as potential biomarkers, Chin. J. Cancer 33 (no. 7) (2014) 330–338.
[22] I.S. Vlachos, K. Zagganas, M. D.Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, A.G. Hatzigeorgiou, DIANA-miRPath v3.0: Deciphering microRNA Function with Experimental Support, vol. 43, Nucleic Acids Research, (Web Server Issue), 2015, pp. W460–W466.
[23] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, J.-H. Yang, Starbase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-scale Clip-seq data, vol. 42, Nucleic Acids Research, (Database Issue), 2014, pp. D92–D97.
[24] M.-Z. Ma, X. Kong, M.-Z. Weng, K. Cheng, W. Gong, Z.-W. Quan, C.-H. Peng, Candidate microRNA biomarkers of pancreatic ductal adenocarcinoma: meta-analysis, experimental validation and clinical significance, J. Exp. Clin. Cancer Res. 32 (no. 1) (2013) 71–84.