Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN
0975-6299

AN MRMR WITH MEAN SCORE FEATURE SELECTION FOR OVARIAN CANCER CLASSIFICATION USING JOINT ANALYSIS

M.ANIDHA[*1] AND DR K.PREMALATHA[2]

*[*1]Research Scholar, Anna University, Chennai, Tamilnadu, India*
[2]Professor, Department of CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India

ABSTRACT

Cancer Classification from microarray expression profiles is a challenging task due to its high dimensionality in the field of biomedicine and bioinformatics. The microarray data experiment contains large number of features and small number of samples, therefore feature selection is an essential task in cancer classification. In this paper, a novel feature selection technique is proposed based on minimum Redundancy Maximum Relevance (mRMR) in which the mean score is introduced to improve the relevance between features. The feature selection is employed in gene expression data and miRNA expression data using joint analysis in ovarian cancer dataset. Joint analysis gives 100% accuracy for ovarian cancer using the classifiers Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The identified signature of miRNAs and genes are useful for finding the stages of ovarian cancer and therapeutic leads to the cancer patients.

KEYWORDS: *Cancer Classification; Feature Selection; mRMR; ovarian cancer; miRNA; mRNA.*

M.ANIDHA*

Research Scholar, Anna University, Chennai, Tamilnadu, India.

***Corresponding Author**

Received on: 01-02-2017
Revised and Accepted on: 22-03-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.2.b495-504

INTRODUCTION

Cancer is a complex disease where molecular mechanism remains elusive[1]. A systematic approach is needed to integrate diverse biological information for the prognosis and therapy risk assessment using mechanistic approach. These assessments understand gene interactions in pathways and networks and functional attributes to unravel the biological behavior of tumors[1].Ovarian cancer is known as a particularly lethal cancer because symptoms can be vague and many women are not diagnosed until the metastasis. Ovarian tumors are a puzzling group of neoplasms that do not fall neatly into benign or malignant categories. Their behaviour is enigmatic, pathogenesis is unclear, and their diagnosis, clinical management prognosis are controversial, especially for borderline epithelial tumors[2]. Ovarian cancer is notoriously difficult to diagnose and treat, resulting in high mortality rate. Researchers at University of California, San Diego School of Medicine and Moores Cancer Center (San Diego) have now identified six mRNA isoforms (bits of genetic material) produced by ovarian cancer cells but not normal cells. They could be possibly used to diagnose early-stage ovarian cancer[3]. In typical, miRNAs are involved in crucial biological process that includes development, differentiation, apoptosis and proliferation[4]. The miRNAs also have possible implications for improving cancer diagnosis. For example, miR-200 family, let-7 family, miR-21 and miR-214 are useful in diagnostic tests to help detect ovarian cancer at an early stage[4].The recently discovered microRNAs (miRNAs) constitute a novel regulatory layer of gene expression and have been implicated in the etiology of various kinds of human cancers. The miRNAs are small (~22bp) endogenous non-coding RNAs and are frequently deregulated in cancer[4]. In general, the expressions of miRNAs in malignant cells is significantly different from that of normal counterpart cells and facilitate the stratification of cancer and the identification of the tissue of origin for poorly differentiated tumors[5]. MicroRNAs (miRNA) can act as oncogenes or tumor suppressors and modulate the expression of approximately one third of all human genes. Several studies have reported that specific miRNA expression signatures can be used as predictors of esophageal cancer diagnosis and prognosis[6]. Moreover, miRNA processing genes have also been associated with the development and survival of multiple cancers, including esophageal cancer[6].

Integrated analysis of miRNA and Gene expression microarray data has proved a milestone in the diagnosis and prognosis of cancer. Ibrahim et al. stated that self-learning and co-training which are semi-supervised machine learning techniques used to enhance the classification performance. Kim et al.,(2013) revealed that mRNA, miRNA and integrated analysis of both mRNA and miRNA using three different classification methods such as RF and SVM-based machine learning algorithm, a modified regression analysis method involving Cox regression, based on initial feature selection (FSCR_REG) a SVM-based machine learning algorithm using gene expression levels multiplied by Cox coefficient (FSCR_SVM) with Fisher Score as feature selection criterion and are used to predict the cancer survival subtype for ovarian cancer data. Ogul implemented the feature selection for miRNA, mRNA and integrated set of both miRNA and mRNA by using Support Vector Machines (SVM) attribute selection, Information Gain based attribute selection, Gain Ratio based attribute selection, chi-squared test-based feature selection and CFS(Correlation based Feature Selection) subset attribute selection. The classification is done by using C4.5 Decision Tree (DT), Artificial Neural Networks (ANN), SVM, Naïve Bayes Multinomial (NBM) classifier, and K-Nearest Neighbours (KNN) methods.

Peng et al., stated that cancer classifications with mRNA profiles are superior than miRNA data profiles. They have used SVM based nRFE (Recursive Feature Elimination) technique for classifying Poorly Differentiated Tumors (PDT). Panagiotis A implemented supervised principal component survival analysis to identify prognostic models which is a reproducible predictor of survival by using 256 advanced stage ovarian cancer samples collected from different sources. Models were independently validated in a 61-patient cohort using a custom array gene chip and a publicly available 229-array dataset. Molecular correspondence of high- and low-risk outcome groups between training and validation datasets was demonstrated using Subclass Mapping. Ding et al., implemented mRMR as a feature selection technique and is classified with Naïve Bayes, Linear discriminant analysis, Logistic regression and SVM on 5 gene expression data sets: NCI, Lymphoma, Lung, Leukemia and Colon cancer datasets. Mandal et al. designed an improved mRMR using the mutual information between a feature and class labels defines the relevance of that feature. The mutual information among different features define the correlation i.e., the redundancy among those features. This technique is tested on benchmark gene expression datasets. Unler et al., presented a hybrid filter-wrapper feature subset selection algorithm in which the filter model is based on mutual information which maximizes the relevance and the modified discrete PSO algorithm to minimize the redundancy and is classified with SVM technique.

METHODS AND MATERIALS

Data Sets

The data sets used in this study are publically available at The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov)as a reliable source of benchmark cancer data sets. This work considers levels 3 (Normalized, Segmented and interpreted expression calls) of both gene expression data and miRNA expression data. The gene expression data set is from AgilentG4502A consisting of 46 samples and 17,814 features and the miRNA expression dataset is from Agilent miRNA_8×15K consisting of 38 samples and 799 features. The dataset has the samples with the following categories: Normal, Ovarian Triplet, Stage I, Prior Malignancy, Prior malignancy of Breast cancer and Subject Positive for Neo-Adjuvant Therapy.

Methods

Introduction to mRMR Feature Selection

The proposed feature selection is based on minimum Redundancy Maximum Relevance (mRMR). In order to combine two effective features which are highly correlated the issue of redundancy may arise. This will lead to inefficient set of features selected for classification since the feature set will consist of a fewer independent or representative features. Secondly, the features selected are not maximally representative of the original space covered by the entire dataset and will represent narrow regions of the relevant space. It is strongly recommended that the dissimilarity between features by maximizing their mutual Euclidean distances, or their pairwise correlations are minimized[12]. Table I shows the criterion function of mRMR optimization conditions.

Table I
Criterion Functions of mRMR optimization conditions

Type	Technique	Formula
Discrete (variables or attributes or Features)	MID-Mutual Information Difference	$\max_{i\in\Omega_S}\left[I(i,h)-\frac{1}{\lvert S\rvert}\sum_{j\in S}I(i,j)\right]$
	MIQ-Mutual Information Quotient	$\max_{i\in\Omega_S}\left\{I(i,h)/\left[\frac{1}{\lvert S\rvert}\sum_{j\in S}I(i,j)\right]\right\}$
Continuous (variables or attributes or Features)	FID-F-test Correlation Difference	$\max_{i\in\Omega_S}\left[F(i,h)-\frac{1}{\lvert S\rvert}\sum_{j\in S}\lvert c(i,j)\rvert\right]$
	FIQ-F-test Correlation Quotient	$\max_{i\in\Omega_S}\left\{F(i,h)/\left[\frac{1}{\lvert S\rvert}\sum_{j\in S}\lvert c(i,j)\rvert\right]\right\}$

A feature is an individual measurable heuristic property of a fact being observed. The feature selection identifies subsets of attributes that are relevant to the parameters used and is normally called as maximum relevance. These subsets contain information which is relevant but redundant and the mRMR address as this problem by removing those redundant subsets. Features can be selected in many different ways. The maximum relevance identifies the subset of features with the highest relevance to the classification variable based on mutual information without considering relationships among the features. The selected features are mutually far away from each other having a high correlation to the classification variable. The minimum Redundancy Maximum Relevance (mRMR) selection is more powerful than the maximum relevance selection. Gene expression dataset is a continuous dataset which contains expression levels genes for samples/conditions. Let $g = \{g1, g2, \dots, gN\}$ with targeted classes $h = \{h1, h2, \dots, hK\}$. For continuous variable, the F-statistic between the genes and the classification variable h is chosen as maximum relevance between the genes and class variable.
The formula for the F-statistic is

$$F = \frac{between - class\ variability}{withih - class\ variabilty}$$

The between-class variability is defined as

$$bc = \frac{\sum_{i=1}^{K} n_i(\overline{gi} - \overline{g})^2}{K-1}$$

where $\overline{gi}$ denotes the sample mean in the i^{th} class, n_i is the number of genes in the i^{th}group, $\overline{g}\,\overline{g}$ denotes the overall mean of the genes and *K* denotes the number of classes.

The within-class variability is defined as

$$wc = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(gij - \overline{gi})^2}{N-K}$$

where *gij* is the j^{th} dimension in the i^{th} gene expression out of *K* groups and *N* is the overall gene size. The F-statistics is shown in equation (1)

$$F(h, gi) = \frac{bc}{wc} \qquad (1)$$

The correlation coefficient indicates the strength and direction of a relationship between two random variables. In general statistical usage, correlation refers to the departure of two random variables from independence. Equation (2) shows the calculation of the correlation coefficient between two variables x and y.

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}}} \qquad (2)$$

There are totally n observations. In gene expression dataset, the variables the genes *gi* and *gj replace* the variables *x* and *y* respectively. The two variables have strong dependency when their correlation coefficient value is close to 1 or -1. When the value is 0, it means that the two variables are not related to each other. In the proposed work, strong dependency is looked for either it is positive or negative. Therefore, in the measurement procedure, the absolute value of the correlation coefficient | r | is used. The minimum redundancy is calculated as

$$R(i,S) = \frac{1}{|S|}\sum_{j\in S} r_{ij} \quad (3)$$

In mRMR the first feature is selected according to F-statistics i.e. the feature with the highest $F(h,gi)$. The rest features are selected in an incremental in which the way earlier selected features remain in the feature set. Suppose m features are already selected for the set *S*, we need to select additional features from the set *G* = *g* − S (i.e. all genes except those already selected). The two conditions are optimized as given below:

$$max\, F(h,gi) \quad (4)$$

$$min\, R(i,S) \quad (5)$$

The condition in equation (4) is equivalent to the maximum relevance condition equation (1) and equation (5) is an approximation of the minimum redundancy condition of equation (2). The selection criterion of a new feature is based on mutual information difference criterion as given below in equation (6):

$$max_{i\epsilon G}\left[F(h,gi) - \frac{1}{|S|}\sum_{j\epsilon S} R(i,j)\right] \quad (6)$$

Mutual Information

The Mutual Information (MI) of two random variables is a measure of the mutual dependence between the two variables. It quantifies the amount of information obtained from one random variable, through the other random variable. The concept of mutual information is linked to the entropy of a random variablethat defines the amount of information held in a random variable. MI determines how similar the joint distribution $p(X,Y)$ is to the products of factored marginal distribution $p(x)p(y)$. The mutual information of two discrete random variables *X* and *Y* can be defined as:

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (7)$$

where $p(x,y)$ is the joint probability distribution function of *X* and *Y*, whereas $p(x)$ and $p(y)$ are the marginal probability distribution functions of *X* and *Y* respectively. If the log base 2 is used, the units of mutual information are the bit. Mutual information is a measure of the inherent dependence expressed in the joint distribution of *X* and *Y* which is relative to the joint distribution of *X* and *Y* under the assumption of independence. In mutual information,if *X* and *Y* are independent random variables then $I(X,Y) = 0$.

Mutual Information can be expressed as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (8)$$

Where

$$H = -\sum_i p_i log_2 p_i \quad (9)$$

Many levels of discretization[19] are introduced to improve the efficiency of MI Algorithms. The relevance between any two features/genes is effectively improved for selecting highly informative features.

Feature Selection using mRMR and Information Gain through Mean Score

Given a miRNA expression dataset *g* and a user-defined number *k*, the problem of finding the complete set of top-*k* high relevant miRNAs in *g* is to discover all the miRNA expressions that maximize the relevancy of features with the class label while it minimizes the redundancy in each class. The steps for identifying top-*k* features using two level feature selection with mRMR and mutual information are given below:

1. Identify top-*q* features using mRMR where *q*>*k*
2. Find the mutual information of top-*q* features as given below:

a. Find the set of mean values

$$mi = \{mi1^-, mi1^+, mi2^-, mi2^+, \ldots, miK^-, miK^+\}$$

for *gi* of its individual targeted classes $h = \{h1, h2, \ldots, hK\}$ in top-*q* ranked features.

Where mij^- represents the number of miRNA expression values less than its mean mij and mij^+ represents the number of miRNA expression values greater than or equal to it mean mij and $1 \le i \le q \& 1 \le j \le K$.

b. The Entropy of *gi* is calculated as given below :

$$H(gi) = -\sum_{i=1}^{2K} p_i log\, p_i$$

$$H(gi) = \left(\frac{mi1^-}{mi1}\log\left(\frac{mi1^-}{mi1}\right) + \frac{mi1^+}{mi1}\log\left(\frac{mi1^+}{mi1}\right) + \ldots + \frac{miK^-}{miK}\log\left(\frac{miK^-}{miK}\right) + \frac{miK^+}{miK}\log\left(\frac{miK^+}{miK}\right)\right)$$

c. Find the Relevant Information Gain $RIG(gi,h)$ using mi from the equation (8).

d. The *RIG* values are added with mRMR values obtained from equation (6) to strengthen the concept of mRMR (i.e) improving the relevance and reducing the redundancy.

3. The top-*k* highest features are selected as highly informative and discriminative features for classification.

Classification

Support Vector Machines (SVM) is a learning method based on statistical learning theory proposed by Vapnik and is a powerful tool for data classification and estimation[17]. SVM can handle a nonlinear classification efficiently by mapping input samples from the input space into a high dimensional feature space with the

nonlinear kernel function. A key factor in SVM is to use kernels to construct nonlinear decision boundary. We use radial basis kernels since it worked well than polynomial and linear kernels. In the feature space, SVM tries to maximize the generalization performance by solving a quadratic programming optimization problem, and finds the optimal separating hyperplane[17].ANNs are excellent computational models that have been implemented to solve different kind of problems. The pattern classification, forecasting and regression problems are areas where the ANN has demonstrated to be an efficient technique[18] ANN has been widely applied in DNA microarrays[18].The 10-fold cross validation method is a scaled down version of the leave one out method, where the dataset is divided into ten partitions. Each of these partitions is used as a testing set while the classifier is trained on the remaining samples.

IMPLEMENTATION AND RESULTS

The feature selection and classification algorithms are implemented in R3.2.5. Initially, AgilentG4502A Gene expression dataset is classified with top ranked feature set selected by mRMR with mean score using SVM and ANN. The same feature selection method and classifiers are applied for Agilent miRNA_8x15K miRNA expression dataset. For joint analysis, the gene expression and miRNA data are combined for feature selection and classification.The SVM-Radial Basis Function (SVM-RBF) is used as a classifier with the parameters cost=10 and gamma=0.001. It is observed that the combined analysis of mRNA and miRNA gives better classification accuracy. Initially, 100, 200, 400, 600 and 1000 features are chosen as subset of features which are ranked using feature selection methods. With the help of empirical analysis it is observed that 600 features in the combined feature set give better performance. In joint analysis, the top ranked genes are selected in different ways:

1. The top ranked genes from mRNA and miRNA datasets are selected as features for classifiers. It gives 90% accuracy.
2. The features from datasets are combined, selected and applied to the classifiers. It gives 100% classification accuracy.

The performance of the proposed work is compared with the existing feature selection techniques such as Improved Mutual Information (IMI), Fisher Score and Hybrid technique which combines IMI and Fisher Score with Standard deviation as a component, Entropy based mean score, mRMR. Table II shows the classification performance of SVM-RBF Classifier. The joint analysis gives 100% accuracy for all feature selection methods and the proposed method gives better accuracy for mRNA dataset.

Table II
Classification performance with SVM-RBF

Feature Selection Method	Sensitivity (%)			Specificity (%)			Classification Accuracy(%)		
	mRNA	miRNA	mRNA+ miRNA	mRNA	miRNA	mRNA+ miRNA	mRNA	miRNA	mRNA+ miRNA
Improved MI	100	98.7	100	100	86.7	100	97.65	96.66	100
F-Score	80.34	78.8	100	90	84.6	100	96.78	98.65	100
Hybrid(F-score +IMI)	87.6	85.6	100	84	90	97.8	98.7	96	100
Entropy Based Mean Score	93.5	90.4	98.7	84.5	92	100	98.3	97.6	100
mRMR	50	70	100	84	95	100	82.6	93.04	100
mRMR with mean score	98.6	90.3	100	98.6	98.4	100	98.8	93.6	100

The ANN is used as a classifier with the following components and their values are listed below: size=2, maxit and decay=0.001.The 95% CI value is 0.8512-1 and the average No Information Rate is 0.9391 with the ANN classifier. Table III shows the classification performance of ANN. The joint analysis gives 100% accuracy except hybrid technique. The proposed method has significant improvement for mRNA compared with other feature selection methods.

Table III
Classification performance of ANN

Feature Selection Method	Sensitivity (%)			Specificity (%)			Classification Accuracy(%)		
	mRNA	miRNA	mRNA+ miRNA	mRNA	miRNA	mRNA+ miRNA	mRNA	miRNA	mRNA+ miRNA
Improved MI	99.8	98.7	100	100	86.7	100	97.63	95.65	100
F-Score	83.34	78.8	100	90	84.6	100	94.78	98.60	100
Hybrid(F-score +IMI)	85.6	85.6	100	84	90	97.8	97.8	96.3	98.7
Entropy Based Mean Score	95.5	90.4	98.7	84.5	92	100	97.3	97.6	100
mRMR	55	70	97.6	84	95	100	83.6	92.04	100
mRMR with mean score	98.6	90.3	100	98.6	98.4	100	98.6	93.04	100

Table IV
Classifier performances with various training – testing partitions

Feature Selection Method	50%-50% Partitions Training-Testing			60%-40% Partitions Training-Testing			80%-20% Partitions Training-Testing		
	mRNA	miRNA	mRNA+miRNA	mRNA	miRNA	mRNA+miRNA	mRNA	miRNA	mRNA+ miRNA
SVM-RBF with EEMSmRMR	96.2	90.2	97.4	96.3	92.4	97.5	98.8	93.6	100
ANN with EEMSmRMR	96	92.3	98.4	97	94	98.2	98.6	93.04	100

Table IV shows the classification performance with various training – testing partitions for SVM and ANN. The 80%-20% ratio gives better result than 50%-50% and 60%-40% training-testing partitions.

Identification of Signature Bio Markers

The functions of RankProd- a Bioconductor package is used for the analysis of microarray data especially to identify the differentially expressed features[1]. It uses non-parametric method based on ranks of fold changes(FC) to identify up-regulated and down-regulated features under one condition against another condition (e.g. Tumor Vs Normal samples) or differentially expressed features under a specific condition (e.g. Stage I samples). This is based on the null hypothesis order of all items and probability of finding a specific item among the top r of n items in a list p = r/ n. Multiplying these probabilities leads to the definition of the rank product $RP = \prod_i \frac{r_i}{n_i}$, where r_i is the rank of the item in the i^{th} sample and n_i is the number of items in the i^{th}sample. The smaller the *RP* value, the smaller the probability that the observed placement of the item at the top of the lists is due to chance. The rank product is equivalent to calculate the geometric mean rank; and to replace the product by the sum leads to a statistics (average rank) that is slightly more sensitive to outlier data and puts a higher premium on consistency between the ranks in various lists[20,21].

Algorithm for Rank and pfp computations

1. Generate *p* permutations of *k* rank lists of length *n*
2. Calculate rank products of the *n* genes in the *p* permutations
3. Count how many times the rank products of the genes in the permutations are smaller or equal to the observed rank product. Let this value be c.
4. Calculate the average expected value for the rank product by $ERP(g) = c/p$
5. Calculate the percentage of false positives by $pfp(g) = \frac{ERP(g)}{rank(g)}$ where *rank* (*g*) is the rank of gene *g* in a list of all *n* genes sorted by increasing *RP*.

Fold-Change values

The Fold Change (FC) is calculated as a ratio of averages from control and test sample values and is a measure describing how much quantity changes from control to test sample values which is used to select the differentially expressed and discriminative genes in a microarray dataset with two biological conditions. For log2-fold change, its formula is $log2FC = log2(B) - log2(A)$. For calculating Fold change from log2, the following formula $power(2, log2\_value)$ is used.

Table 5 shows significant genes with various conditions. Mammaglobin B (SCGB2A1) is a novel tumor antigen highly differentially expressed in all major histological types of ovarian cancer which may represent a novel and attractive target for the immunotherapy of patients harboring recurrent disease resistant to chemotherapy[23,24]. It is observed that SCGB2A1 is highly expressed in Prior Malignancy of Breast Cancer, Stage I, Ovarian Triplet and is less expressed in Fallopian Normals. Matrilysin (MMP7) is overexpressed in all stages of ovarian cancer including epithelial ovarian cancer (EOC) invasion and metastasis[25]. From the Table V, it is observed that it is also highly expressed Gene next to the SCGB2A1. SCGB2A1, MMP7, EMX2, FCGR3A, DAPL1, RGS1, CD163, EHF are highly expressed, up-regulated and commonly present in Stage-I, Ovarian Triplet and Prior Malignancy of Breast Cancer cases. The SPON1 is up-regulated, highly expressed gene present in Ovarian Triplet patients.

Table V
Signature Genes related to various conditions of Ovarian Cancer

Various states of Ovarian Cancer	Signature Genes			
	Up-Regulated Genes with Fold- change values		Down-Regulated with Fold- change values	
Stage I	SCGB2A1	293.5831	TTR	0.0159
	MMP7	213.3754	C3orf57	0.0152
	EMX2	185.6868	DKK1	0.0143
	FCGR3A	140.6954	AHSG	0.0125
	DAPL1	127.0298	TYR	0.0117
	CD163	128.7888	FGB	0.011
	RGS1	116.4501	RPS4Y1	0.01
	EHF	115.9796	FABP1	0.0095
Positive to Neo-Adjuvant Therapy	FLJ22655	287.0818	C10orf81	0.0369
	OGN	96.93	LAMP3	0.0336
	MMRN1	73.3262	CCNA1	0.0369
	LRRC17	65.1782	C1orf172	0.0332

	HS3ST2	50.6136	VTCN1	0.0256
	ANGPTL5	56.018	TMPRSS4	0.0222
Prior Malignancy	CD163	265.9249	IGFBP1	0.0064
	FCGR3A	200.0633	FGL1	0.0098
	RGS1	192.9723	FGB	0.0099
	SCGB2A1	180.1783	DKK1	0.0104
	EMX2	146.7661	RPS4Y1	0.011
	ZBED2	145.6009	TYR	0.0113
	MMP7	100.4355	FABP1	0.0119
	DAPL1	101.3929	DSCR8	0.0125
	GGTA1	91.2419		
Fallopian Normal	SCGB2A1	154.1038	BUB1	0.0324
	FAM81B	124.7401	HBG1	0.0299
	ARMC3	125.7912	DSCR8	0.0305
	MMP7	124.5705	C3orf57	0.0273
	DAPL1	105.9425	PAH	0.0275
	CRISP3	93.6339	IGFBP1	0.025
	FLJ44379	94.9876	AHSG	0.0253
	TTC29	90.3322	RPS4Y1	0.0239
			FABP1	0.0198
			DKK1	0.0167
Ovarian Triplet	SCGB2A1	172.319	IGFBP1	0.0056
	MMP7	170.883	FGL1	0.0094
	RGS1	134.9849	RPS4Y1	0.0119
	SPON1	129.3304	FABP1	0.0169
	EMX2	118.7361	FGB	0.0141
	FCGR3A	104.0767	AHSG	0.0143
	CD163	91.6076	TYR	0.0147
	EHF	76.3206	UTS2	0.0183
			DKK1	0.0219
			TTR	0.0183
Prior Malignancy of Breast Cancer	SCGB2A1	379.2888	HBG1	0.0155
	MMP7	303.5474	TTR	0.0152
	EMX2	198.0159	DKK1	0.013
	FCGR3A	167.7239	TYR	0.011
	DAPL1	160.7336	AHSG	0.011
	RGS1	159.9741	FGB	0.0112
	CD163	157.5652	FABP1	0.0092
	EHF	141.8555	RPS4Y1	0.0081

Table VI

Signature miRNAs related to various conditions of Ovarian Cancer

Various states of Ovarian Cancer	Signature miRNAs			
	Up-Regulated miRNAs with Fold- change values		Down-Regulated with Fold- change values	
Stage I	hsa-miR-923	31167.77	hsa-miR-1231	23.3493
	hsa-miR-21	29632.374	kshv-miR-K12-9	23.4797
	hsa-let-7a	13804.045	hsa-miR-323-5p	23.5784
	hsa-miR-141	12638.893	ebv-miR-BART17-5p	23.6334
	hsa-let-7b	10428.578	ebv-miR-BHRF1-2	23.7664
	hsa-let-7f	7699.379	hsa-miR-31	38.8823
	hsa-miR-29a	6953.276	kshv-miR-K12-6-3p	23.9013
	hsa-miR-16	6789.771	ebv-miR-BHRF1-3	23.749
Positive to Neo-Adjuvant Therapy	hsa-miR-126*	9.1377	hsa-miR-141	0.0188
	hsa-miR-9*	7.042	hsa-miR-200b	0.0173
	hsa-miR-9	6.7634	hsa-miR-205	0.0236
	hsa-miR-1	6.2224	hsa-miR-200c	0.0276
	hsa-miR-143	6.5736	hsa-miR-449a	0.0461
	hsa-miR-424	4.4578	hsa-miR-923	0.042
	hsa-miR-145*	4.9248	hsa-miR-200a	0.0435
	hsa-miR-99a	4.7276	hsa-miR-429	0.0443
Prior Malignancy	hsa-miR-21	38980.089	hsa-miR-449a	10.0889
	hsa-miR-923	28871.556	kshv-miR-K12-9	22.759
	hsa-let-7a	12923.208	hsa-miR-1231	23.1114
	hsa-miR-451	16800.45	ebv-miR-BART17-5p	23.2136
	hsa-let-7b	11123.556	hsa-miR-587	23.2049
	hsa-miR-141	10836.939	hsa-miR-941	23.2472
	hsa-let-7f	7232.857	kshv-miR-K12-3*	23.269
	hsa-miR-29a	6338.624	hsa-miR-620	23.2838
Fallopian Normal	hsa-miR-923	87160.926	hsa-miR-7	13.6039
	hsa-let-7b	26438.749	hsa-miR-182	15.0308
	hsa-miR-141	15931.234	hsa-miR-335*	18.6641
	hsa-miR-145	11499.646	hsa-miR-135a	13.9609
	hsa-miR-21	10809.32	hsa-miR-183	20.5429
	hsa-miR-181a	12771.121	hsa-miR-182*	22.4469
	hsa-miR-22	10558.831	hsa-miR-183*	22.9277

	hsa-miR-26a	10497.886	hsa-miR-218-2*	22.9183
	hsa-miR-21	36762.682	kshv-miR-K12-9	23.1295
Ovarian Triplet	hsa-miR-923	35658.253	hsa-miR-1231	23.3222
	hsa-let-7a	17316.912	ebv-miR-BART13*	23.5037
	hsa-let-7b	14115.604	hsa-miR-219-1-3p	23.5787
	hsa-miR-29a	10778.812	hsa-miR-941	23.5381
	hsa-let-7f	8774.276	ebv-miR-BART17-5p	23.609
	hsa-miR-141	8943.648	kshv-miR-K12-3*	23.6002
	hsa-miR-24	7556.264	hsa-miR-208a	23.6724
Prior Malignancy of Breast Cancer	hsa-miR-21	32151.352	kshv-miR-K12-9	22.9717
	hsa-miR-923	37407.246	ebv-miR-BART2-3p	23.3753
	hsa-miR-141	15205.033	ebv-miR-BHRF1-3	23.3864
	hsa-let-7a	12656.845	hsa-miR-1231	23.2317
	hsa-let-7b	11216.015	kshv-miR-K12-3*	23.3754
	hsa-miR-29a	9256.809	hsa-miR-323-5p	23.4036
	hsa-let-7f	6525.759	ebv-miR-BHRF1-2	23.6184
	hsa-miR-24	6885.949	hsa-miR-452*	23.7059

Table VI shows the signature miRNAs related with various conditions of Ovarian Cancer. The let-7 family, miR-21 and miR-141 are the most predominant and abundantly expressed miRNAs across the different development stages[25]. It is observed that let-7 family and miR-21 are present in Ovarian Triplet, Stage-I, Prior Malignancy of Breast Cancer and Prior Malignancy. MiR-200a, miR-200b, miR-200c and miR-141 are overexpressed in ovarian cancer markers[6]and are observed in Prior Malignancy of Ovarian Cancer. In this analysis the Prior Malignancy of Ovarian Cancer and Positive to Neo-Adjuvant Therapy are analyzed with Rank Product two-class case and all others are analyzed with Rank Product one class case method.

Table.VII

Performance analysis of proposed method usingjoint analysis with existingtechnique

Reference	Techniques	Data Sets	% of Accuracy
Shinuk Kim et al., (2013)	RF&SVM, FSCR_REG, FCCR_SVM	Ovarian Cancer from TCGA	86.36
Hasan Oğul et al.,(2013)	Information Gain, Gain Ratio, Chi squared, CFS	Ovary	93.3
Sihua Peng et al., (2009)	nREF	Ovary	93.3(mRNA) 85.4(miRNA)
This Work	EEMSmRMR - Improved mRMR	Ovarian Cancer from TCGA	100

Table VII analyses the performance of various techniques with the proposed technique. It is proved that mRMR with mean score technique achieved 100% result with Joint Analysis of mRNA and miRNA data sets from TCGA.

CONCLUSION

Specific Gene / miRNA expression signatures in ovarian cancer are associated with diagnosis, prognosis, and therapy response. Moreover it has been suggested that Biomarker miRNAs and Genes are essential molecular pathways in the initiation and/or progression of human cancers. A better understanding of miRNA and Gene expressions in cancer may uncover novel molecular pathways, or novel mechanisms of activation for known pathways. It is proved that the integration of mRNA and miRNA data yields better classification accuracy than mRNA and miRNA data separately. Based on the experimental results, it is shown that the classification of ovarian cancer dataset with the mRMR with mean score feature selection technique significantly improves the prediction performance and provides a better classification accuracy for mRNA data. The joint analysis gives 100% accuracy for selected feature selection methods. The Rank Product function from Bioconductor package is effectively used for identifying signature genes and miRNAs which are significant in the discovery of different states of ovarian cancer and therapeutics.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Kumar G, Breen EJ, Ranganathan S,Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework,BMC Systems Biology,2013 Feb,6;7:12.
2. Lavanya R and Ramesh.S, Immunohistochemical expression of ki-67 in ovarian tumors& correlation with clinicopathological factors,Int J Pharm Bio Sci 2016 Jan; 7(1): 67 – 73.
3. http://ucsdnews.ucsd.edu/pressrelease/ovarian_cancer_specific_markers_set_the_stage_for_early_diagnosis_personalized_treatments.
4. Zaman MS, Maher DM, Khan S, Jaggi M, Chauhan SC. Current status and implications of microRNAs in ovarian cancer diagnosis and therapy. Journal of ovarian research. 2012 Dec 13;5(1):44.
5. Di Leva G, Croce CM. The role of microRNAs in the tumorigenesis of ovarian cancer. Frontiers in oncology. 2013 Jun 13;3:153.

6. George GP, Mittal RD. MicroRNAs: Potential biomarkers in cancer. Indian Journal of Clinical Biochemistry. 2010 Jan 1;25(1):4-14.
7. Ibrahim R, Yousri NA, Ismail MA, El-Makky NM. miRNA and gene expression based cancer classification using self-learning and co-training approaches. InBioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on 2013 Dec 18 (pp. 495-498). IEEE.
8. Kim S, Park T, Kon MA. Computational methods for cancer survival classification using intermediate information. InIWBBIO 2013 (pp. 517-525).
9. Oğul H, Altındağ O. Integrating MicroRNA and mRNA Expression Data for Cancer Classification, ICPRAM2013 International Conference on Pattern Recognition Applications and Methods, 2013, pp: 503-507.
10. Peng S, Zeng X, Li X, Peng X, Chen L. Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. Journal of Genetics and Genomics. 2009 Jul 31;36(7):409-16.
11. Konstantinopoulos PA, Cannistra SA, Fountzilas H, Culhane A, Pillay K, Rueda B, Cramer D, Seiden M, Birrer M, Coukos G, Zhang L. Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. PLoS One. 2011 Mar 29;6(3):e18202.
12. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology. 2005 Apr;3(02):185-205.
13. Mandal M, Mukhopadhyay A. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. Procedia Technology. 2013 Jan 1;10:20-7.
14. Unler A, Murat A, Chinnam RB. mr 2 PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Information Sciences. 2011 Oct 15;181(20):4625-41.
15. Ding CH. Analysis of gene expression profiles: class discovery and leaf ordering. InProceedings of the sixth annual international conference on Computational biology 2002 Apr 18 (pp. 127-136). ACM.
16. Goh L, Song Q, Kasabov N. A novel feature selection method to improve classification of gene expression data. InProceedings of the second conference on Asia-Pacific bioinformatics-Volume 29 2004 Jan 1 (pp. 161-166). Australian Computer Society, Inc...
17. Xu R, Xu J, Wunsch DC. MicroRNA expression profile based cancer classification using Default ARTMAP. Neural Networks. 2009 Aug 31;22(5):774-80.
18. Garro BA, Rodríguez K, Vázquez RA. Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing. 2016 Jan 31;38:548-60.
19. Anidha M, Premalatha K. A Hybrid Gene Selection Technique Using Improved Mutual Information and Fisher Score for Cancer Classification Using Microarrays. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2016 Apr 1;10(3):585-8.
20. M. Anidha, K. Premalatha,An Entropy Based Mean Score Feature Selection method for Identification of Biomarkers using miRNA expression profiles for Cancer Classification,Asian Journal of Information Technology, Accepted on 25th May 2016.
21. Hong F. Bioconductor Rank Prod Package Vignette,2010[Updated on Oct 17,2016] Available From: https://pdfs.semanticscholar.org/bb59/a69a587d90917e3afaab1d494119a5eae58f.pdf.
22. Breitling R, Rank Product (updated on Jan, 92017) Available from: https://en.wikipedia.org/wiki/Rank_product.
23. Bellone S, Tassi R, Betti M, English D, Cocco E, Gasparrini S, Bortolomai I, Black JD, Todeschini P, Romani C, Ravaggi A. Mammaglobin B (SCGB2A1) is a novel tumour antigen highly differentially expressed in all major histological types of ovarian cancer: implications for ovarian cancer immunotherapy. British journal of cancer. 2013 Jul 23;109(2):462-71.
24. Fischer K, von Brünneck AC, Hornung D, Denkert C, Ufer C, Schiebel H, Kuhn H, Borchert A. Differential expression of secretoglobins in normal ovary and in ovarian carcinoma–Overexpression of mammaglobin-1 is linked to tumor progression. Archives of biochemistry and biophysics. 2014 Apr 1;547:27-36.
25. Li Y, Fang Y, Liu Y, Yang X. MicroRNAs in ovarian function and disorders. Journal of ovarian research. 2015 Aug 1;8(1):51.
26. Lisowska KM, Olbryt M, Dudaladava V, Pamuła-Piłat J, Kujawa K, Grzybowska E, Jarzab M, Student S, Rzepecka IK. Gene expression analysis in ovarian cancer–faults and hints from DNA microarray study, Frontiers in Ontology,2014,Vol.4
27. MASOUM S, GHAHERI S. Feature selection and classification of microarray gene expression data of ovarian carcinoma patients using weighted voting support vector machine. Iranian Journal of Mathematical Chemistry. 2013 May 1;4(2):163-75.
28. Hong JH, Cho SB. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. Artificial intelligence in Medicine. 2006 Jan 31;36(1):43-58.
29. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000 Oct 1;16(10):906-14.
30. Kumar M, Rath NK, Swain A, Rath SK. Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor. Procedia Computer Science. 2015 Jan 1;54:301-10.

Reviewers of this article

Asst.Prof.Dr. Sujata Bhattacharya

Assistant Professor, School of Biological and Environmental Sciences, Shoolini University, Solan (HP)-173212, India

Dr C.Gunavathi

Associate Prof,School of information Technology and Engineering,VIT University,Vellore-632014 9442091979

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.