J Biochem Tech (2017) 8(4): 1116-1120
ISSN: 0974-2328

Novel Gene Selection Method for Breast Cancer Classification

Geetika Munjal, Madasu Hanmandlu, Sangeet Srivastava

Received: 14 March 2017 / Received in revised form: 02 May 2017, Accepted:12 May 2017, Published online: 27 May 2017

Abstract

Classifying different breast cancer subtypes is of great importance in breast cancer diagnosis and prevention, and Microarray data helps in cancer diagnosis and prognosis. However due to large number of features, its processing has become very complex. The current work has been done to i) find the smallest set of genes that can ensure accurate classification of breast cancer subtypes from microarray data with the help of novel gene selection method, ii) express similarity and dissimilarity among breast cancer samples based on ER and PR status using selected genes.

Keywords: Breast Cancer, Classification, Receptor Status, Feature Selection

Introduction

Breast cancer is a large, heterogeneous class of cancer in which a group of cells undergo uncontrolled growth, and destroy the adjacent tissues (Saini, 2014). Classifying different breast cancer subtypes is of great importance in breast cancer diagnosis, prevention, or drug discovery. The microarray technology has been effectively used in diagnosis, prognosis, or treatment outcome prediction of breast cancer. However, it has generated large amount of raw data related to cancer that is hard to analyze in clinical practice (Desper *et al.*, 2004). Therefore, a computational approach may help us to reduce this effort by finding meaningful information in a time efficient manner like classifying cancer subtypes, identifying similarity and dissimilarity among breast cancer patients.

Breast cancer can be diagnosed by testing the expression level of the Estrogen Receptor(ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2(HER2) (Dai *et al.*, 2014; Parise and Caggiano, 2014). The ER, PR and HER2 are activated by the hormone estrogen, and they modulate the activity levels of various genes. Status of receptors is expressed positively or negatively in breast cancer patients. The receptor's status is valuable for both diagnosis and prognosis of breast cancer, and also helps in categorizing the disease into various molecular classes. In this work, the relevant genes that can classify and cluster breast cancer cases according to ER and PR status have been identified. The clustering helps in finding intra-tumor heterogeneity among breast cancer patients (Saini, 2014).

To proceed further, there is a need to know various clinical factors present in breast cancer microarray data described in table 1.

Table 1- clinical factors of breast cancer

Clinical Covariate	Description	Level
Size	Size of tumor	-
Age	Age of patient	18-60
Histological Grade	Tumor History	IDC.TUB IDC: Invasive ductile Carcinoma ILC: Invasive Lobular Carcinoma
ER	Estrogen Receptor status	1=positive, 0=negative, null
HER2	Human Epidermal growth Receptor status	1=positive, 0=negative, null
PR	Progesterone Receptor	1=positive, 0=negative, null
Treatment	Treatment available	CT: Chemotherapy RT: Radiation Therapy HT: Hormone therapy
DMFS	Distant Metastasis Free survival	0. No Metastasis 1. Metastasis
RFS	Recurrence Free Survival	0- No Recurrence 1- Recurrence
DFS	Disease Free survival	0- No disease 1- Disease
OS	Overall Survival	0- Surviving 1- Dead

Geetika Munjal*, Sangeet Srivastava
The NorthCap University, Gurugram, India.
*Email: munjal.geetika@gmail.com

Madasu Hanmandlu
Indian Institute of Technology, Delhi, India,

Along with parameters described in table 1, there are also thousands of genes with their expression values. The monitoring of all three receptors, i.e. ER, PR and HER2 has shown a significant rise in patient survival. In most of the cases, ER positive (ER+) has had a better survival than ER negative (ER-), which signifies that ER is an important factor in survival analysis of breast cancer patients (Saeys *et al.*, 2007). Therefore, ER status identification is a crucial component of breast cancer treatment strategy. The IHC (ImmunoHistoChemistry) markers including ER, PR and HER2, have been extensively used in finding the intrinsic differences among breast cancer subgroups based on mRNA and miRNA. The classification of IHC surrogates is useful in survival analysis of patients (Saini, 2014).

In all the studies, whether identifying cancer classes or heterogeneity in the form of clusters, performance is highly dependent on the selected genes. It has also been observed that in cancer disease, the driver genes change the cancer progression and they even affect the participation of other genes. In this study, considering the importance of genes, it has been focused on finding relevant genes that can classify and cluster breast cancer according to their receptor status. For which experiments have been done on Microarray data of breast cancer in the current study.

Literature Review

Microarray gene data enables researchers to examine several genes in parallel. Microarray data comes in different flavors and forms as they are produced by various companies (Affymetrix, Agilent, etc.) and in different laboratories (Saeys *et al.*, 2007). Gene expression data has some characteristics which make them complex and challenging compared to other types of data. The major limitation is that sample size is very small compared to the number of gene features. Some of the genes are redundant with no known significance, for a particular study. Hence, feature extraction and feature selection are the two main approaches that can resolve these issues. Feature extraction focuses on dimensionality reduction and comes up with the most relevant information from the genes. Feature selection is the process of selecting the subset of relevant genes (Guyon *et al.*, 2002; Chandra and Gupta, 2011).

Feature selection methods aim at identifying gene signatures that can help in cancer classification, molecular subtyping or identifying heterogeneity in cancer. A lot of work has been done in identifying gene signatures relevant to breast cancer sub-typing including the 70-gene mammaprint and the 21-gene signatures oncotype, that have predictive as well as prognostic values (Arranz *et al.*, 2012). Predictive analysis of microarray (PAM50) (Desper *et al.*, 2004; Arranz *et al.*, 2012) has also been extensively used in identifying molecular subtyping of cancer. Currently, molecular subtyping of breast cancer has led to Luminal, Basal, and HER2-enriched subtypes, which have specific clinical behavior. The existing gene signatures, viz., mammaprint and oncotype have been extensively used in classifying breast cancer subtypes (Arranz *et al.*, 2012). But, these gene signatures have their own limitations including the associated cost, requirement to send samples to a reference center, and the existence of new subtype (Arranz *et al.*, 2012). To overcome these problems, computational approaches can be explored for gene feature identification.

Gene Feature Selection

Feature selection using computational approaches has gained the attention of researchers in various fields. It is a process of selecting a subset of significant features algorithmically, that are used in building a classifier (Chandra and Gupta, 2011). Feature selection decreases time and space complexity of an algorithm. Therefore, these techniques are more efficient, and are used to extract only relevant and unique features. Three fundamental feature selection techniques are the filter, wrapper, and the embedded methods which can be applied to gene feature selection as well.

A filter based approach, integrated prognosis, and risk estimation (IPRE) (Saini, 2014) have been developed to achieve higher classification accuracy for good and poor prognosis in breast cancer patients. Good prognosis patients remain free from the recurrence of breast cancer for at least five years, whereas poor prognosis patients may have a relapse of breast cancer within five years. To select the gene signature, based on prognosis, virtual chromosome score was calculated for each gene, IPRE has achieved 82% accuracy, 88% specificity and 95% specificity on various dataset (Saini, 2014). A rank based method was used to select 231 genes and they were ranked in the descending order of their correlation coefficient (Saeys *et al.*, 2007). Of these 231 genes, five genes have been repeatedly taken from the top of the rank ordered list and added to the prognosis classifier to optimize it. LOOCV method was applied for the evaluation of the prognosis classifier, and 70 genes were selected to form the gene signature. Cancer genes can also be selected on the median absolute deviation or coefficient variation (Chandra and Gupta, 2011), this approach has been used to discover cancer subtypes where genes are assigned weights according to the page ranking algorithm. It has provided solution to cluster breast cancer subtypes using microarray data. Statistical approach has also been used for selecting genes with higher weights where weights are assigned according to class discrimination capability (Chandra and Gupta, 2011; Lu, 2003). T-test can also be used for feature selection in a two-class problem. In this method, genes are ranked according to their t-statistic value to select important features. Class separability measure can further refine the selected genes to improve classification performance (Wang *et al.*, 2007).

Material and Method

The literature has suggested that to obtain a robust gene signature, there is a need to reduce the gap between the sample size and the number of gene features. This requires increasing the number of samples by integrating various datasets and using gene feature selection (Arranz *et al.*, 2012). To get the highest possible performance based on high gene relevance score, the $gene_{score}$ algorithm has been introduced and applied on the increased sample size by integrating three datasets.

For experimental purposes, the datasets used provided a comparable ratio between the numbers of ER/PR positive samples to ER/PR negative samples. HER2 status was not balanced in any of the datasets as almost in every dataset, 70-80% of the samples were HER2 negative. The selected datasets were: GSE25055, GSE20271 and GSE21974 described in Table 5.2; that have been downloaded from NCBI (Kim *et al.*, 2004). The sample selection was independent of age, tumour grade and other clinical parameters except for ER, PR and HER2 status, therefore the samples with missing ER, PR and HER2 status have been eliminated from the structure of the paper

To integrate the three datasets, all of them were needed to have the same gene names instead of probe-ids. Therefore, the probe-ids have been mapped into gene names, and some probes were deleted, as they did not have corresponding gene names. Many gene names were repeated because multiple gene expressions existed which were corresponding to a particular gene name. The mean was calculated for such genes so that a unique gene expression value corresponding to a particular gene name could be obtained. To integrate the three datasets, their values were normalized using the min-max technique.

In the integrated dataset, there were a total of 925 samples whose ER, PR and HER2 status was known, and all the experiments were done on them or their subsets. This dataset has been described as per its ER, PR and HER2 status count in Table 2.

Support Vector Machine (SVM) classifier has been used to compare the results of the $gene_{score}$ algorithm with the existing feature selection techniques. The reason for selecting SVM was that it has the capability to analyze the broad patterns of microarray gene expression data, and classify the cancer classes (Maltseva *et al.*, 2013). SVM is a supervised learning algorithm which has a robust performance on the noisy and sparse data and has been used in a wide number of applications. This classifier considers the correlations in the microarray data without taking into account the structure of data.

Table 2- dataset description

Dataset name	Number of samples	ER+	ER-	PR+	PR-	HER2+	HER2-
GSE20194	279	161	106	118	107	55	203
GSE20271	178	97	80	83	79	26	151
GSE25055	508	300	207	243	238	6	498

The process of the proposed model has been shown in the form of a flow chart in Fig. 1. The input gene data in $gene_{score}$ was normalized using min-max function expressed in Eq. (1).

$$g_i = \frac{g_i - g_{\min}}{g_{\max} - g_{\min}} \quad (1)$$

where g_i the gene intensity of i_{th} gene, g_{min} is the minimum gene intensity of i_{th} gene.

$$r = \frac{\sum_{i=1}^{n}(x_i - x_{mean})(y_i = y_{mean})}{\sqrt{\sum_{i=1}^{n}(x_i - x_{mean})^2}\sqrt{\sum_{i=1}^{n}(y_i - y_{mean})^2}} \quad (2)$$

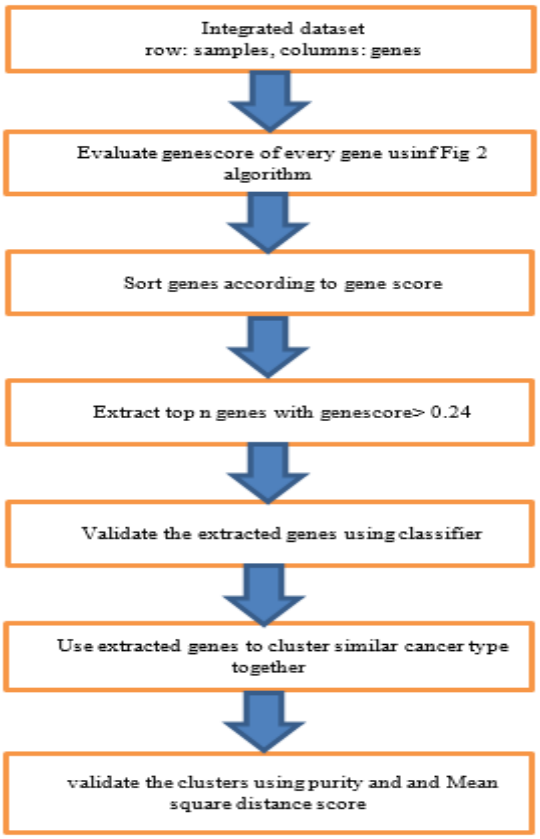

Fig 1. Flowchart of proposed approach

Algorithm $gene_{score}$ (m, n, CSscore)

```
for i= 1: m
        for j= i+1: n
        find corr_score of gen (i, j) using Eq. (2)
        end
end
if corr_score <=avg(corr_score)
        gene_new =gene
for i= 1: n
Evaluate the CS_score of gene_new using Eq. (5.3)
end
gene_new1=gene_new sorted according to CS_score
evaluate gene_score for gene_new1 using Eq. (5)
```

Fig 2. Algorithm $gene_{score}$
(gene (i, j) is gene intensity of i^{th} gene and j^{th} sample, m is the number of samples, n is the number of gene features, and $CS_{score\ is}$ output class seperability score)

After normalization, the correlation score among genes was obtained using Eq. (2) where x_i and $y_{i\ were}$ a pair of genes, in Fig 2 $gene_{new}$ is a set of features that are least correlated with any other gene, i.e. the genes features having the correlation score less than the average correlation core. After obtaining class separability (Wang *et al.*, 2007) score of each selected gene, $gene_{new}$ was obtained for two classes, i.e. ER+ and ER- as expressed in Eq. (3)

$$CS_g = \frac{D_g}{S_g} \quad (3)$$

where Dg is the sum of squares of the difference between the two samples of different classes, i.e. ER+ and ER- (interclass difference) expressed in Eq. (4), where m is the possible number of classes and y¯_g expressed in Eq. (5) is the mean of each gene in a particular class. S_ (g) is the sum of squares expressed in Eq. (6) as difference between the same class samples (intra class difference) for each gene (Wang *et al.*, 2007). CS (class separability score) was computed for every gene, and a higher class separability score indicated the capability of a gene in identifying classes.

$$D_g = \sum_{g=1}^{m} (\overline{y_{gm}} - \overline{y_g})^2 \quad (4)$$

$$\overline{y_g} = \sum_{f=1}^{n} \frac{y_g}{n} \quad (5)$$

$$S_g = \sum_{m=1}^{l} \sum_{f \in G_m} (y_g - \bar{y}_g)^2 \quad (6)$$

The class separability score again ranked the genes. After that, the genes were sorted according to this score. The sorted list of genes reflected the highly associated genes for any class on the top. Finally, $gene_{score}$ was calculated using Eq. (7). It was evaluated as an inverse square root of the number of genes n, $gene_{score}$ as follows:

$$\mathbf{gene_{score}} = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \mathbf{gene_{new1}}\right) \quad (7)$$

The top genes having a maximum gene score would help in identifying the relevant genes. The results of $gene_{score}$ algorithm were compared with other feature selection methods.

Results and Discussions

The proposed method was validated on various parameters including accuracy, sensitivity, and specificity of the top selected genes. The aim of this study was to select the minimum number of genes and also, show the relationship between the various cancer subtypes in form of clusters. The analysis was started by selecting the top 111 with a gene score of 0.24 and above using genescore to classify samples according to ER status, and {ER, PR} status. The reason for selecting the genes having a robust score above 0.24 was to find an optimum number of features. If this score is reduced to 0.23 or below, then the number of features is doubled, but the performance remains almost the same for the samples.

The results of $gene_{score,}$ were compared with other feature selection methods, i.e., mRMR (Chandrashekar and Sahin, 2014), fast correlation based feature selection (Yu and Liu, 2003), and class seperability (Peng, 2005). Tables 3 to 5 show that the results of the selected genes for ER, PR and HER2 status from t-test. It can be seen that the classification results of $gene_{score}$ algorithm were clearly better in terms of accuracy, sensitivity than those of mRMR, FCBF, class seperability.

The top 87 genes have been further selected with a robust score of 0.24 and above using $gene_{score}$ to classify patients according to PR status. The results of the selected top 87 were then compared with those of correlation based FCBF, mRMR, class seperability. It can be seen in table that the performance of $gene_{score}$ is comparable to those of mRMR, FCBF and class seperability methods considering ER and PR status. Whereas for HER2 receptor, none of the methods have performed very well, the reason was the imbalance between HER2 negative and HER2 positive samples.10-folds cross validation has been applied to validate the results in this study.

Table 3- Accuracy of genes selected using $gene_{score}$ and other methods according to various receptor status

Status	No of Features	FCBF	Class seperability	mRMR	$gene_{score}$
ER status	111	89.1	87	87.2	90.3
PR status	87	87.3	86.27	92.2	89.2667
HER2	156	78.2	70	75.01	58.2

Table 4- sensitivity of genes selected using gene score and other methods according to various receptor status

Recceptor	No of Features	FCBF	Class seperability	mRMR	$gene_{score}$
ER status	111	87	79.3	89.2	88
PR status	87	86.27	88	87	87.3
HER2	156	70	90.1	63.3	62

Table 5- specificity of genes selected using $gene_{score}$ and other methods according to various receptor status

Receptor	No of Features	FCBF	Class seperability	mRMR	$gene_{score}$
ER status	111	82.1	88	87	89.1
PR status	101	86.8	87.3	86.2	87.3
HER2	156	65.6	62	88	70.3

The first aim in this study was to select the minimum possible features with the highest $gene_{score}$, and the second aim was to form the cluster of cancer subtypes. The shortlisted genes for ER and PR status have been combined, giving a total of 165 genes. Cluster of ER-PR-, ER-PR+, ER+PR- and ER+PR+ was formed using hierarchical clustering algorithm, so we have validated the clusters based on purity metrics. The purity gave a measure of how many samples of a particular class were clustered together. It has been expressed in Eq. (8).

$$\text{purity} = \frac{\max(\text{number of sample correctly clustered})}{\text{number of samples}} \quad (8)$$

The cluster formed using $gene_{score}$ has performed well in grouping tumors for a particular type, although there was some contamination in each case. The clusters have been validated based on purity metrics. The average purity of $gene_{score}$ algorithm in Table 6 was 0.83, whereas purity of mRMR was 0.80, FCBF was 0.75 and the class seperability was 0.78. None of the algorithms has achieved a very high purity as the distinction between intertumoural and intratumoral heterogeneity was not clear cut. The accuracy assessment of the clusters (ER-PR-, ER+PR-, ER-PR+ and ER+PR+) was done on the basis of Mean Squared Distance (MSD). The MSD within class where low MSD (intra cluster distance) depicted compact cluster has been examined.

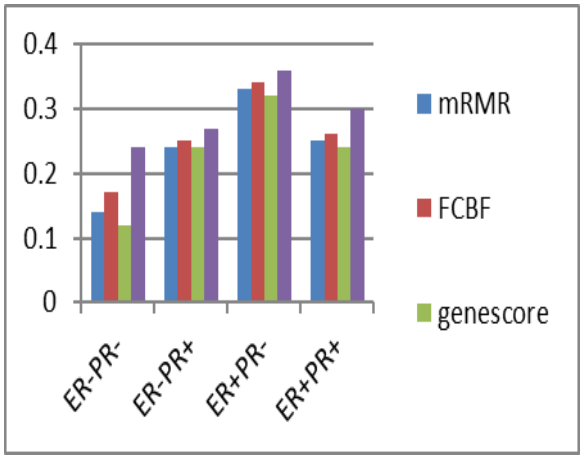

Fig 3. Assessment of Breast Cancer Cluster Quality based on Clustering Common Labels using Mean Square Distance

Fig 3 shows that the network inferred by the gene selected from $gene_{score}$ shows better results than those of t-test and Fisher test methods. The quality of tree was highly dependent on the selected genes however, the results can improve using some correction and optimization methods (Vijver *et al*., 2007; Yu and Liu, 2003).

Table 6- Purity of Sample of Fig 5.5

Method	ER-PR-	ER-PR+	ER+PR-	ER+PR+
$gene_{score}$	0.79	0.78	0.91	0.84
mRMR	0.74	0.76	0.85	0.86
FCBS	0.71	0.86	0.73	0.72
Class seperability	0.69	0.76	0.83	0.86

Conclusion

This paper has provided an insight into microarray gene expression data related to breast cancer. A very low number of samples were available in the microarray gene expression data as compared to the number of genes, and to reduce this gap, the multiple datasets of breast cancer has been integrated. The proposed $gene_{score}$ algorithm was applied to this integrated data which achieved a high classification performance in terms of accuracy, sensitivity, and specificity for classifying ER+ and ER- groups and PR+ and PR- groups as compared to the other filter based methods including t-test and Fisher score that selected the genes that were the least correlated with other genes and had a high class separability score.

The selected genes were used to group patient samples based on ER, PR status. The $Gene_{score}$ was also used to find the relevant genes to classify samples based on PR status. The genes for ER and PR status were combined to find cluster of patients with ER-PR-, ER+PR-, ER-PR+ and ER+PR+ cancer subtype. The cluster of samples expressed heterogeneity among cancer subtypes with respect to ER, PR status.

Acknowledgement

The research was a part of funded project under the department of science and technology with number: SR/WOS-A/ET-1015/2015(G).

References

Arranz E. E., Vara J. Á. F., Gámez-Pozo A., and Zamora P., “Gene Signatures in Breast Cancer: Current and Future Uses,” Transl. Oncol., vol. 5, no. 6, pp. 398–403, 2012.

B. Chandra and Gupta M., “An efficient statistical feature selection approach for classification of gene expression data,” J. Biomed. Inform., vol. 44, no. 4, pp. 529-535, 2011.

Dai X., Chen A., and Bai Z., “Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling,” Scientific Reports, vol. 4, no.1, 2014.

Desper R., Khan J. and Schäffer A., "Tumor classification using phylogenetic methods on expression data", Journal of Theoretical Biology, vol. 228, no. 4, pp. 477-496, 2004.

G. Chandrashekar and Sahin F., “A survey on feature selection methods,” Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.

Guyon I., Weston J., Barnhill S., Vapnik V., “Gene Selection for cancer classification using Support Vector Machines,” Mach. Learn., vol. 46, no. 1.3, pp. 1–39, 2002.

H. Peng, “Feature Selection Based on Mutual Information:” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.

Kim C. et al., “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.,” N. Engl. J. Med., vol. 351, no. 27, pp. 2817–26, 2004.

L. Wang, F. Chu, and Xie W., “Accurate cancer classification using expressions of very few genes,” IEEE Trans. Comput. Biol. Bioinforma., vol. 4, no. 1, pp. 40–53, 2007.

L. Yu and Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, pp 856–863, 2003

Maltseva D. V et al., “High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples,” J. Clin. Bioinforma., vol. 3, no. 1, p. 13, 2013.

Parise C. A. and Caggiano V., “Breast Cancer Survival Defined by the ER / PR / HER2 Subtypes and a Surrogate Classification according to Tumor Grade and Immunohistochemical Biomarkers,” J. Cancer Epidemiol., vol. 2014, pp. 1–12, 2014.

Saeys Y., Inza I., and Larranaga P., “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

Saini A., “Breast Cancer Prognosis Risk Estimation using Integrated Gene Expression and Clinical Data,” Biomed Res. Int., vol. 2014, pp. 34–50, 2014.

Vijver V. et al., “Gene expression profiling for prognosis of breast cancer,” Breast Cancer Research, vol. 9, no. S1, 2007.

Y. Lu, “Cancer Classification Using Gene Expression Data,” Inf. Syst., vol. 28, no. 4, pp. 1–35, 2003.