Iranian J Publ Health, Vol. 43, No.8, Aug 2014, pp. 1091-1098

Original Article

A Gene Selection Method for Survival Prediction in Diffuse Large B-Cell Lymphomas Patients using 1D Discrete Wavelet Transform

Maryam FARHADIAN [1], ****Hossein MAHJUB*** [2], ***Abbas MOGHIMBEIGI*** [3], ***Jalal POOROLAJAL*** [3], ***Muharram MANSOORIZADEH*** [4]

1. *Dept. of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan. Iran.*
2. *Research Center for Health Sciences and Dept. of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
3. *Modeling of Noncommunicable Diseases Research Center and Dept. of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
4. *Dept. of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamadan, Iran*

***Corresponding Author:** Email: mahjub@umsha.ac.ir

(Received 21 May 2014; accepted 09 July 2014)

Abstract

Background: An important aspect of microarray studies includes the prediction of patient survival based on their gene expression profile. To deal with the high dimensionality of this data, use of a dimension reduction procedure along with the survival prediction model is necessary. This study aimed to present a new method based on wavelet transform for survival relevant gene selection.

Methods: The data included 2042 gene expression measurements from 40 patients with Diffuse Large B-Cell Lymphomas (DLBCL). The pre-processing gene expression data is decomposed using third level of the 1D discrete wavelet transform. The detail coefficients at levels 1 and 2 are filtered out and expression data reconstructed using the approximation and detailed coefficients at the third level. All the genes are then scored based on the *t* score. Then genes with the highest scores are selected. By using forward selection method in Cox regression model, significant genes were identified.

Results: The results showed wavelet-based gene selection method presents acceptable survival prediction. Using this method, six significant genes were selected. It was indicated the expression of *GENE3359X* and *GENE3968X* decreased the survival time, whereas the expression of *GENE967X*, *GENE3980X*, *GENE3405X* and *GENE1813X* increased the survival time.

Conclusion: Wavelet-based gene selection method is a potentially useful tool for the gene selection from microarray data in the context of survival analysis.

Keywords: Survival analysis, One dimensional wavelet transform, Microarray data, DLBCL

Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma among adults with an annual incidence of 7-8 cases per 100,000 people (1, 2). The duration of survival in patients with DLBCL is very different (3). "In order to predict treatment success and explain disease heterogeneity, clinical features have been employed for prognostic purposes. But these fea-

tures have had only modest predictive performance" (3). It is estimated that high-dimensional gene expression data could noticeably enhance the predictive ability of such survival models (4).
Survival analysis is concerned with the relationship of the covariates and the time to events of interest. The typical challenge when relating survival times to gene expression data is a relatively small number of individuals compared to a large number of predictors. In addition, microarray data often possess a great deal of noise (5). From the biological aspect, only a small portion of genes have predicting power for phenotypes. If all or most of the genes is considered in the predictive model, they can induce substantial noise and thereby lead to poor predictive performance (5, 6). Thus, a crucial step towards the application of microarrays for survival prediction is the dimensional reduction from the gene expression profiles. In recent years, both feature selection and feature extraction methods have been widely used to relate censored survival time to gene expression data (6). Recent studies show that wavelet-based methods have also been used to solve the dimension reduction problem. One dimensional discrete wavelet transform (DWT) is frequently used for feature extraction in the analysis of high dimensional biomedical data (7). This method has acceptable performance in the field of feature extraction in the classification framework (8, 9). Wavelets have also used for feature selection in some of studies. Jose et al. present a wavelet-based feature selection method that assigns scores to genes for differentiating samples between two classes (10). Prabakaran et al. used the Haar wavelet power spectrum to gene selection based on expression data in the context of disease classification (11). Zhou et al. used mutual information and setting thresholds to select the most relevant features. However, due to high-dimensionality and censoring, building a predictive model for time to event is more difficult than the classification problem (12). Few studies have used wavelet transform in the area of survival analysis. Liu et al. used continuous wavelet transform combined with a genetic algorithm to select genes related to survival in colon cancer (7).

In regard to improve survival prediction, the main objective of this study was to investigate whether or not a wavelet based pre-processing method is able to remove noise from microarray data. In this study, a novel method has been introduced for gene selection based on one dimensional discrete wavelet transform in survival framework.

Materials and Methods

Data

The proposed method was applied to a DLBCL dataset (13). The dataset includes expression measurements of 4026 genes from 40 patients with DLBCL including 22 death and 18 censoring times. The survival time of the patients ranges from 1.3 to 129.9 months. The median survival time, using Kaplan–Meir approach, was 32.5 months. The expression of gene was not specified for a large part of the dataset because of missing data. Since these genes were deleted. After deleting this part of the dataset, the number of remaining genes reduced to 2042 genes. The data have been publicly published at http://llmpp.nih.gov.

Cox proportional hazards model

Cox proportional hazards regression is the most widely used method of survival analysis, which is not based on any assumptions concerning the nature or shape of the survival distribution. The Cox proportional hazards model is given by:

$$h(t,x) = h_0(t)\exp(\beta^T x)$$

Where $h_0(t)$ represents the unknown baseline hazard function and β is the unknown vector of coefficients. The unknown coefficient vector β is estimated by maximizing the partial likelihood function as follow:

$$l(\beta) = \prod_{j=1}^{k}\left(\frac{\exp(\beta^T x_j)}{\sum_{l\in R_j}\exp(\beta^T x_l)}\right)$$

Where, R_j represents all patients at risk at the jth failure time and k is the number of distinct failure times (14).

Wavelet transform

In signal processing, a transformation technique is used to transfer a data in another domain where hidden information can be extracted. Wavelets have a nice feature of local description and separation of signal characteristics, and give a tool for the analysis of transient or time-varying signal (15). Wavelet transform is an efficient time-frequency representation method which transforms a signal in time domain to a time-frequency domain. A wavelet is a set of orthonormal basis functions generated from dilation and translation of a single scaling function or father wavelet (φ), and a mother wavelet (ψ).

Wavelet transforms are classified into two different categories: the continuous wavelet transforms (CWT) and the discrete wavelet transforms (DWT). DWT is a linear operation that operates on a data vector, transforming it into a wavelet coefficient. The idea underlying DWT is to express any function $f(t) \in L^2(R)$ in terms of $\varphi(t)$ and $\psi(t)$ as follows:

$$f(t) = \sum_k c_0(k)\varphi(t-k) + \sum_k \sum_{j=1} d_j(k) 2^{\frac{-j}{2}} \psi(2^{-j}t - k)$$
$$= \sum_k c_{j_0}(k) 2^{\frac{-j_0}{2}} \varphi(2^{-j_0}t - k) + \sum_k \sum_{j=j_0} d_j(k) 2^{\frac{-j}{2}} \psi(2^{-j}t - k)$$

where $\varphi(t)$, $\psi(t)$, c_0 and d_j represent the scaling function, mother wavelet function, scaling coefficients (approximation coefficients) at scale zero, and detail coefficients at scale j, respectively. The variable k is the translation coefficient for the localization of gene expression data. The scales denote the different (low to high) scale bands. The variable symbol j_0 is scale (level) number selected (8, 15).

One-dimensional DWT decomposes a signal as a sum of wavelets at different time shifts and scales (frequencies) using DWT. For this purpose, the signal is passed through a series of high pass and low pass filters in order to analyze low as well as high frequencies in the signal as follows:

$$c_{j+1} = \sum_m h(m-2k)c_j(m)$$
$$d_{j+1} = \sum_m h_1(m-2k)c_j(m)$$

Where $h(m-2k)$ and $h_1(m-2k)$ are the low-pass filters and high-pass filters.

At each level, the high pass filter produces detail coefficients (wavelet coefficients) d_1, while the low pass filter associated with the scaling function produces approximation coefficient (scaling coefficients) c_1. Then the approximation coefficients c_1 are split into two parts by using the same algorithm and are replaced by c_2 and d_2, and so on. This decomposition process is repeated until the required level is reached. The coefficient vectors are produced by down sampling and are only half the length of the signal or the coefficient vector at the previous level.

From a viewpoint of time-frequency, the approximation coefficients are corresponding to the larger-scale low-frequency components, and the detail coefficients are corresponding to the small-scale high-frequency components. Generally, the former can be used to approximate the original signal, and the latter represents some local details of the original signal (10, 11). The decomposed components can be assembled back into the original signal without loss of information; is called reconstruction or synthesis. The mathematical manipulation, that effects synthesis is called the inverse discrete wavelet transform (IDWT).

There are different families of wavelets symlet, coiflet, daubechies and biorthogonal wavelets. They vary in various basic properties of wavelets, like compactness. Haar wavelets belonging to Daubechies wavelet family are most commonly used wavelets in database literature because they are easy to be comprehended and fast to be computed.

Model building

Firstly, the median of survival time is estimated based on Kaplan-Meier estimator, and any patient who lived longer than the median survival time (32.5 months) is placed into the class1, otherwise, into the class2. Then, the samples are grouped such that samples belonging to each class are arranged together. For investigating the effect of the order of samples within groups on the proposed method, the pre-grouped data within each class is shuffled100 times independently. The proposed

DWT-based feature selection method consists of the following steps:

1- The expression data corresponding to each gene are decomposed by the one-dimensional DWT to the specific level (second or third level in this study) using the selected mother wavelets. Then, all the detail coefficients in the lower levels are filtered out and the signal is reconstructed using just the approximation and detail coefficients in the last level.

2- An absolute value of the independent t-test statistics of the reconstructed signal is taken as the score of the gene. All the genes are ranked according to their corresponding mean t-scores and the required numbers of genes (20 genes in this study) are selected from the list.

3-Selected genes in pervious step are added to the Cox regression model and forward stepwise selection method is used for selecting the most significant genes ($\alpha < 0.05$).

4-Multiple Cox regression model including the significant genes is constructed for evaluating the performance of these selected significant genes. The predictive performance of a fitted Cox model based on selected genes is evaluated using Likelihood Ratio statistic, R^2 statistic, AIC and C index.

Note that, in the first step of proposed method, the wavelet transform is examined using db1, db3, db4, db7, sym1, sym2, coif1 and coif3 wavelets. Moreover, the numbers of selected genes in the second step are considered proportional to the sample size. The method is implemented using MATLAB r2012a software and R statistical package.

Model evaluation criteria

R^2 statistic

R^2 statistic measures the percentage of the variation in survival time that is explained by the model. Thus, when comparing models, one would prefer the model with the larger R^2 statistic (16). R^2 values are those provided by the coxph R function.

C index

Concordance, or C-statistic, is a valuable measure of model discrimination in analyses involving survival time data. Consider random pairs of patients that for each pair we inspect whether the model correctly predicts an order, e.g. a higher model score for the better result. Concordance is then the fraction of pairs for which the model is correct. A completely random prediction would have a concordance of 0.5, a perfect rule a concordance of 1(17).

AIC

Akaike information criterion (AIC) is as follows:

$$AIC = -2\,LogL + kp$$

Where the number of regression parameter in the model is p, k is some predetermined constant and L is the usual likelihood function. Models with smaller AICs are preferred (14).

Likelihood Ratio Test Statistic

The likelihood ratio test is a global goodness-of-fit *test statistic* for a *Cox* regression *model*. The test statistic for the likelihood ratio test is given as follows:

$$LR = -2lnL_R - (-2lnL_F)$$

Where R denotes the reduced (PH) model obtained when all β's are 0, and F denotes the full model. Thus, the performance is good when LR is large (14).

Results

Daubechies wavelet db-3 presents better survival prediction than the other wavelets. Therefore, the results of survival prediction model are illustrated based on db-3 for the third level of decomposition. Twenty genes have great mean absolute score, and six number of them are selected based on forward stepwise selection, using Cox regression model (P<0.05). The Predictive performance of Cox model based on the best selected genes based on discrete wavelet db3 is shown in Table 1. Table 2 shows the coefficients, hazard ratios and their 95% confidence intervals for the selected genes based on proposed method. The expression of *GENE3359X* and *GENE3968X* decreased the

survival time, whereas the expression of genes *GENE967X*, *GENE3980X*, *GENE3405X* and *GENE1813X* increased the survival time.

To further examine whether clinically relevant groups can be identified by the selected genes, the risk scores ($f(x) = \hat{\beta}'x$) estimated for the patients based on their gene expression levels of the six genes in the predictive model. We used mean of this score as a cutoff point of the risk scores and divided the patients into two groups based on whether they have positive or negative risk scores. Fig.1. shows the Kaplan-Meier curves for the two groups of patients. A significant difference was observed in overall survival between the high risk group (22 patients) and low risk group (18 patients) (P_{value}=2.96e-07). The estimated means of survival time for high and low risk patients are 24.5 and 93.9 months, respectively.

Fig. 2. indicates expression for *GENE3405X* in low and high risk patients in original and reconstructed data based on discrete wavelet db3.

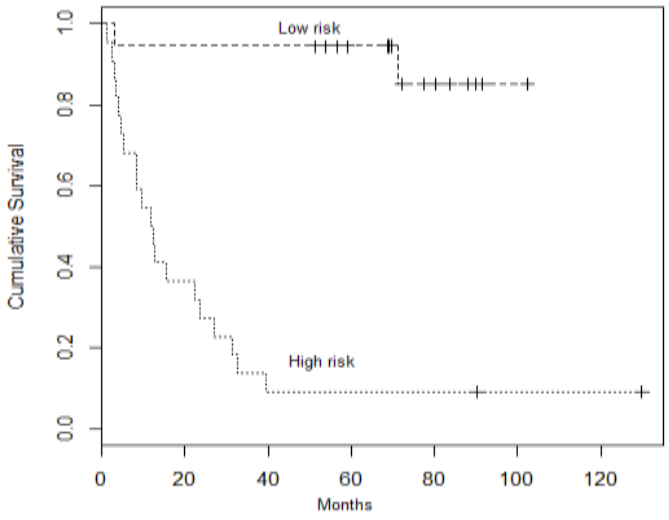

Fig.1: Kaplan-Meier plot for the high and low risk groups defined by the estimated scores

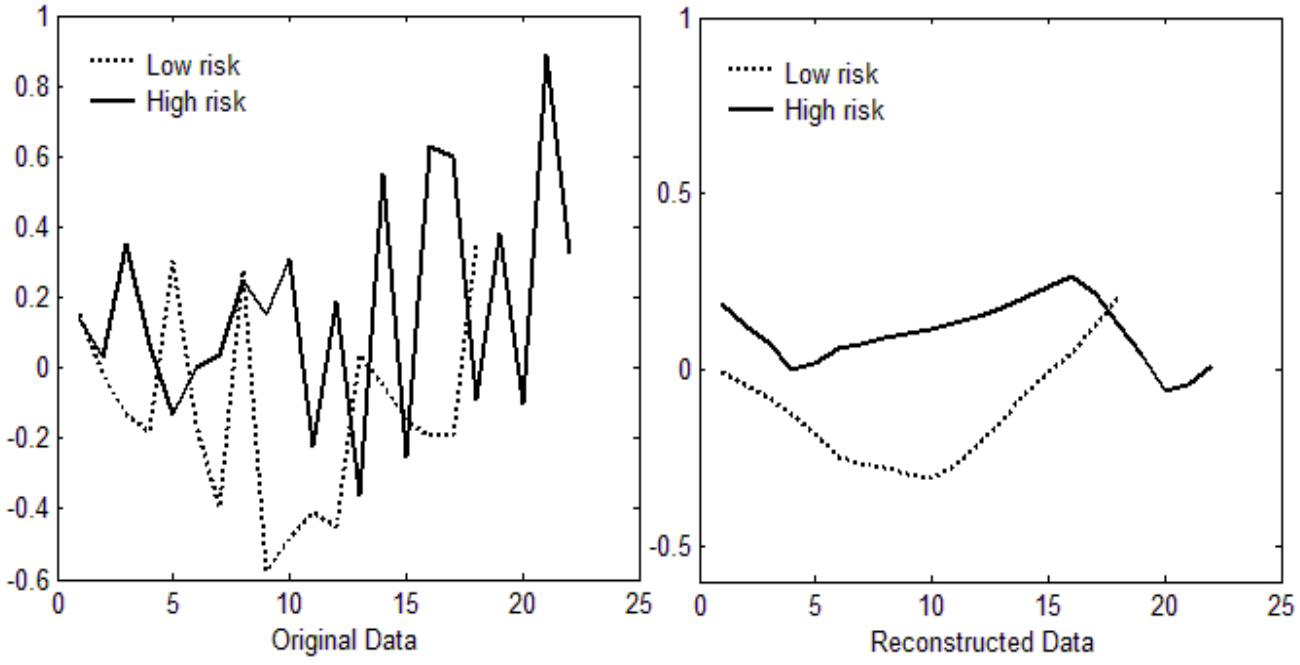

Fig.2: A comparison of the expression for *GENE3405X* in the original (without wavelet transform) and reconstructed data based on discrete wavelet db3

Table1: Model evaluation criteria for Cox model based on the discrete wavelet db3 and comparison with other studies

Method	*Preselect gene*	#Sig gene	C index	R^2	L R	AIC
Discrete wavelet db3(level 3)	*GENE3359X*,GENE3603X,GENE3391X,GENE3699X, GENE3296X,GENE377X,GENE967X*,GENE3349X, GENE3968X*,GENE3328X,GENE1146X,GENE3393X, GENE1182X,GENE1812X,GENE2638X,GENE3399X, GENE3980X*,GENE3405X*,GENE3228X,GENE1813X*,*	6	0.889	0.745	54.69	104.242
Khoshhali et al	*GENE3807X*, GENE3555X*, GENE3228X*,GENE1551X**	4	0.810	0.565	33.32	121.612
Sha et al	*GENE148X*, GENE3561X*, GENE2537X*, GENE3228X**	4	0.696	0.225	7.920	104.862

***** Significant gene in the Cox model

Table2: Estimated parameters for the selected genes using Cox regression model

Genes selected (index)	β_i	Standard Error	*P* value
GENE3359X	-2.694	0.779	0.0005
GENE967X	5.435	1.536	0.0004
GENE3968X	-2.305	0.811	0.0045
GENE3980X	4.516	1.126	6.13e-05
GENE3405X	5.840	1.392	2.74e-05
GENE1813X	1.475	0.551	0.0074

Discussion

In this study, a one-dimensional DWT-based gene selection method was proposed. The proposed method was applied to the DLBCL data of Alizadeh et al. (13). A Cox proportional hazards model based on the selected genes provided a good predictive performance for patient survival.

We found that the Daubechies wavelet db-3 presents good prediction results. In general, when wavelets are constructed as filters to remove noises in the signal, the wavelet-scaling function should have properties similar to the original signal (10). The best wavelets for selecting genes may be depending on the platforms and samples used in the microarray experiments (9).

Wavelet analysis can often condense or de-noise a signal without appreciable degradation. Normally, noise hidden in microarray profiles is obtained at acquisition. Wavelet detail coefficients have small energy and contain noise in the acquisition of microarray data (8). In wavelet transform, the main components are kept in low frequency space (approximation coefficients) and in high frequency space (detail coefficients), the extracted components hold small energy, which normally noise is hidden in. Therefore, microarray data in original data space contain noise and redundant information, to make it easier to find the significant genes, were moved the small changes existed in the high frequency part (detail coefficients) based on wavelet decomposition. If the detail coefficients in the first and second levels of the decomposition can be used to eliminate a large part of "small change," the successive approximations appear less and less "noise". Therefore approximation coefficients compress the microarray data and hold the major information on data (18).

Khoshhali et al. applied seven dimension reduction methods in order to predict survival in patients with DLBCL using gene expression dataset. Totally, their results showed that the ridge regression had best performance (19). Sha et al. proposed a Bayesian variable selection approach. They selected a set of four genes as being associated with DLBCL survival (20).

Comparison our results with the Khoshhali et al. and Sha et al. studies related to the DLBCL data set showed the Cox model based on selected genes using wavelet method has higher capability for survival prediction (Table 1). However, the methods proposed by the other studies may have their own desirable properties (19, 20). Also it was observed the two risk groups identified by the estimated risk scores show significant difference in risk of death. The results indicate that the risk

score which was built based on the proposed method can be used for predicting the risk of developing an event in future patients.
Some of the identified genes play a role as protective factors and some others as risk factors, thus, they can be used for prediction of survival time in patients with DLBCL and estimating their prognosis. Furthermore, identifying predisposing factors may be the first step for preparation and production of new treatment. However, further investigations need assess the role of these genes in promoting and prognosis of DLBCL (4, 6).
The wavelet-based gene selection method can be used to identify a set of genes for survival prediction. Expression levels of influential genes on the survival time play a role as either risk factors or preventive factors. Therefore, they can be considered as prognostic factors in secondary prevention (6). In this study gene expression data were studied as predictors. However, prediction performance of survival model may be improved by adding other covariates such as age, sex, and stage.

Conclusion

Wavelet based gene selection method is a valuable tool for identify a set of highly discriminate genes. The results demonstrated the proposed de-noising pre-processing method has potential to remove possible noise contain in microarray data. The Cox model based on selected genes by 1D wavelet method has acceptable prediction performance. The performance of proposed method exhibits the possibility of developing new tools using wavelets for the gene selection in the context of survival analysis.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study is a part of PhD thesis in Biostatistics. Therefore, the authors thank the Vic-chancellor of Research and Technology of Hamadan University of Medical Sciences, Iran, for approving the project and providing financial support. The authors declare that there is no conflict of interests.

References

1. Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD and Linet MS (2006). Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. *Blood,* 107(1): 265–276.
2. Rosenwald A, Wright G, Chan WC, Connors JM and Campo E (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large Bcell lymphoma. *N Engl J Med*, 346:1937–1947.
3. Segal MR (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7(2):268–285.
4. Bovelstad, HM, Nygard S, and Storvold, HL (2007). Predicting survival from microarray data-a comparative study. *Bioinformatics*, 23: 2080-2087.
5. Li L and Li H (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, 20(18): 3406–3412.
6. Wessel N, Wieringen V, Kuna D, Hampelb R and Boulesteix A L (2009). Survival prediction using gene expression data: A review and comparison. *Comput Stat Data An*, 53: 1590-1603.
7. Liu Y, Aickelin U, Feyereisl J and Durrant LG (2013). Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data. *Knowledge-Based Systems*, 37: 502–514.
8. Liu Y (2012). Dimensionality reduction and main component extraction of mass spectrometry cancer data. *Knowledge-Based Systems*, 26: 207–215.
9. Nanni L and Lumini A (2011). Wavelet selection for disease classification by DNA microarray data. *Expert Syst Appl*, 38: 990-995.

10. Jose A, Mugler D and Duan ZH (2009). A gene selection method for classifying cancer samples using 1D discrete wavelet transform. *Int J Comput Biol Drug Des*, 2(4): 398-411.
11. Prabakaran S, Rajendra S and Shekhar V (2006). Feature selection using Haar wavelet power spectrum. *BMC Bioinformatics*, 7: 432-443.
12. Zhou X, Wang X and Dougherty ER (2004). Nonlinear probit gene classification using mutual information and wavelet based feature extraction. *J Biol Syst*, 12(3): 371-386.
13. Alizadeh A, Eisen MB, Davis RE, Ma C, Losses IS and Resenwald A (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature,* 403: 503-511.
14. Klein JP and Moeschberger ML (2003). Survival Analysis. Techniques for Censored and Truncated Data. New York, *Springer-Verlag.*
15. Liu Y (2009). Wavelet feature extraction for high-dimensional microarray data. *Neurocomputing*, 72: 985–990.
16. Heller G (2012). A measure of explained risk in the proportional hazards model. *Biostatistics*, 13(2): 315–325.
17. Pencina MJ and Agostino RB (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statist Med*, 23: 2109–2123.
18. Li X, Li J and Yao X (2007). A wavelet-based preprosecing analysis approach in mass spectrometry. *Comput Biol Med,* 37: 509-516.
19. Khoshhali M, Mahjub H, Saidijam M, Poorolajal J and Soltanian AR (2012). Predicting the survival time for diffuse large B-cell lymphoma using microarray data. *J Mol Genet Med,* 6: 287-292.
20. Sha N, Tadesse MG and Vannucci M (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics,* 22(18): 2262–2268.