Wavelet-based gene selection method for survival prediction in diffuse large B-cell lymphomas patients

Maryam Farhadian

Department of Biostatistics and Epidemiology,
School of Public Health,
Hamadan University of Medical Sciences,
Hamadan, Iran
Email: m.farhadian@umsha.ac.ir

Hossein Mahjub*

Research Center for Health Sciences
Department of Biostatistics and Epidemiology,
School of Public Health,
Hamadan University of Medical Sciences,
Hamadan, Iran
Email: mahjub@umsha.ac.ir
*Corresponding author

Abbas Moghimbeigi

Modeling of Noncommunicable Disease Research Center,
Department of Biostatistics and Epidemiology,
School of Public Health,
Hamadan University of Medical Sciences,
Hamadan, Iran
Email: moghimbeigi@umsha.ac.ir

Paulo J.G. Lisboa

School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Liverpool, UK
Email: P.J.Lisboa@ljmu.ac.uk

Jalal Poorolajal

Modeling of Noncommunicable Diseases Research Center,
Department of Biostatistics and Epidemiology,
School of Public Health,
Hamadan University of Medical Sciences,
Hamadan, Iran
Email: poorolajal@umsha.ac.ir

Muharram Mansoorizadeh

Department of Computer Engineering,
Faculty of Engineering,
Bu-Ali Sina University,
Hamadan, Iran
Email: mansoorm@basu.ac.ir

Abstract: Microarray technology allows simultaneous measurements of expression levels for thousands of genes. An important aspect of microarray studies includes the prediction of patient survival based on their gene expression profile. This naturally calls for the use of a dimension reduction procedure together with the survival prediction model. In this study, a new method based on wavelet transform for survival-relevant gene selection is presented. Cox proportional hazard model is typically used to build prediction model for patients' survival using the selected genes. The prediction model will be evaluated with the R^2, concordance index, likelihood ratio statistic and Akaike information criteria. The results proved that good performance of survival prediction is achieved based on the selected genes. The results suggested the possibility of developing more advanced tools based on wavelets for gene selection from microarray data sets in the context of survival analysis.

Keywords: survival analysis; gene selection; one-dimensional wavelet transform; microarray data; DLBCL; diffuse large B-cell lymphoma.

Reference to this paper should be made as follows: Farhadian, M., Mahjub, H., Moghimbeigi, A., Lisboa, P.J.G., Poorolajal, J. and Mansoorizadeh, M. (2015) 'Wavelet-based gene selection method for survival prediction in diffuse large B-cell lymphomas patients', *Int. J. Data Mining and Bioinformatics*, Vol. 13, No. 2, pp.197–210.

Biographical notes: Maryam Farhadian is PhD in Biostatistics from Hamadan University of Medical Sciences, Hamadan, Iran. Her current research interests include multivariate analysis, survival analysis in high dimensional data, bioinformatics and machine learning.

Hossein Mahjub is Professor of Biostatistics, Hamadan University of Medical Sciences, Hamadan, Iran. His current research interests include multivariate analysis, survival analysis and longitudinal analysis.

Abbas Moghimbeigi is Associated Professor of Biostatistics, Hamadan University of Medical Sciences, Hamadan, Iran. His current research interests include categorical data analysis, survival analysis, multilevel analysis, sampling design and QTL mapping.

Paulo J.G. Lisboa is Professor in Industrial Mathematics in School of Computing and Mathematical Sciences, Liverpool John Moores University, UK. His current research interests include pattern recognition methods for decision support and exploratory data analysis to medicine, business and engineering, Survival modelling following tumour excision and brain tumour grading.

Jalal Poorolajal is Associate Professor of Epidemiology, Hamadan University of Medical Sciences, Hamadan, Iran. His current research interests include epidemiology.

Muharram Mansoorizadeh is Assistant Professor of computer engineering of Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamadan, Iran. His current research interests include neural networks.

1 Introduction

Diffuse Large B-Cell Lymphoma (DLBCL) is a cancer of B cells, a type of white blood cell responsible for producing antibodies. It is the most common type of non-Hodgkin lymphoma among adults with an annual incidence of seven to eight cases per 100,000 people (Morton et al., 2006; Smith et al., 2011). The duration of survival in patients with DLBCL is very different (Segal, 2006). In order to predict treatment success and explain disease heterogeneity, clinical features have been employed for prognostic purposes. But these features have had only modest predictive performance (Segal, 2006). It is estimated that high-dimensional gene expression data could noticeably enhance the predictive ability of such survival models (Bovelstad et al., 2007). RNA-Seq is a novel technology that has recently emerged to measure gene expression (Zwiener et al., 2014). Rapaport et al. (2013) provided a broad review on current research in differential gene expression analysis methods for RNA-seq data. Recently, application of RNA-seq to DLBCL has revealed numerous genes that are recurrent targets of somatic point mutation in this disease (Morin et al., 2013). Microarray studies which examine thousands of genes simultaneously are traditionally performed, to find genes that are used for disease subtypes classification. However, a new emphasis is on finding genes predictive of survival.

Survival analysis is concerned with the relationship of the covariates and the time to events of interest. However, due to the large variability in survival times between cancer patients and the amount of genes on the microarrays unrelated to outcome, developing accurate prediction models remains a challenge. 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 (Li and Li, 2004). From the biological aspect, only a small portion of genes has predicting power for phenotypes. If all or most of the genes are considered in the predictive model, they can induce substantial noise and thereby lead to poor predictive performance (Li and Li, 2004). Thus, a crucial step towards the application of microarrays for survival prediction is the dimensional reduction from the gene expression profiles.

In this regard, both feature selection and feature extraction methods have been widely used to relate censored survival time to gene expression data (Wessel et al., 2009; Du et al., 2013). Rosenwald et al. (2002) described a feature selection approach for identifying genes related to survival time that fits Cox Proportional Hazards (PH) models to each gene and selects those that pass a threshold for significance. Li and Li (2004) proposed a dimension reduction strategy which combines principle component analysis and sliced inverse regression to identify useful linear combinations of genes. Li and Luan (2005) developed a penalised estimation procedure for the Cox model using kernels, under the assumption that the covariate effects are smooth functions of gene expression levels. Nguyen and Rojo (2009) considered the problem of relating survival time to gene expression by reduction of the dimensionality via Partial Least Squares (PLS) method. Liu and Jiang (2009) developed Cox proportional hazard model with L_p penalty method for simultaneous gene selection and survival prediction. Bonato et al. (2011) proposed Bayesian ensemble methods for survival prediction using gene expression data. Yang and Zou (2013) have used an elastic net approach for variable selection under the Cox proportional hazard model.

Different studies show that machine learning methods such as neural networks, Bayesian networks and support vector machines are used to improve the survival prediction models (Biganzoli et al., 2002; Lisboa et al., 2008; Van Belle et al., 2010). Wavelet-based methods have also been used to solve the dimension reduction problem. The primary intuition for applying wavelets in the case of gene expression is that genes are often co-expressed in groups (Tokuyasu et al., 2003). It would be useful to treat the group as a single variable, akin to the motivation behind methods such as principal component analysis. One-dimensional Discrete Wavelet Transform (DWT) is frequently used for feature extraction in the analysis of high-dimensional biomedical data (Liu et al., 2013). Studies showed that this method has acceptable performance in the field of feature extraction in the classification framework (Liu, 2012; Nanni and Lumini, 2011). Li et al. (2007) used a wavelet-based pre-processing method for feature extraction of high-dimensional mass spectrometry data. Wavelets have also been used for feature selection in some studies. Jose et al. (2009) present a wavelet-based feature selection method that assigns scores to genes for differentiating samples between two classes. Prabakaran et al. (2006) used the Haar wavelet power spectrum to gene selection based on expression data in the context of disease classification. Li et al. (2006) took the one-dimensional DWT of the gene expression levels and applied an algorithm based on setting a threshold to choose the most important wavelet coefficients. Zhou et al. (2004) 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. Few studies have used wavelet transform in the area of survival analysis. Liu et al. (2013) used continuous wavelet transform combined with a genetic algorithm to select genes related to survival in colon cancer.

In regard to improving 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 DWT in survival framework.

2 Method

2.1 Data

The proposed method was applied to a set of gene expression data with survival information on the DLBCL data set (Rosenwald et al., 2002).

The data set includes expression measurements of 7399 genes from 240 patients with DLBCL, with their survival times. A total of 138 (57.5%) deaths have been observed with the median death time of 2.8 years. A detailed description of DLBCL data can be found in the original publication (Rosenwald et al., 2002). The missing data were imputed using the mean expression level of the nearest eight genes (Li and Luan, 2005). The data have been published at http://llmpp.nih.gov/lymphoma/data.shtml.

2.2 Cox proportional hazard model

Cox proportional hazard 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 hazard model is given by

$$h(t,x)=h_0(t)\exp\left(\beta^T x\right) \tag{1}$$

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 maximising the partial likelihood function as follows:

$$l(\beta)=\prod_{j=1}^{k}\left(\frac{\exp\left(\beta^T x_j\right)}{\sum_{l\in R_j}\exp\left(\beta^T x_l\right)}\right) \tag{2}$$

where R_j represents all the patients at risk at the jth failure time and k is the number of distinct failure times (Klein and Moeschberger, 2003).

2.3 Wavelet transform

A wavelet is a 'small wave' which has its energy concentrated in time. A wavelet system is generated from a single scaling function or wavelet by simple scaling and translation. In signal processing, a transformation technique is used to transfer data in one domain into another where hidden information can be extracted. Wavelets have nice features of local description and separation of signal characteristics, and give a tool for the analysis of transient or time-varying signal (Liu, 2012).

Wavelet transforms are classified into two different categories: Continuous Wavelet Transforms (CWT) and DWT. According to DWT, a time-varying function $f(t) \in L^2(R)$ can be expressed in terms of $\varphi(t)$ and $\psi(t)$ as follows:

$$\begin{aligned} f(t) &= \sum_k c_0(k)\varphi(t-k)+\sum_k\sum_{j=1} d_j(k)2^{\frac{-j}{2}}\psi\left(2^{-j}t-k\right) \\ &= \sum_k c_{j_0}(k)2^{\frac{-j_0}{2}}\varphi\left(2^{-j_0}t-k\right)+\sum_k\sum_{j=j_0} d_j(k)2^{\frac{-j}{2}}\psi\left(2^{-j}t-k\right) \end{aligned} \tag{3}$$

where $\varphi(t)$, $\psi(t)$, c_0 and d_j represent the scaling function, wavelet function, scaling coefficients (approximation coefficients) at scale 0 and detail coefficients at scale j, respectively. The variable k is the translation coefficient for the localisation of gene expression data. The scales denote the different (low to high) scale bands. The variable symbol j_0 is the scale (level) number selected.

The wavelet analysis involves two compounds: approximations and details. For one-dimensional wavelet decomposition, starting from the signal, the first step produces two sets of coefficients: approximation coefficients (scaling coefficients) c_1 and detail coefficients (wavelet coefficients) d_1. These coefficients are computed by convolving signal with the low pass filter for approximation, and with the high pass filter for detail.

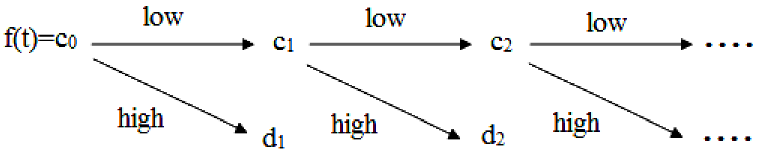
The convolved coefficients are down-sampled by keeping the even indexed elements. Then the approximation coefficients c_1 are split into two parts by using the same algorithm and are replaced by c_2 and d_2 and so on as follows:

$$c_{j+1} = \sum\nolimits_m h(m-2k)c_j(m) \tag{4}$$

$$d_{j+1} = \sum\nolimits_m h_1(m-2k)c_j(m) \tag{5}$$

where $h(m - 2k)$ and $h_1(m - 2k)$ are the low pass filters and high pass filters. 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. Conversely, approximations and details are constructed inverting the decomposing step by inserting zeros and convolving the approximation and detail coefficients with the reconstruction filters. The whole process of obtaining the wavelet transform of $f(t)$ using the pyramid algorithm is shown in Figure1.

Figure 1 The one-dimensional wavelet decomposition process



The main advantage of the wavelet transform is that each basis function is localised jointly in both the time and the frequency domains. 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 (Nanni and Lumini, 2011).

The presence of noise is fairly common in microarray processing, which makes the identification of the transient change more complicated. Wavelets have an important application in signal denoising (Li et al., 2007). After wavelet decomposition, the high-frequency sub-bands contain most of the noise information and little signal information. The ability to selectively reconstruct the signal using altered or chosen coefficients makes the wavelet analysis extremely useful for dealing with noise in the signals (Li et al., 2007). There are different families of wavelets: symlet (sym), coiflet (coif), daubechies (db) and biorthogonal (bior). They vary in various basic properties of wavelets, like compactness.

2.4 Model building

Firstly, the median of survival time is estimated based on the Kaplan–Meier estimator, and any patient who lived longer than the median survival time (3.9 years) is placed into class 1, otherwise into class 2. 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 are shuffled 100 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 (24 or 48 in this study) are selected from the list.

3 Selected genes in previous 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 A multiple Cox regression model including the significant genesis 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 the likelihood ratio statistic, R^2 statistic, AIC and *C* index.

Note that, in the first step of the 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.

2.5 Model evaluation criteria

2.5.1 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 (Heller, 2012). R^2 values are those provided by the coxph() R function.

2.5.2 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 and a perfect rule a concordance of 1 (Pencina and Agostino, 2004).

2.5.3 AIC

Akaike Information Criterion (AIC) is as follows:

$$AIC = -2 \log L + 2p \qquad (6)$$

where p is the number of regression parameter in the model and L is the usual likelihood function. Models with smaller AICs are preferred (Klein and Moeschberger, 2003).

2.5.4 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 as follows:

$$LR = -2 \ln L_R - \left(-2 \ln L_F\right) \qquad (7)$$

where R denotes the reduced (PH) model obtained when all β are 0 and F denotes the full model. Thus, the performance is good when LR is large (Klein and Moeschberger, 2003).

3 Results

The results showed that 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. Various numbers of genes, including 24 and 48 genes, were tested and similar predictive results were observed. Model evaluation criteria for using 24 and 48 genes with the greater mean absolute *t*-score is shown in Table 1. Forty-eight genes have great mean absolute score, and seven of them are selected based on forward stepwise selection, using Cox regression model ($p < 0.05$). Table 2 shows the coefficients, hazard ratios and their 95% confidence intervals for the selected genes. The expression of genes 3825 (known as AA714513), 5301 (known as AA181040) and 4131 (known as AA805575) decreased the survival time, whereas the expression of genes 4540 (known as AA487453), 1825 (known as AA262133), 7070 (known as LC_33732) and 5614 (known asNM_014456) increased the survival time. The Gene Bank ID and a description of seven genes selected by our proposed method are shown in Table 3. Also, the results of other studies based on the same data set are shown in Table 4. It can be seen that the proposed method has higher capability for prediction than the methods used in other studies (Segal, 2006; Gui and Li, 2004; Engler and Li, 2009; Lai et al., 2013).

Table 1 Model evaluation criteria in the original and transformed data based on the discrete wavelet db3

Method	*level*	*#Pre-select gene*	*#Significant gene*	*C index*	*R^2*	*Likelihood ratio*	*AIC*
Discrete wavelet db3	2	24	5	0.71	0.25	69.09	1320.46
	2	48	7	0.73	0.33	94.92	1298.63
	3	24	6	0.71	0.26	71.39	1320.16
	3	48	7	0.74	0.34	94.54	1293.69

Table 2 Estimated parameters for the selected genes using Cox regression model

Genes selected (index)	β_i	*Standard error*	P_{value}	*exp*(β_i)	*95% CI for exp* (β_i)	
					Lower	*Upper*
4540	0.300	0.092	0.001	1.350	1.126	1.618
3825	–0.359	0.086	2.99e–05	0.698	0.589	0.826
1825	1.397	0.258	6.24e–08	4.046	2.439	6.713
5301	–0.245	0.108	0.023	0.782	0.633	0.967
7070	0.622	0.154	5.23e–05	1.862	1.378	2.517
5614	0.265	0.121	0.027	1.304	1.029	1.652
4131	–0.121	0.046	0.008	0.886	0.809	0.970

Table 3 Gene bank ID and descriptions of the seven genes selected by the proposed method

Index	*Gene bank ID*	*Signature*	*Description*
4540	AA487453	–	GRO2 oncogene
3825	AA714513	MHC	Major histocompatibility complex, class II, DR beta 5
1825	AA262133	PF	septin 1
5301	AA181040	Lymph	secreted protein, acidic, cysteine-rich (osteonectin)
7070	LC_33732	–	–
5614	NM_014456	–	Programmed cell death 4 (neoplastic transformation inhibitor)-H731 nuclear antigen: PDCD4
4131	AA805575	Germ	ESTs, weakly similar to A47224-Thyroxine-binding globulin precursor [H.sapiens]

Notes: Germ: germinal centre B-cell signature; MHC: MHC class II signature; Lymph: lymph-node signature; PF: proliferation signature

Table 4 Comparison results with the other studies

Method	*# Gene*	*Genes selected (index)*	*C index*	R^2	*Likelihood ratio*	*AIC*
Segal (2006)	4	6720,5054,3811,5342	0.668	0.172	45.34	1342.21
Gui and Li (2004)	10	3243,3799,3825,4056,4131,5027,5054,5342,5408	0.702	0.238	65.11	1334.44
Engler and Li (2009)	7	80, 992, 1456, 3216, 5442, 6402, 6909	0.657	0.168	44.12	1349.42
Lai et al. (2013)	8(sig)	2182,3785,3825,3921,3951,4138,4552,4728	0.696	0.254	70.42	1325.13

To further examine whether clinically relevant groups can be identified by the selected genes, the risk scores ($f(x) = \hat{\beta}'x$)) are estimated for the patients based on the gene expression levels of the seven genes in the predictive model. We used zero as a cut-off point of the risk scores and divided the patients into two groups based on whether they have positive or negative risk scores. Figure 2 shows the Kaplan–Meier curves for the two groups of patients. A significant difference (p = 1.94e–14) was observed in the overall survival between the high-risk group (120 patients) and low-risk group (120 patients). The estimated medians of survival time for high- and low-risk patients are 1.2 and 16.9 years, respectively. The results indicate that the model which was built based on the proposed method can be used for predicting the risk of developing an event in future patients.

Figure 2 Kaplan–Meier curves for the high and low risk groups defined by the estimated scores

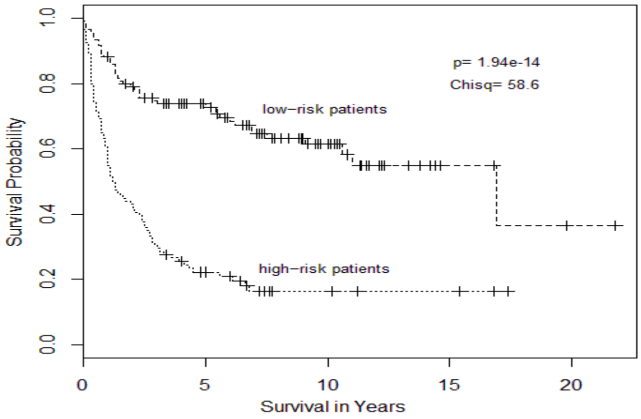
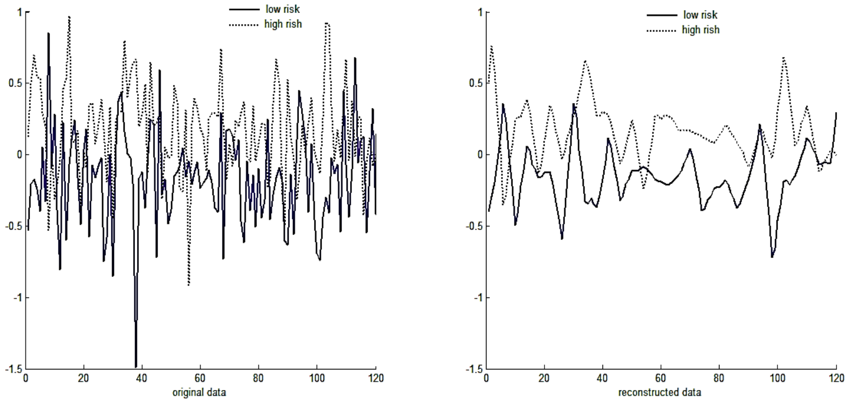


Figure 3 indicates expression for gene 1825 in low- and high-risk patients in the original and the reconstructed data based on discrete wavelet db3.

Figure 3 A comparison of the expression for gene 1825(AA262133) in the original (without wavelet transform) and reconstructed data based on discrete wavelet db3



4 Discussion

In this study, a one-dimensional discrete wavelet-transform-based gene selection method was proposed. The proposed method was applied to the DLBCL data of Rosenwald et al. (2002). A Cox proportional hazard 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 (Jose et al., 2009). The best wavelets for selecting genes may be depending on the platforms and samples used in the microarray experiments (Nanni and Lumini, 2011).

Comparison of our results with the other studies related to the DLBCL data set showed the Cox model based on selected genes using wavelet method has the best prediction performance. However, the methods proposed by the other studies may have their own desirable properties.

Rosenwald et al. (2002) used hierarchical clustering to identify four signature groups whose expressions were correlated with the survival times. The four groups are MHC class II (MHC), Proliferation signature (PF), Lymph node signature (Lymph) and Germinal centre B-cell signature (Germ). The genes in these groups were shown to be related to the risk of death for the DLBCL patients.

In the current study, seven significant genes are identified. Four of the selected genes belong to the three groups of gene expression signature defined by Rosenwald et al. (2002). These genes included genes number 4131, 3825, 5301 and 1825, which belong to Germinal-centre B-cell signature, MHC class II signature, Lymph node signature and Proliferation signature group, respectively. In this study, the results showed that the estimated coefficients for the genes number 3825, 4131 and 5301 were all negative, indicating that high expression levels of these genes reduce the risk of death among the patients with DLBCL. This agrees with what Rosenwald et al. (2002) has found.

The genes number 1825 and 4131 are identified as high ranked genes by Li and Luan (2005) using the partial likelihood-based scores. Also, genes number 3825 and 4131 are identified as interesting candidate genes by Gui and Li (2004) using the LARS-Lasso method. Moreover, gene 3825 is identified as interesting candidate by Lai et al. (2013) using the lasso regression method.

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 in estimating their prognosis (Lin et al., 2013). Furthermore, identifying predisposing factors may be the first step for preparation and production of new treatment. However, further investigations need to assess the role of these genes in promoting and prognosis of DLBCL (Bovelstad et al., 2007; Wessel et al., 2009). It was observed that the two risk groups identified by the estimated risk scores show more significant difference in risk of death than the groups defined by Gui and Li (2004) who have used the LARS-Lasso method (*p*-value of 1.94e–14 for current study vs. p = 0.000379). In order to assess the effect of the number of genes on the evaluation criteria, further reviews were done. The results showed that the model with the top four selected genes based on the proposed method also came up with better performance than Segal's model.

However, some studies have shown that there are many combinations of genes that seem to be truly related to cancer outcome, but are actually random (Venet et al., 2011). To check whether the selected genes are meaningful in describing B-cell lymphoma outcome and free from false discovery, 100 sets, each consisting of seven genes, were generated randomly from 7399 genes. The Cox model was fitted to each set and the model evaluation criteria were calculated. The results showed that the performance of all the randomly generated data sets was poorer than that of the selected genes based on the model evaluation criteria. So we can say that the final result is meaningful in describing B-cell lymphoma outcome and is free from false discovery.

Wavelet analysis can often condense or denoise 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 (Liu, 2009; Liu, 2012). In wavelet transform, the main components are kept in low-frequency space (approximation coefficients) and in high-frequency space (detail coefficients), and the extracted components hold small energy, which normally noise is hidden. Therefore, microarray data in the original data space contain noise and redundant information; to make it easier to find, the significant genes were moved to where the small changes existed in the high-frequency part (detail coefficients) based on wavelet decomposition. The motivation is that coefficient details are more likely due to noise and approximation coefficient due to important signal features. If the detail coefficients in the first and second levels of the decomposition can be used to eliminate a large part of the 'small change', the successive approximations appear less and less 'noisy'. Therefore, approximation coefficients compress the microarray data and hold the major information on data (Li et al., 2007).

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 the role of either risk factors or preventive factors. Hence, determining the expression levels of such genes might be helpful for primary prevention programmes. On the other hand, the expression levels of these genes could influence the survival time. Therefore, they can be considered as prognostic factors in secondary prevention (Wessel et al., 2009). 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.

5 Conclusion

The wavelet-based gene selection method is a valuable tool for identifying a set of highly discriminate genes. The results demonstrate that the proposed denoising pre-processing method has the potential to remove possible noise contained in microarray data. The Cox model based on selected genes by a one-dimensional wavelet method has acceptable prediction performance compared with the other studies. The performance of the proposed method exhibits the possibility of developing more advanced tools using wavelets for gene selection in the context of survival analysis.

Acknowledgements

This study is a part of a PhD thesis in biostatistics. Therefore, the authors thank the Vice-chancellor of Research and Technology of Hamadan University of Medical Sciences, Iran, for approving the project and providing financial support.

References

Biganzoli, E., Boracchi, P. and Marubini, E. (2002) 'A general framework for neural network models on censored survival data', *Neural Networks*, Vol. 15, pp.209–218.

Bonato, V., Baladandayuthapani, V., Broom, B.M., Sulman, E.P., Aldape, K.D. and Do, K.A. (2011) 'Bayesian ensemble methods for survival prediction in gene expression data', *Bioinformatics*, Vol. 27, No. 3, pp.359–367.

Bovelstad, H.M., Nygard, S. and Storvold, H.L. (2007) 'Predicting survival from microarray data: a comparative study', *Bioinformatics*, Vol. 23, pp.2080–2087.

Du, W., Sun, Y., Wang, Y., Cao, Z., Zhang, C. and Liang, Y. (2013) 'A novel multi-stage feature selection method for microarray expression data analysis', *International Journal of Data Mining and Bioinformatics*, Vol. 7, No. 1, pp.58–77.

Engler, D. and Li, Y. (2009) 'Survival analysis with high-dimensional covariates: an application in microarray studies', *Statistical Applications in Genetics and Molecular Biology*, Vol. 8, No. 1, Article 14.

Gui, J. and Li, H. (2004) 'Partial Cox regression analysis for high dimensional microarray gene expression data', *Bioinformatics*, Vol. 20, pp.208–215.

Heller, G. (2012) 'A measure of explained risk in the proportional hazards model', *Biostatistics*, Vol. 13, No. 2, pp.315–325.

Jose, A., Mugler, D. and Duan, Z.H. (2009) 'A gene selection method for classifying cancer samples using 1D discrete wavelet transform', *International Journal of Computational Biology and Drug Design*, Vol. 2, No. 4, pp.398–411.

Klein, J.P. and Moeschberger, M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Lai, Y., Hayashida, M. and Akutsu, T. (2013) 'Survival analysis by penalized regression and matrix factorization', *The Scientific World Journal*, Vol. 2013, Article ID 632030.

Li, H. and Luan, Y. (2005) 'Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data', *Bioinformatics*, Vol. 21, pp.2403–2409.

Li, L. and Li, H. (2004) 'Dimension reduction methods for microarrays with application to censored survival data', *Bioinformatics*, Vol. 20, No. 18, pp.3406–3412.

Li, S., Liao, C. and Kwok, J.T. (2006) 'Wavelet-based feature extraction for microarray data classification', *IEEE International Joint Conference on Neural Networks*, 16–21 July, Vancouver, BC, pp.5028–5033.

Li, X., Li, J. and Yao, X. (2007) 'A wavelet-based pre-processing analysis approach in mass spectrometry', *Computers in Biology and Medicine*, Vol. 37, pp.509–516.

Lin, D., Tilahun, A., Abrahantes, J.C., Shkedy, Z., Molenberghs, G., Gohlmann, H.W.H., Talloen, W. and Bijnens, L. (2013) 'Comparison of methods for the selection of genomic biomarkers', *International Journal of Data Mining and Bioinformatics*, Vol. 8, No. 1, pp.24–41.

Lisboa, P.J.G., Etchells, T.A., Jarman, I.H., Hane Aung, M.S., Chabaud, S., Bachelot, T. et al. (2008) 'Time-to-event analysis with artificial neural networks: an integrated analytical and rule-based study for breast cancer', *Neural Networks***,** Vol. 21, pp.414–426.

Liu, Y. (2009) 'Wavelet feature extraction for high-dimensional microarray data', *Neurocomputing*, Vol. 72, pp.985–990.

Liu, Y. (2012) 'Dimensionality reduction and main component extraction of mass spectrometry cancer data', *Knowledge-Based Systems*, Vol. 26, pp.207–215.

Liu, Z. and Jiang, F. (2009) 'Gene identification and survival prediction with Lp Cox regression and novel similarity measure', *International Journal of Data Mining and Bioinformatics*, Vol. 3, No. 4, pp.398–408.

Liu, Y., Aickelin, U., Feyereisl, J. and Durrant, L.G. (2013) 'Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data', *Knowledge-Based* Systems, Vol. 37, pp.502–514.

Morin, R.D., Mungall, K., Pleasance, E., Mungall, A.J., Goya, R., Huff, R.D., Scott, D.W., Ding, J., Roth, A. and Chiu, R. (2013) 'Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing', *Blood*, Vol. 15, pp.1256–1265.

Morton, L.M., Wang, S.S., Devesa, S.S., Hartge, P., Weisenburger, D.D. and Linet, M.S. (2006) 'Lymphoma incidence patterns by WHO subtype in the United States, 1992–2001', Blood, Vol. 107, No. 1, pp.265–276.

Nanni, L. and Lumini, A. (2011) 'Wavelet selection for disease classification by DNA microarray data', *Expert Systems with Applications*, Vol. 38, pp.990–995.

Nguyen, D.V. and Rojo, J. (2009) 'Dimension reduction of microarray data in the presence of a censored survival response: a simulation study', *Statistical Applications in Genetics and Molecular Biology*, Vol. 8, No. 1.

Pencina, M.J. and Agostino, R.B. (2004) 'Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation', *Statistics in Medicine*, Vol. 23, pp.2109–2123.

Prabakaran, S., Rajendra, S. and Shekhar, V. (2006) 'Feature selection using Haar wavelet power spectrum', *BMC Bioinformatics*, Vol. 7, pp.432–443.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D. and Betel, D. (2013) 'Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data', *Genome Biology*, Vol. 14, p.R95.

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M. and Campo, E. (2002) 'The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma', *New England Journal of Medicine*, Vol. 346, pp.1937–1947.

Segal, M.R. (2006) 'Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited', *Biostatistics*, Vol. 7, No. 2, pp.268–285.

Smith, A.D., Howell, R., Patmore, R., Jack, A. and Roman, E. (2011) 'Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network', *British Journal of Cancer*, Vol. 105, No. 11, pp.1684–1692.

Tokuyasu, T.A., Albertson. D., Pinkel, D. and Jain, A. (2003) 'Wavelet transforms for the analysis of microarray experiments', *Proceedings of the IEEE Bioinformatics Conference*, 11–14 August, Stanford, CA, pp.429–430.

Van Belle, V., Pelckmans, K., Suykens, J.A.K. and Van Huffel, S. (2010) 'Additive survival least-squares support vector machines', *Statistics in Medicine*, Vol.29, pp.296–308.

Venet, D., Dumont, J.E. and Detours, V. (2011) 'Most random gene expression signatures are significantly associated with breast cancer outcome', *PLoS Computational Biology*, Vol. 7, No. 10, p.e1002240.

Wessel, N., Wieringen, V., Kuna, D., Hampelb, R. and Boulesteix, A.L. (2009) 'Survival prediction using gene expression data: a review and comparison', *Computational Statistics and Data Analysis*, Vol. 53, pp.1590–1603.

Yang, Y. and Zou, H. (2013) 'A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions', *Statistics and Its Interface*, Vol. 6, pp.167–173.

Zhou, X., Wang, X. and Dougherty, E.R. (2004) 'Nonlinear probit gene classification using mutual information and wavelet based feature extraction', *Journal of Biological Systems*, Vol. 12, No. 3, pp.371–386.

Zwiener, I., Frisch, B. and Binder, H. (2014) 'Transforming RNA-Seq data to improve the performance of prognostic gene signatures', *PLOS ONE*, Vol. 9, p.e85150.