A gene selection method for classifying cancer samples using 1D discrete wavelet transform

Adarsh Jose and Dale Mugler

Department of Biomedical Engineering,
University of Akron, Akron, OH 44236, USA
E-mail: adarshjos@gmail.com
E-mail: dmugler@uakron.edu

Zhong-Hui Duan*

Integrated Bioscience Program,
Department of Computer Science,
University of Akron, Akron, OH 44236, USA
E-mail: duan@uakron.edu
*Corresponding author

Abstract: Selecting a set of discriminant genes for biological samples is an important task for designing highly efficient classifiers using DNA microarray data. The wavelet transform is a very common tool in signal-processing applications, but its potential in the analysis of microarray gene expression data is yet to be explored fully. In this paper, we present a wavelet-based feature selection method that assigns scores to genes for differentiating samples between two classes. The gene expression signal is decomposed using several levels of the wavelet transform. The genes with the highest scores are selected to form a feature set for sample classification. In this study, the feature sets were coupled with *k*-nearest neighbour (*k*NN) classifiers. The classification accuracies were assessed using several real data sets. Their performances were compared with several commonly used feature selection methods. The results demonstrate that 1D wavelet analysis is a valuable tool for studying gene expression patterns.

Keywords: microarray gene expression profiles; feature selection; wavelet applications; biological sample classification.

Reference to this paper should be made as follows: Jose, A., Mugler, D. and Duan, Z-H. (2009) 'A gene selection method for classifying cancer samples using 1D discrete wavelet transform', *Int. J. Computational Biology and Drug Design*, Vol. 2, No. 4, pp.398–411.

Biographical notes: Adarsh Jose earned his Bachelor's Degree in Science, Electronics and Biomedical Engineering from the Cochin University of Science and Technology in India and an MS in Biomedical Engineering from the University of Akron, Akron, Ohio. He is currently a PhD student of the Bioinformatics and Computational Biology Graduate Programme at the Iowa State University. His research interests include bioinformatics and biomedical signal processing.

Dale Mugler earned his PhD in Mathematics from the Northwestern University. He is the Dean of the Honors College at the University of Akron.

He is also a Professor of Applied Mathematics and Biomedical Engineering at the University of Akron. His research interests include wavelet theory and applications, computational methods in signal processing, and numerical algorithms and scientific computing.

Zhong-Hui Duan received her MS in Computer Science and a PhD in Applied Mathematics from the Florida State University. She is currently an Associate Professor of Computer Science at the University of Akron and a faculty member of the Integrated Bioscience PhD Programme at the University of Akron. Her research interests include algorithms, bioinformatics, data mining and computer simulations.

1 Introduction

One of the challenges in cancer biology is the early detection of malignant tumour and identification of distinct tumour sub-types. DNA microarray technology has been used to identify differentially expressed genes and characterise their expression patterns for many solid and haematological malignancies, such as colon cancer, breast cancer, prostate cancer, leukaemia and lymphoma (Ramaswamy et al., 2003). The expression patterns of the identified genes for normal vs. cancerous samples, samples from different stages of malignancy and different sub-types of tumour allow the application of systematic and unbiased pattern recognition techniques for cancer classification. Furthermore, the identification of these genes provides biologists insights into understanding the molecular mechanism of carcinogenesis and metastasis.

In pattern recognition, to design an unbiased and low variance classifier that performs well over the entire population, the sample size is expected to be comparable with the number of features that are used to characterise each sample. However, in a typical biological study utilising microarray technology, the number of genes examined is usually much larger than the number of samples under study. The microarray gene expression levels are highly noisy due to non-specific binding occurred during DNA hybridisation. To reduce the noise and ensure the quality of the data for further analysis, raw data from microarray experiments goes through a sequence of pre-processing steps. Typically, the number of genes in the data set will be reduced dramatically. Nevertheless, the number of genes will typically still be over 10 times larger than the number of samples. This raises several critical issues regarding the design of classifiers using microarray data, such as overfitting and peaking in classification accuracy as the number of features increases (Zongker and Jain, 1997; Dougherty et al., 2005). Selecting a subset of decisive genes from the large number of potential ones becomes one of the critical issues in microarray data analysis.

Many feature selection methods are already in the machine-learning literature. Some have been modified for applications in microarray data analysis. In addition, new methods have been developed to solve the classification problem using microarrays (Saeys et al., 2007). A recent review of the different gene expression data analysis methods (Jafari and Azuaje, 2006) has identified the *t*-test and ANOVA as the most common feature selection tools. The two methods are typical cases of univariate parametric methods, in which each feature is assumed to have a certain underlying distribution. Several modifications of the *t*-test and ANOVA have been proposed to

address the small sample size issues (Churchill et al., 2003). The limitation of sample sizes and the uncertainty associated with the underlying assumptions about the expression levels have led to the use of model free non-parametric methods for univariate gene selection. Several scoring methods such as the Wilcoxon rank sum method and the method using the ratio of Between Sums of Squares (BSS) to Within Sums of Squares (WSS) have been proposed in the literature (Dudoit et al., 2002; Troyanskaya et al., 2002).

Univariate methods such as *t*-test and BSS/WSS method are simple to use and intuitive to comprehend, but they ignore redundancy in the data. There are several multivariate methods in the literature, which account for gene interaction by considering the correlation between the expression levels of genes. They include correlation-based feature selection methods including the minimum redundancy method (Ding and Peng, 2005), the uncorrelated shrunken centroid approach (Yeang et al., 2001) and a simple bivariate method in which two genes are considered together at a time (Bo and Jonassen, 2002).

Feature selection methods that use classifiers embedded in them have also been proposed. These methods are called wrapper methods. In a wrapper method, the gene space is explored to obtain the best feature set that optimises the performance of the embedded classifier. Methods used to attain the optimal feature set include the hill climbing method, the forward search method and sequential floating forward search methods (Dougherty et al., 2005). Wrapper methods have given good results, but they are computationally more demanding and their performances depend on the embedded classifiers.

Wavelet-based methods have also been used to address the gene selection problem. Subramani et al. used the Haar wavelet power spectrum to rank each gene (Prabakaran et al., 2006). Li et al. (2006) took the 1D discrete wavelet transform 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.

In this study, we use wavelet-smoothing operations for selecting differentially expressed genes. The genes are scored based on their capability to discriminate two classes of microarray gene expression profiles. We provide a brief overview of the wavelet-smoothing operations and the proposed scoring scheme used in this study. We applied the method combined with the *k*NN method to several real microarray data sets. We present the results obtained using the third-order approximation of the expression levels using Daubechies-3 (db-3), Daubechies-8 (db-8) and Coiflet-3 (coif-3) wavelets.

2 Methods

2.1 Wavelet transform

The wavelet transform decomposes a signal using scaled and translated versions of localised waveforms called wavelets (Addison, 2002). Since wavelets can be used to construct filters for both stationary and non-stationary signals, the wavelet transform has been widely applied in signal processing in many fields. The discrete wavelet transform is defined by the two functions, the wavelet function that presents the details

in the signal and the scaling function that provides an approximation of the signal. A dyadic grid wavelet and a scaling function are given in equations (1) and (2). The scaling term is a power of 2 and the translation term a factor of 2^m.

$$\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n), \quad m \in Z, n \in Z \tag{1}$$

$$\phi_{m,n}(t) = 2^{-m/2}\phi(2^{-m}t - n), \quad m \in Z, n \in Z \tag{2}$$

where ψ is the wavelet function, ϕ is the scaling function, m corresponds to the level of approximation whereas 2^{-m} represents the resolution, and Z stands for the integer set.

The discrete wavelet transform decomposes a signal using discrete values of dilation and translation parameters to obtain the detail coefficients $T_{m,n}$ and approximation coefficients $S_{m,n}$ as presented in equations (3) and (4).

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)\,\mathrm{d}t, \quad m \in Z,\ n \in Z \tag{3}$$

$$S_{m,n} = \int_{-\infty}^{\infty} x(t)\phi_{m,n}(t)\,\mathrm{d}t, \quad m \in Z,\ n \in Z \tag{4}$$

where $x(t)$ represents the original signal under study. The detail signals and approximation signals at the mth level can be reconstructed from the detail and approximation coefficients $T_{m,n}$ and $S_{m,n}$ using equations (5) and (6).

$$d_m(t) = \sum_{n \in Z} T_{m,n}\psi_{m,n}(t), \quad m \in Z \tag{5}$$

$$a_m(t) = \sum_{n \in Z} S_{m,n}\,\phi_{m,n}(t), \quad m \in Z \tag{6}$$

where $d_m(t)$ corresponds to the detail signal at level m and $a_m(t)$ represents the approximation signal at level m. The approximation signal is created such that the approximation at level m_0 can be written as the combination of detail signals of levels higher than m_0 (equation (7)).

$$a_{m_0}(t) = \sum_{m > m_0} d_m(t), \quad m_0 \in Z. \tag{7}$$

The original signal $x(t)$ can be then reconstructed at the level m_0 by adding the approximation signal at that level and all the detail signals at that level and lower levels as shown in equation (8):

$$x(t) = a_{m_0}(t) + \sum_{m \le m_0} d_m(t). \tag{8}$$

The process of breaking down a signal into its approximation and detail parts is called multi-resolution analysis. This analysis allows one to filter out the noise in the original signal and keep the signal resolution to an appropriate level.

The wavelets are assumed to have a finite number of characteristic coefficients. It has been proved that the approximation coefficients and detail coefficients at the level $m + 1$ can be estimated from the approximation coefficients at the next lower level (Mallat, 1989).

$$T_{m+1,n} = \frac{1}{\sqrt{2}} \sum_{k \in Z} b_k S_{m,2n+k}, \quad n \in Z \tag{9}$$

$$S_{m+1,n} = \frac{1}{\sqrt{2}} \sum_{k \in Z} c_k S_{m,2n+k}, \quad n \in Z. \tag{10}$$

This reduces the calculation of detail and approximation coefficients to a simple filtering operation, where the coefficient $(1/\sqrt{2})b_k$ corresponds to the high-pass filter and $(1/\sqrt{2})c_k$ correspond to the low-pass filter. We used the wavelet toolbox of MATLAB for wavelet analysis.

Feature selection

In this study, the features (genes) are selected based on the approximation signal of gene expression levels. The expression levels of each gene in different samples first went through a sequence of pre-processing steps. The pre-processed expression levels are then arranged based on their sample class labels, so that the expression levels in the samples belonging to one class are grouped together. The expression signals corresponding to each gene are decomposed using the 1D discrete wavelet transform to the third level. In the study, the transform is performed using db-3, db-8 and coif-3 wavelets separately. The detail coefficients at levels 1 and 2 are filtered out. The expression signal is reconstructed using the approximation coefficients at the third level. All the genes are then scored based on the difference between the average expression levels of two classes as shown in equation (11).

$$\text{Score (gene}_i) = \left| \frac{1}{n_1} \sum_{j \in C_1} e_{ij}^l - \frac{1}{n_2} \sum_{j \in C_2} e_{ij}^l \right|, \tag{11}$$

where e_{ij}^l represents the expression level of gene i in sample j after passing the lth level of wavelet filter (we note that in this study l is taken to be 3); n_1 and n_2 stand for the number of samples in classes 1 and 2, respectively; C_1 and C_2 represent the two sets of samples from classes 1 and 2, respectively. All the genes are ranked according to their corresponding scores and the genes with scores higher than a specified threshold are selected to form a feature set for sample classification. The scoring method applied to the leukaemia data set is illustrated in Figures 1 and 2. The leukaemia samples are arranged into two groups along the x-axis. The first 43 samples are Acute Lymphoblastic Leukaemia (ALL) samples. The second group consists of 20 Acute Myeloid Leukaemia (AML) samples. Figure 1(s) shows the expression levels of CST3 (Cystatin C) from leukaemia data set. As we can see, the expression level is very high in AML samples compared with ALL. The signal is decomposed using the wavelet db-8 to the third level. Figure 1(a1), (a2) and (a3) shows the approximation signal at levels 1, 2 and 3, respectively. Figure 1(d1), (d2) and (d3) explains the detail signals at their corresponding levels. We clearly see that the signal at the third level (Figure 1(a3)) captures the differential expression levels between the two classes. The expression signal past this

level of filtering is used to calculate the score of a feature, representing its discriminatory power. Figure 2 presents a comparison of the scores of two genes LDHA and SEMA3C from the lymphoma data set. Figure 2(a) shows the original expression signal of LDHA. Figure 2(b) shows the expression signal after the third-level filtration using db-8 wavelet. Figures 2(c) and 5(d) illustrate the original expression signal of SEMA3C and its expression signal after the third-level filtration. We can see LDHA is differentially expressed between DLBCL and FL samples and provides constructive information for differentiating the two cancer classes. On the other hand, the expression level of SEMA3C apparently varies randomly among the samples and may not be able to differentiate the two classes.

Figure 1 The process of scoring a feature. The plot shows the scoring process for gene CST3 (Cystatin C). The expression levels of CST3 in different samples are arranged into two groups along the *x*-axis based on their classes. (S) shows the original expression levels of CST3. (a1), (a2), and (a3) represent the signal passed the Daubechie wavelet db-8 filter at levels 1, 2 and 3, respectively. (RS) illustrates the overlay of the original signal (S) and its level 3 approximation (a3). (d3), (d2) and (d1) represent the detail signals at their corresponding levels

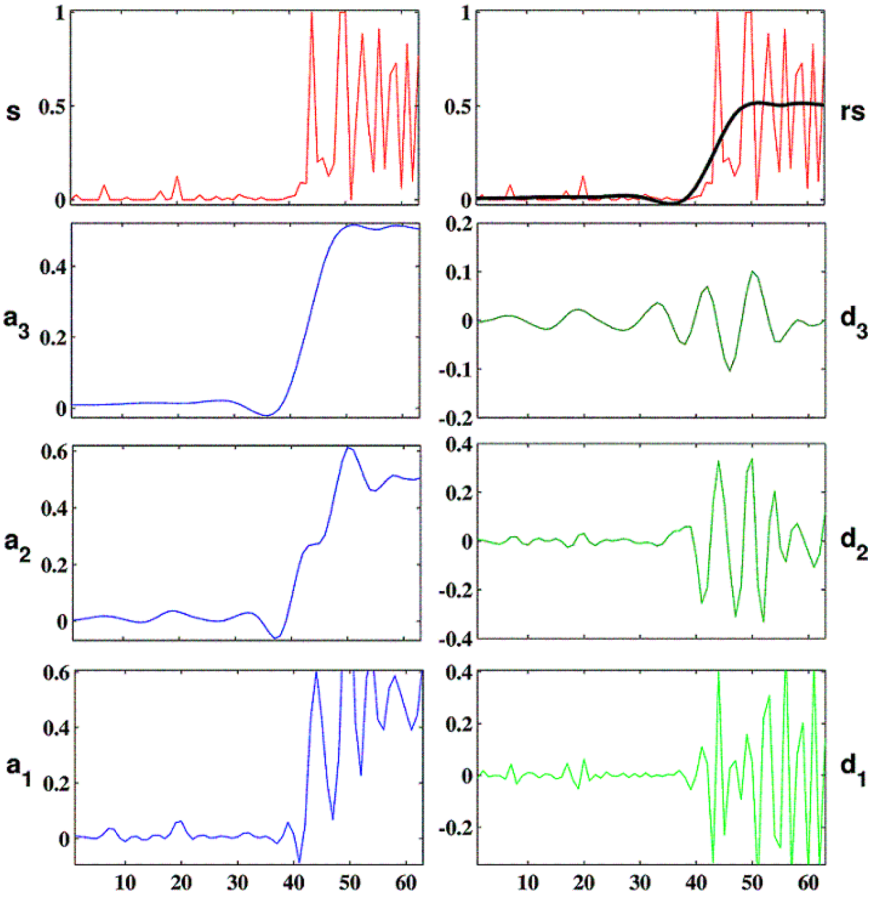
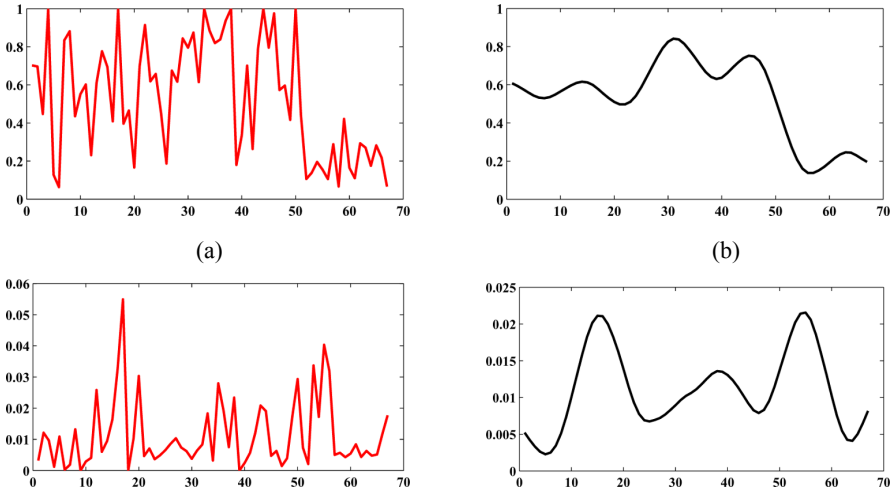


Figure 2 A comparison of the expression signals and the scores of an informative gene and non-informative gene. The plot illustrates the original expression signal of an informative gene LDHA (Lactate dehydrogenase A): (a) and its third-level approximation obtained using db-8 wavelet (b). The original expression signals of a non-informative gene SEMA3C (Semaphorin E) and its third-level approximation are shown in (c) and (d), respectively



K-nearest neighbour classifier

To test the effectiveness of wavelet-based feature selection method, the classical *k*NN classification method (Duda et al., 2001) was utilised. In the method, the Euclidean distances between a sample of unknown class and all training samples are calculated using the selected features. The *k*NN of the uncharacterised sample are identified. Each neighbour has a vote of its own class for the sample. The class of the uncharacterised sample is determined by the majority vote of its neighbours.

3 Results

To examine the efficiency of the proposed wavelet-based feature selection method, we carried out experiments to compare the performance of the method with the commonly used methods *t*-test and BSS/WSS method. Three publicly available microarray data sets were used in the experiments. The data sets consist of the leukaemia data set (Golub et al., 1999), the lymphoma data set (Shipp et al., 2002) and the colon cancer data set (Alon et al., 1999). The leukaemia data set contains 73 samples of human acute leukaemia, including 48 Acute Lymphoblastic Leukaemia (ALL) samples and 25 AML samples. Each sample consists of the expression levels of 7129 probes for 6817 genes. The lymphoma data set contains 77 lymphoma samples, including 58 B-cell lymphomas (DLBCL) and 19 Follicular Lymphomas (FL). The same microarray chips were used for this data set as for the leukaemia data set. Therefore, each lymphoma sample consists of the expression levels of 7129 probes for 6817 genes. The third data set is the colon cancer data set. It contains 62 samples from colon cancer patients, including

40 samples of cancer tissues and 22 normal biopsies samples. Each sample in the colon cancer data set consists of 2000 selected genes from the original microarray data.

Samples in each data set went through a sequence of pre-processing steps, involving elimination of non-informative genes and scaling (Dudoit et al., 2002). First, genes that satisfy one of the following conditions were removed from the analysis:

1 housekeeping genes (endogenous and other miscellaneous controls)

2 genes whose expression values are marked as 'present' for less than 20 samples.

The data sets were threshold at the levels used in the original papers through scaling. More specifically, it involves

1 setting the expression levels to be 16,000 if they are above the threshold for leukaemia and lymphoma data set. The threshold is 8000 for colon cancer data set

2 setting the expression levels to be 20 if they are below 20 for the lymphoma and colon cancer data sets and setting the levels to be 100 when the level is below 100 for the leukaemia data set

3 scaling the expression levels to be in the range of [–1, 1].

To evaluate the effectiveness of the wavelet-based feature selection method, *k*NN classifiers with *k* being 1, 3 and 5 were employed. The performances were compared with the standard parametric feature selection method *t*-test and a standard non-parametric BSS/WSS method. The *k*NN classifiers were combined with the top features selected through each of the three selection methods. Euclidean distances calculated via the selected features were used to identify the neighbours of a sample. The majority vote from the neighbours was used to make the class assignment for the sample under classification. The average classification errors were obtained using the Monte Carlo cross-validation sub-sampling strategy (Boulesteix et al., 2008). The strategy picks *m* randomly selected samples from the sample space out of training. The remaining samples are used as training data to select the informative genes. The *m* left-out samples are then used as test samples to evaluate the classification performance. In this study, the *m* is taken to be 10 and the sub-sampling was repeated 250 times for each data set. The results are presented in Figures 3–5 and Tables 1 and 2.

Table 1 Top 10 genes selected by db-8 wavelet

	Leukaemia data set		*Lymphoma data set*		*Colon cancer data set*	
Rank	*Gene symbol*	*Average score*	*Gene symbol*	*Average score*	*EST ID*	*Average score*
1	CST3	0.475	MTL5	0.443	T95018	0.236
2	MPO	0.404	LDHA	0.428	M63391	0.199
3	AZU1	0.387	ENO1	0.407	T58861	0.170
4	IL8	0.384	CTSB	0.370	T61609	0.168
5	FTL	0.340	PKLR	0.318	T92451	0.153
6	GPX1	0.297	PAPLN	0.310	U14971	0.141
7	TCL2	0.294	CLU	0.305	T57619	0.137
8	CFD	0.292	MIF	0.298	R22197	0.133
9	LYZ	0.288	IFI30	0.296	M87789	0.132
10	SDC1	0.285	APOE	0.292	T72863	0.131

Table 2 Number of common genes between the top 100 genes identified by different feature selection methods

	Leukaemia data set				
Feature selection method	*db-3*	*db-8*	*coif-3*	*BSS/WSS*	*T-test*
db-3		99	100	42	31
db-8	99		99	42	31
coif-3	100	99		43	31
BSS/WSS	42	42	43		53
T-test	31	31	31	53	
	Lymphoma data set				
Feature selection method	*db-3*	*db-8*	*coif-3*	*BSS/WSS*	*T-test*
db-3		99	99	32	30
db-8	99		99	33	30
coif-3	99	99		32	30
BSS/WSS	32	33	32		65
T-test	30	30	30	65	
	Colon cancer data set				
Feature selection method	*db-3*	*db-8*	*coif-3*	*BSS/WSS*	*T-test*
db-3		99	99	34	33
db-8	99		100	33	33
coif-3	99	100		33	33
BSS/WSS	34	33	33		82
T-test	33	33	33	82	

Figure 3 Classification error for the leukaemia data set. The plot compares the average classification errors for the leukaemia data set. The classifications were performed using 3NN classifiers coupled with *t*-test, BSS/WSS, and three wavelet-based feature selection methods, respectively. The averages were done over 250 repetitions of Monte Carlo sub-sampling

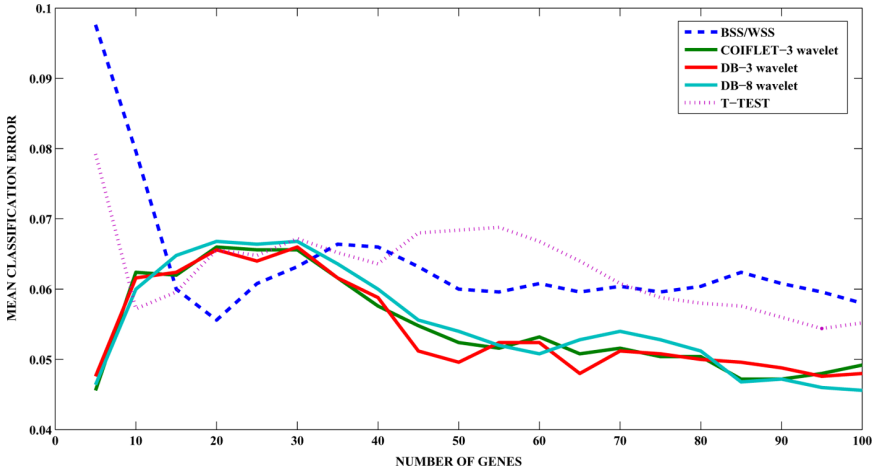


Figure 4 Classification error for the lymphoma data set. The plot compares the average classification errors for the lymphoma data set. The classifications were performed using 3NN classifiers coupled with *t*-test, BSS/WSS, and three wavelet-based feature selection methods respectively. The averages were done over 250 repetitions of Monte Carlo sub-sampling

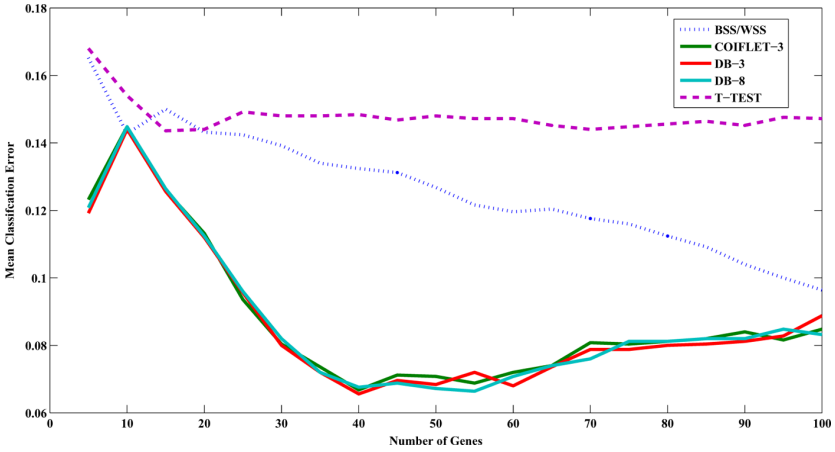
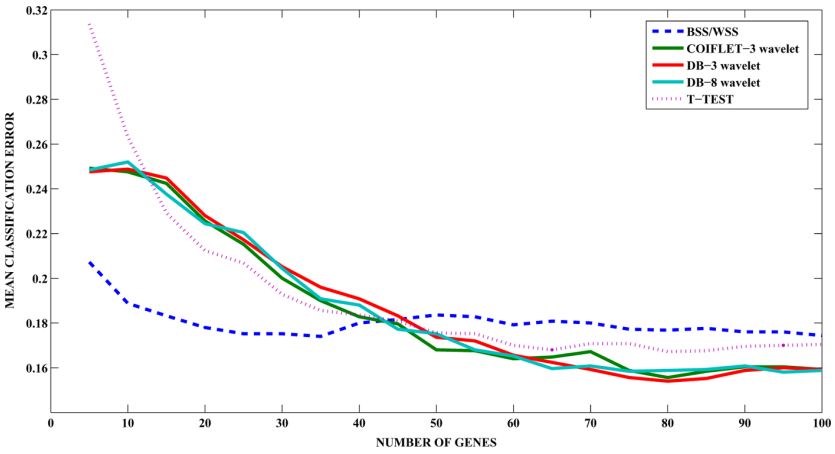


Figure 5 Classification error for the colon cancer data set. The plot compares the average classification errors for the colon cancer data set. The classifications were performed using 3NN classifiers coupled with *t*-test, BSS/WSS, and three wavelet-based feature selection methods, respectively. The averages were done over 250 repetitions of Monte Carlo sub-sampling



4 Discussion

Figure 3 presents and compares the average classification errors for the leukaemia data set using 3NN classifier with features selected through *t*-test, BSS/WSS and the three wavelet transforms using db-3, db-8 and coif-3. The classification errors were averaged over 250 runs of the experiments using Monte Carlo cross-validation sub-sampling. The *x*-axis represents the number of top features used in the classification whereas *y*-axis shows the average classification error. As we can see clearly, although the BSS/WSS

method performed better when the number of features was chosen to be between 15 and 35, the 3NNs coupled with wavelet transforms consistently outperform both *t*-test and BSS/WSS when more than 40 genes are used. The wavelet-based feature selection method performed especially well when over 50 top genes were used. The classification errors using *t*-test and BSS/WSS were reduced by about 25%. The same experiment was performed with the lymphoma data set. The results are shown in Figure 4. The results demonstrate even more the remarkable advantage of using the wavelet transforms. The 3NN classifier coupled with the wavelet transforms outperformed *t*-test and BSS/WSS over the entire spectrum of selected genes. The biggest differences were observed when about 40 genes were selected. We observed that the classification error from the BSS/WSS method decreased with the number of selected genes. However, the wavelet-based methods consistently outperform BSS/WSS until the number of selected genes is about 100. We also observed that the BSS/WSS method performs roughly the same as the wavelet-based methods when the number of selected genes is between 100 and 200. The outcome for the colon cancer data set presents a slightly different picture (Figure 5). The wavelet-based methods outperformed the *t*-test and BSS/WSS when the number of selected features is more than 50. We note that the available features from this data set are relatively fewer than a normal microarray gene chip provides. The 2000 features were preselected through other pre-processing steps, which might provide some bias towards the *t*-test and BSS/WSS for using a relatively small number of features in the classification.

The wavelet coefficients of a signal do depend on the order in which the samples are arranged within a class. The gene selection process should not. We investigated the effect of the sample arrangement within groups on the proposed method by shuffling the pre-grouped training data within each class 100 times independently. The lists of top 100 selected features were compared between each round. It was found that more than 87% of the genes were consistent across all the 100 gene lists. It confirms that the score of each gene calculated using equation (11) do not vary much with the order of samples within a class.

Furthermore, other wavelets such as Haar and Sym4 were explored in this study (Mallat, 1989). We found that the Daubechies wavelet db-3, db-8 and the Coiflet coif-3 present better classification 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. Wavelets db-3, db-8 and coif-3 apparently share similar shape, whereas the Haar and Sym4 wavelets have significantly different properties such as continuity and smoothness when compared with db-3, db-8 and coif-3. We suspect that the best wavelets for selecting genes for microarray sample classification may depend on the platforms and samples used in the microarray experiments.

*k*NN classifiers with different values of k ($k = 1, 3, 5$) were also examined for testing classification accuracy. The nearest neighbour method ($k = 1$) and 3NN method gave comparable results. The results show that the wavelet-based method outperforms the two other methods for all three data sets. The performances of all feature selection methods dropped when $k = 5$, suggesting the votes from the distant neighbours should not be included directly in making a class assignment. In addition, weighted *k*NN methods were also applied, in which each selected gene carries a weight in the distance calculation. The weight was calculated based on the relative score of the gene as described in equation (12):

$$\text{weight}(\text{gene}_i) = \frac{\text{score}(\text{gene}_i)}{\sum_{\substack{\text{selected} \\ \text{genes}}} \text{score}(\text{gene}_j)}. \tag{12}$$

It was observed that the classification error was significantly reduced when the weighted *k*NN method was used.

Table 1 presents the top genes that were selected most frequently during the sub-sampling process. We note that the top genes selected by db3, db-8 and coif-3 wavelets were almost identical for each data set. The results are reflected in the classification error plots presented in Figures 3–5. The performances of both wavelets were roughly the same. For the leukaemia data set, CST3, IL8, AZU1, LYZ have been identified as some of the most important genes in distinguishing ALL from AML in the original study (Golub et al., 1999). MPO and TCL2 have also been identified to be associated with AML and ALL, respectively. Many of the genes reported as important in the original study were ranked very highly by the wavelet-based methods and featured in the top 100 genes list in all the 250 re-samplings. For the lymphoma data set, several of the genes found to be informative by the original study were ranked very highly by the wavelet-based methods. Lactose Dehydrogenase A (LDHA) is a known biomarker for B-cell lymphoma as reported in the original paper (Shipp et al., 2002). CTSB, CLU and ENO1 in the top 10 list had also been identified as very important. Many of the genes identified as relevant in the original study received very high scores and most of them featured in the top 100 list in all the 250 re-sampling studies. For example, HMG-1 ranked 12th and CTSD ranked 16th. In the case of colon data, the original study reports the importance of soft-muscle-related ESTs in identifying colon cancer. They use the average intensity levels of 17 ESTs as a muscle index, with a lower index identifying tumours (Alon et al., 1999). The wavelet-based method gave high scores to most of these 17 ESTs.

The top 100 genes were compared with the ones identified by *t*-test and BSS/WSS. A matrix showing the numbers of common genes selected by different methods for the three data sets is presented in Table 2. As we can see, almost 100% of the genes found by the three wavelets are the same, indicating the consistency of the three wavelet methods. About 32% of the top 100 genes obtained by the wavelet-based methods and the *t*-test were found to be common in the three data sets. About 38% of the top 100 genes obtained by the wavelet-based methods and BSS/WSS were common. On the other hand, the number of common genes selected by the *t*-test and the BSS/WSS method is much higher. The relative low percentage of common genes identified by wavelet-based methods and by *t*-test or BSS/WSS indicates that the wavelet-based methods offer a different perspective in terms of differentially expressed genes. Therefore, the proposed wavelet-based gene selection method can facilitate the identification of differentially expressed genes, which might be otherwise neglected.

5 Conclusions

A 1D discrete wavelet-transform-based gene selection method was proposed. It was developed based on the observation that the third-level 1D wavelet approximation captures the differential microarray gene expression between sample classes. The genes that exhibit high differences between the average wavelet approximations of

the expression levels are selected to form a feature set for sample classification. Our experiments illustrate that the wavelet-based method consistently outperformed the standard methods in all the three data sets studied, particularly when the number of selected genes was more than 40. We conclude that the wavelet analysis is a valuable tool for studying gene expression patterns. The wavelet-based gene selection method can be used to identify a set of highly discriminant genes for microarray data classification.

Acknowledgements

This research is funded in part by the graduate assistantship from the Department of Biomedical Engineering at UA, NSF DUE 0410727, and UA faculty research fellowship.

References

Addison, P.S. (2002) *The Illustrated Wavelet Transform Handbook*, Taylor & Francis, New York.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Brarra, S.Y., Mack, D. and Levine, A.J. (1999) 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proc. Natl. Acad. Sci., USA*, Vol. 96, pp.6745–6750.

Bo, T. and Jonassen, I. (2002) 'New feature subset selection procedures for classification of expression profiles', *Genome Biology*, Vol. 3, No. 4, pp.17.1–17.11.

Boulesteix, A.L., Strobl, C., Augustin, T. and Daumer, M. (2008) 'Evaluating microarray-based classifiers: an overview', *Cancer Informatics*, Vol. 6, pp.77–97.

Churchill, G.A., Xiangqin, C. and Gary, A. (2003) 'Statistical tests for differential expression in cDNA microarray experiments', *Genome Biology*, Vol. 4, p.210.

Ding, C. and Peng, H. (2005) 'Minimum redundancy feature selection from microarray gene expression data', *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, pp.185–205.

Dougherty, E.R., Datta, A. and Sima, C. (2005) 'Research issues in genomic signal processing', *IEEE Signal Processing Magazine*, Vol. 22, No. 6, pp.46–68.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*, 2nd ed., John Wiley & Sons, Inc.

Dudoit, S., Fridlyand, J. and Speed, T. (2002) 'Comparison of discrimination methods for the classification of tumors using gene expression data', *J. Amer. Statist. Assoc.*, Vol. 97, pp.77–87.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligirui, M., Bloomfield, C. and Lander, E. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, No. 5439, pp.531–537.

Jafari, P. and Azuaje, F. (2006) 'An assesment of recently published gene expression data analysis: reporting experimental design and statistical factors', *BMC Medical Informatics Decision Making*, Vol. 6, p.27.

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

Mallat, S.G. (1989) 'A theory for multiresolution signal decomposition: the wavelet representation', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp.674–693.

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

Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003) 'A molecular signature of metastasis in primary solid tumors', *Nat. Genet*., Vol. 33, No. 1, pp.49–54.

Saeys, Y., Inza, I. and Larranaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, Vol. 23, No. 19, pp.2507–2517.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J. and Golub, T. (2002) 'Diffuse large b-cell Lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine*, Vol. 8, No. 1, pp.68–74.

Troyanskaya, O., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. (2002) 'Nonparametric methods for identifying differentially expressed genes in microarray data', *Bioinformatics*, Vol. 18, pp.1454–1461.

Yeang, C., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R., Angelo, M., Reich, M., Lander, E., Mesirov, J. and Golub, T. (2001) 'Molecular classification of multiple tumor types', *Bioinformatics*, Vol. 17, Suppl. 1, pp.S316–S322.

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.

Zongker, D. and Jain, A.K. (1997) 'Feature selection: evaluation, application and small sample perfomance', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 2, pp.153–158.