

Discrete Wavelet Transform Based Feature Extraction for Tissue Classification Using Gene Expression Data

¹Xiaoying DOND ¹Guangmin SUN

¹Department of Electronic Engineering

Beijing University of Technology

Beijing, 100022, China

xiaoying_dong@emails.bjpu.edu.cn

gmsun@bjpu.edu.cn

²Guandong XU

²Department of Etiology and Carcinogenesis

Peking Union Medical College

Beijing, 100021, China

xgdxl@hotmail.com

Abstract

DNA microarrays can be used to measure the expression levels of thousands of genes simultaneously. In this paper, the gene expression data were processed by a signal processing method. A discrete wavelet transform (DWT) based feature extraction method for cancer classification was introduced, by which micro-array data are transformed into time-scale domain and used as classification features. Finally, some test and comparison experiments for the feature extraction method have been made by using the weighted voting classification scheme[1]. Experiment results show that the correct rate is over 90% in tumor vs normal classification by using the feature extraction method.

Keywords

gene expression, discrete wavelet transform, feature extraction, tissue classification

1 Introduction

With the advance of hybridization array technology scientists can measure expression levels of thousands of genes across different conditions in parallel. The conditions might be:(1). pathological tissue specimens from patients;(2).internal cellular physiology from different cell lines;(3).diverse physiological conditions in an intact

organism; or(4).serial time points following a stimulus to a cell or organism [2].

Gene expression data offer potential insight into gene function and regulatory mechanisms and aid in better understanding of carcinogenesis. Normal cells can evolve into malignant cancer cells through a series of abnormality in genes that control the cell cycle, apoptosis, and genome integrity, to name only a few [3].As determination of cancer type and stage is momentous to the assignment of appropriate treatment and evaluation of treatment outcomes, one of the most central goal of gene expression data analysis is to classify heterogeneous tissues (e.g., tumors vs. normal) on molecular level.

Gene expression data usually present as a matrix, in which rows represent genes and columns represent samples or observations (e.g. a single micro-array experiment). A novel way to think of micro-array data is as a signals set. The number of genes is the length of signals. From this point of view, information and signal processing technique can be used to micro-array data analysis. In signal processing field there is an impressive

arsenal of tools. Perhaps the most well-known is Fourier analysis, which is extremely useful when the signal's frequency contents are of great importance. At the same time Fourier analysis has a serious drawback. In transforming to the frequency domain, time information is lost. That is, one can find what frequency contents in a signal by Fourier transform, but it is impossible to tell when a particular frequency component took place. This disadvantage makes against finding cancer related genes. In an effort to correct this deficiency, we adapt wavelet transform, which has attracted increasing attention in recent years for its inherent multi-resolution approach to signal analysis.

The rest of the paper is as follows. The wavelet transform theoretical framework is discussed in Section 2. The proposed method, that is feature extraction in time-scale domain, is described in Section 3. We apply the selected features in tissue classifier and give the results in Section 4. followed by a summary in Section 5.

2 Discrete Wavelet Transform

The continuous wavelet transform (CWT) of a finite energy signal $x(t)$ ($x(t) \in L^2(\mathbb{R})$) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function ψ :

$$WT_x(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi^* \left(\frac{t - \tau}{a} \right) dt \quad (1.1)$$

$$WT_x(a, \tau) = \langle x(t), \varphi_{a\tau}(t) \rangle \quad (1.2)$$

where $WT_x(a, \tau)$, a function of scale and position, are many wavelet coefficients.

Calculating wavelet coefficients at every possible scale is a fair amount of work. So only some discrete scales and positions are chosen in practice, which is called DWT. Generally, scales and positions are based upon powers of two ---so-called dyadic discrete wavelet transform.

The full DWT for signal $x(t)$ can be represented in terms of a shifted version of a scaling function $\phi_{j,k}$ and a shifted and dilated version of a so-called mother wavelet function $\psi_{j,k}$. The representation of the DWT can be written as:

$$x(t) = \sum_{k \in \mathbb{Z}} u_{j_0, k} \phi_{j_0, k}(t) + \sum_{j=-\infty}^{j_0} \sum_{k \in \mathbb{Z}} w_{j, k} \psi_{j, k}(t) \quad (3)$$

where $w_{j,k}$ are the wavelet coefficients and $u_{j,k}$ ($j < j_0$) are the scaling coefficients. These coefficients are given by the inner product in $L^2(\mathbb{R})$, i.e.,

$$u_{j,k} = \langle x, \phi_{j,k} \rangle, \quad w_{j,k} = \langle x, \psi_{j,k} \rangle \quad (4)$$

where $\phi_{j,k}(t) = 2^{-j/2} \phi(2^{-j/2}t - k)$ is a family of scalar functions and $\psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j/2}t - k)$ a family of wavelet functions [16].

Once a mother wavelet is selected, the wavelet transform can be used to decompose a signal according to scale, allowing to separate the fine-scale behavior from the large-scale behavior of the signal. The relationship between scale and signal behavior is: low scale 'a' \rightarrow compressed wavelet \rightarrow rapidly changing details \rightarrow high frequency; high scale 'a' \rightarrow stretched wavelet \rightarrow slowly changing, coarse features \rightarrow low frequency [5]. Signal decomposition is typically done in an iterative fashion using the scales $a = 2, 4, 8, \dots, 2^L$, with successive approximations being

decomposed in turn, so that one signal is broken down into many lower resolution components.

Suppose x_i is the i th signal (sample) in micro-array data, and N is the signal length (number of genes). Denote A_L and D_L are level L ($a=2^L$) approximation and detail signals reconstructed from coefficients cA_L and cD_L respectively. With the fast wavelet transform making use of filters, developed in 1988 by Mallat [6], we can implement signal decomposition in an efficient way, as shown in Fig.1.

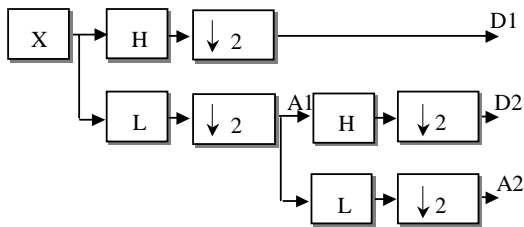


Fig.1. DWT fast algorithm schematic diagram. Where H is high-pass filter and L is low-pass filter.

In this paper, we extract features by using DWT followed by a correlation coefficients ranking method, and then use the weighted voting scheme proposed by Golub et al. [1] to classify the colon tissue sample data set published by Alon et al. [7]. Fig.2 is our tissue classification system block.

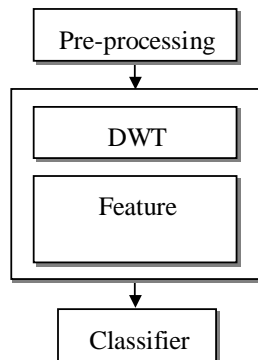


Fig.2. Tissue classification system block

3 DWT Based Features extraction

Most of genes are not relevant to the distinction between cancerous and normal tissues, and furthermore, they introduce noise to the system. Moreover, Finding small gene sets that are sufficiently informative to distinguish between cells of different types is a requirement of diagnosis in practice. Furthermore, it is very important to pathologist to isolate genes which are potentially intimately related to the tumor makeup and pathomechanism [8]. From classification point of view, reduction the dimension of the feature space can help overcome the risk of overfitting. Overfitting problem arises frequently in tissue classification problem where the dimension of the feature vectors (in our case thousands of genes) is typically several orders of magnitude larger than the number of training patterns (in our case a few dozen tissue samples). In such a situation, classification performance on a test set is much more poor than on training set.

Before describing DWT based feature extraction method proposed in this paper, we briefly list some dimensional reduction scheme that have proven to be useful in the micro-array data analysis context. Multidimensional scaling (MDS) is usually used to high-dimensional data display, which can project high-dimensional data points onto 2 or 3 dimensions while preserving the space structure within data set. Principal component analysis (PCA) is an usual feature extraction tool, which generates a new set of principal components by combining original variables linearly. All the principal components are orthogonal to each other so there is no redundant

information in the new low-dimensional space. There are also many statistical approaches, most of which are based upon feature-ranking techniques. These approaches include ranking with correlation coefficients, ranking with disorder, ranking with likelihood, ranking with TNoM (Threshold Number of Misclassification), etc.[9]. In recent years some machine learning based feature extraction techniques have emerged, such as genetic algorithm based, SVM based, ANN based, etc.

In our DWT based feature extraction scheme, signals (tissue samples) are firstly transformed to time-scale domain by multilevel wavelet decomposition. Both of the coefficients extracted and signals reconstructed are used as raw features to be selected. One of the signal and its DWT results are shown in Fig.3,

and the corresponding reconstructed signals by inverse discrete wavelet transform (IDWT) are shown in Fig.4.

In this paper, classification performance is compared with the commonly used method with and without DWT processing stage. At the same time decomposition level is selected tentatively. Several families of wavelets that have proven to be especially useful in many applications[9-12] are compared in our experiments to find a more suitable wavelet to our classification system. These mother wavelets are from the Haar, Daubechies, Symlet, biorthogonal, and Coiflet families[13]. By varying the mother wavelet, the classification accuracy can be greatly affected. Experimental results are given in section 4.

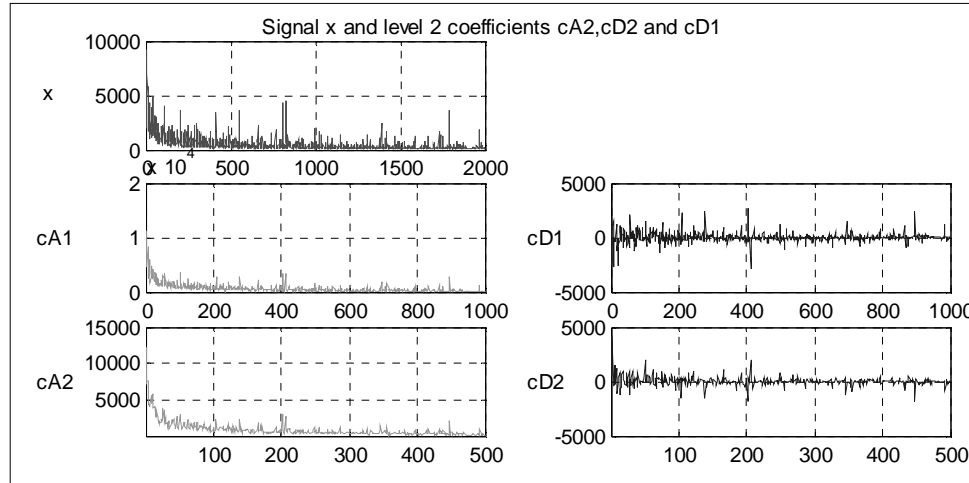


Fig.3. Approximation and detail coefficients extracted by wavelet decomposition with haar wavelet at level 2.

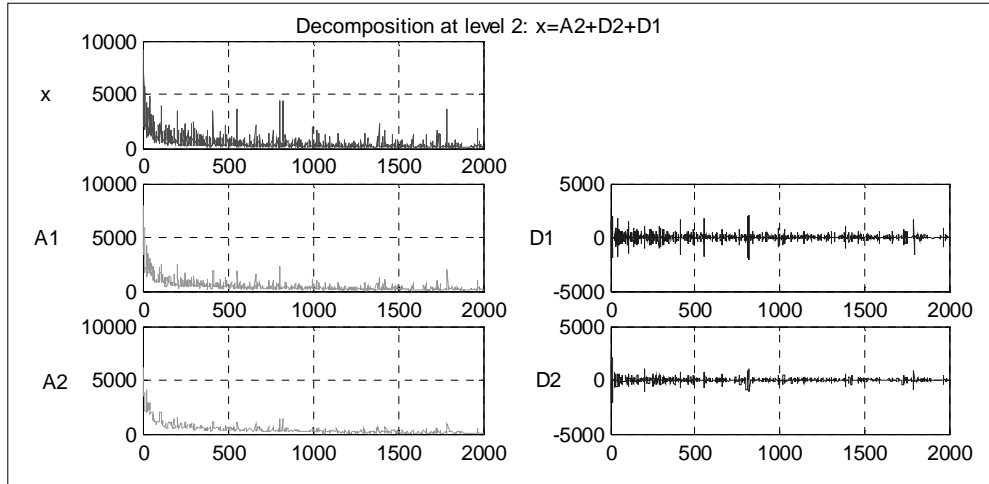


Fig.4. Approximation and detail signals reconstructed by IDWT with haar wavelet at level 2.

4. Classification Experiments And Results

We evaluate the performance of the approach discussed in the previous section on colon cancer data set reported by Alon et al. [7]. This data set consists of more than 6,500 human genes and 62 tissue samples of colon epithelial cells collected from colon cancer patients, including 40 tumor and 22 normal samples. The tumor and normal biopsies were collected from tumors and healthy parts of the same patients respectively.

Because the number of samples is very small, we use a common statistical tool, leave one out cross validation (LOOCV), to test the accuracy of the classifier. Reader would prefer to [14] for details about LOOCV.

Firstly, we explore several wavelets, including sym2 from Symlet family, bior1.1, bior2.2 and bior6.8 from biorthogonal family, and coif2, coif4 from Coiflet family. With these wavelets, signal is broken down into many lower resolution components by iterated

decomposition process. Decomposition is implemented at the first, the second and the third level respectively. One can also choose a suitable number of levels according to the nature of the signal, or based on entropy criterion [15].

Table 1. shows the percent of correctly classified samples in the LOOCV evaluation. Group1 comprises DWT based feature extraction methods with a choice of diverse wavelet,decomposition level and different combination of approximation and detail signals. Group 2 lists partial results reported in [3] obtained by some existing feature extraction methods, covering correlation coefficients ranking (CCR) based, clustering based, nearest neighbor, support vector machine (SVM) with linear and quadratic kernel inner product functions and boosting. The comparison results demonstrate that space transformation procedure using DWT prior to feature selection can lead to significant improvements in classification accuracy.

Experiments demonstrate that

reconstructed signals as features excel coefficients universally. Then we adopt A3, D1, D2, D3, D1&D2, D1&D2&D3 and A3&D3 as feature selection inputs respectively. Fig 5. depicts the correct rate versus different feature selection set. The results are shown for different feature set coming from DWT using wavelet db1 (haar), sym2, coif2, bior2.2 and bior6.8. As we can see, varying the mother wavelet can affect classification accuracy greatly. In addition, detail signals, especially the first level detail signal D1, may be regarded as more suitable feature selection set. As the same time, approximation signals have no much weightiness in the problem of tissue classification by gene expression data.

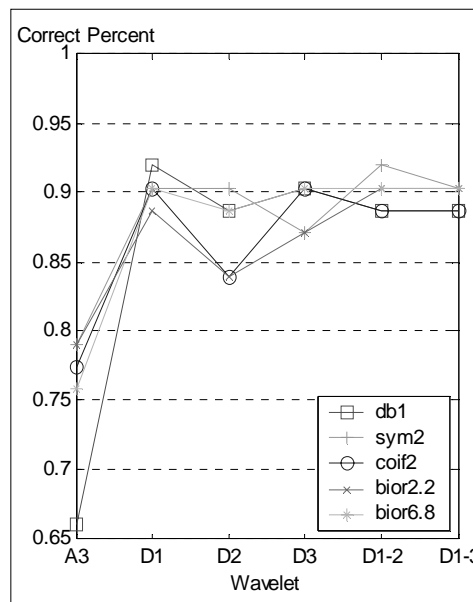


Fig.5. Correct rate with various feature selection set

Table1. Classification performance of different methods.

	Method	Correct Rate(%)
DWT Based	Level=2, db2, D ₁ &D ₂	92.0
	Level=2, db1, D ₁	92.0
	Level=3, db1, D ₁	92.0
	Level=2, db1, D ₁ &D ₂	90.3
	Level=2, db6, D ₁	90.3
	Level=3, db2, D ₂	90.3
	Level=3, sym2, D ₁ &D ₂	92.0
	Level=3, bior1.1, D ₁	92.0
	Level=3, coif2, D ₁	90.3
	Level=3, coif4, D ₁	90.3
	Level=3, bior2.2, D ₁ &D ₂	90.3
	Level=3, bior6.8, D ₁	90.3
No DWT	CCR without DWT	88.7
	Clustering	88.7
	Nearest Neighbor	80.6
	SVM, linear kernel	77.4
	SVM, quad kernel	74.2
	Boosting, 100 iter.	72.6

5. Summary and Outlook

With the development of information processing techniques, it is necessary to introduce some novel and advanced signal processing techniques into gene expression data analysis to use the full potential of micro-array tool and facilitate correlative research field.

In this paper, we deal with gene expression data from a signal processing perspective. DWT based feature extraction scheme is advanced in this paper and we demonstrate correct rate of at least 90% in tumor vs normal classification. We highlighted the new angle of processing gene expression data though this approach is maybe not the best bet at the present time as compared with some existing methods.

Further work comprises an in-depth study of wavelet transform based feature extraction method, including feature selection after WT and so on. Other

aspects that deserve further investigation include WT in conjunction with some other distinguished pattern recognition methods, such as SVM, ANN and so on.

Finally we note that the approach presented here has a general applicability in various classification problems in biology field.

References

- [1] T.R. Golub, D. K. Slonim et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(15): 531-537, 1999.
- [2] Raychaudhuri, Patrick D. Sutphin et al., Basic microarray analysis: Grouping and feature reduction, *TRENDS in Biotechnology*, Vol.19 No.5: 189-193, May 2001.
- [3] Amir Ben-Dor, Laurakay Bruhn et al., Tissue classification with gene expression profiles, *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000.
- [4] Mallat, S., A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Pattern Anal. and Machine Intell.*, vol. 11, no. 7: 674-693, 1989.
- [5] U.Alon, N.Barkai et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.*, Vol.96, pp. 6745-6750, June 1999.
- [6] Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *BIOWulf Technical Report*, 2000.
- [7] M. Nadir Kurnaz, Tamer Ölmez, Determination of features for heart sounds by using wavelet transforms, 15th IEEE Symposium on Computer-Based Medical Systems (CBMS'02), June, 2002.
- [8] L.M.Bruce, C.H.Koger,J.Li, Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction, *IEEE Transactions on geoscience and remote sensing*, vol.40, No.10, October, 2002.
- [9] P. Xu and A. K. Chan, Fast and robust neural network based wheel bearing fault detection with optimal wavelet features, *International Joint Conference on Neural Networks (IJCNN)*, 2002.
- [10] C.Wang,R.X.Gao,Wavelettransform with spectral post-processing for enhanced feature extraction, *IEEE Instrumentation and Measurement Technology Conference*, 2003.
- [11] Juan Liu, Hitoshi Iba et al., Selecting informative genes with parallel genetic algorithms in tissue classification, *Genome Informatics*, 12: 14–23 , 2001.
- [12] Duda, R.O. & Hart, P.E., *Pattern recognition and scene analysis*, John Wiley & Sons, New York, 1973.
- [13] Coifman, R.R., M.V Wickerhauser, Entropy-based algorithms for best basis selection, *IEEE Trans. on Inf. Theory*, vol.38,2, pp.713-718, 1992.
- [14] Magnus Orn Ulfarsson, Jon Atli Benediktsson et al., Wavelet feature extraction and genetic feature selection for multisource data, *IEEE* , 2002.