Classification of Temporal Gene Expression Data Using Wavelet Domain Feature in Kernel SVM

S. A. Fattah[1*], M. Maksud[2], A. K. Biswas[3], N. Zannat[4], Y. Luo[5], and S.-Y. Kung[5]
[1]Department of Electrical and Electronic Engineering, BUET, Dhaka, Bangladesh
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, TX
[3]Department of Electrical and Computer Engineering, Virginia Commonwealth University, VA
[4] Department of Electrical and Computer Engineering, North Carolina State University, NC
[5]Department of Electrical Engineering, Princeton University, NJ
E-mail: *fattah@eee.buet.ac.bd

***Abstract*—In this paper, an effective feature extraction scheme is proposed to classify cell-cycle regulated genes into two major phases: G1 and non-G1. In order to overcome the lack of regularity in temporal as well as corresponding spectral pattern of the gene expression data, the discrete wavelet transform (DWT) is employed on the temporal data to extract features. Because of low frequency dominancy in the gene expression profile, only approximate DWT coefficients are considered. For the purpose of classification, kernel based support vector machine (SVM) is employed where radial basis kernel function is utilized. Performance of some well-known classifiers is also investigated using the proposed wavelet domain feature. From extensive simulation based on leave one out cross validation technique, it is found that the proposed scheme provides very low error rates in comparison to some of the existing methods for the case of the most widely used yeast microarray gene expression database containing genes with or without having the biologically proven ground truth on cell-cycle information.**

***Index Terms*— Cell-cycle, discrete wavelet transform, feature extraction, gene expression, kernel, support vector machine.**

I. Introduction

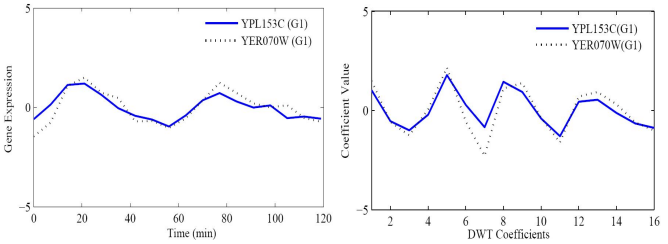
THE temporal pattern analysis of gene expression levels can play an important role in diagnosing and predicting diseases in medical research. Microarray gene expression data exhibit different expression levels during the progression of the cell cycle. Identification of cell-cycle regulated genes from a large gene dataset has been attempted by several researchers [1]-[4]. In this regard, Yeast Saccharomyces Cerevisiae genes are most commonly used. In [1], based on periodicity and correlation algorithms, 800 genes are identified from 6178 Yeast genes. In [2] and [3], spectral domain analysis and in [4], some bio-molecular genomic experiments are carried out. Among different phases of a cell cycle, G1 (Gap-1) phase is biologically most significant as some genes associated with G1 play a key role in proliferation, oncogenic transformation and cell death. Hence, in most of the research works, G1 versus non-G1 gene classification is performed [5]-[10]. For the purpose of classification, several classifiers are used, such as hierarchical clustering, k-means clustering, and the support vector machine (SVM) [11]. There are few methods that deal with the two class (G1 versus non-G1) gene clustering problem utilizing time domain information [5]-[10]. In [5], functional logistic regression tool based on functional principal components is employed. The local wavelet-vaguelette based functional statistical method proposed in [6] provides an option for dimension reduction. Kernel-induced random forest algorithm is extended in [7] by defining some kernels for functional data and using functional principal component analysis (FPCA). In [8], the FPCA in functional data analysis framework is utilized. Some nonparametric analysis techniques are employed in [9]-[10].

The main objective of this paper is to introduce an efficient feature extraction algorithm followed by a robust classifier for the purpose of classifying cell-cycle regulated genes into two phases. Instead of using directly either the temporal or the spectral feature, the discrete wavelet transform (DWT) is employed on the temporal data, which offers the advantage of both time and frequency localization. The quality of the proposed DWT feature is also investigated in terms of within class compactness and between class separation. In the proposed scheme, kernel SVM classifier is employed with the radial basis function (RBF) kernel. Performance of some commonly used classifiers is also investigated. The classification performance is tested on genes with or without having biologically proven cell-cycle information taken from Spellman's yeast microarray database.

II. Proposed Method

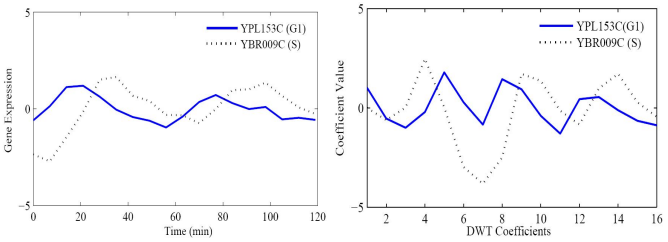
In case of temporal gene expression data, spectral analysis does not give any information about how the expression signal changes with respect to time. It is only suitable for measuring the periodicity. On the contrary, it is found that the wavelet transformation is a measure of similarity between the basis functions (wavelets) and gene expression profiles, and the calculated DWT coefficients refer to the closeness of the gene expression profile to the wavelet at the current scale. In brief, the idea is to obtain an alternative way to extract expression patterns in temporal gene expression data using discrete wavelet analysis. Next, instead of using conventional classifiers, kernel based SVM classifier is employed.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE



(i) (ii)

Fig. 1. Representation of gene expression profile of two similar genes in (i) time and (ii) DWT domain.



(i) (ii)

Fig. 2. Representation of gene expression profile of two dissimilar genes in (i) time and (ii) DWT domain.

A. Motivation behind the DWT Feature

The alphabet used for representing different phases of cell cycle data are: G1 (Gap-1), S (Synthesis), G2 (Gap-2), and M (Mitosis). The two major groups G1 and non-G1 are not always clearly distinguishable. The non-G1 group includes four phases: S, S/G2, G2/M and M/G1. To illustrate the effect of wavelet transformation on the quality of extracted features, two genes from G1 phase, namely YPL153C and YER070W, and one S phase gene, namely YBR009C are considered. In Figs. 1 and 2, time domain microarray gene expression data and corresponding DWT representations with db4 basis function are shown, respectively. It is evident that within group similarity and between group separation are much higher in wavelet domain in comparison to time domain.

B. Proposed Feature Extraction

With a view to obtain time frequency characteristics of the microarray gene data, we propose to employ the DWT which is a multi-resolution transform having a very fast implementation. DWT is a lossless linear transformation of a signal or data into coefficients on a basis of wavelet functions. Wavelets are a family of basis functions that can be used to approximate other functions by expansion in orthonormal series. They offer combination of some powerful properties, such as orthonormality and localization both in time and scale (frequency). One of the key advantages of wavelets is their ability to spatially adapt to features of a function, such as discontinuities and varying frequency behavior. They offer compact support which means each wavelet basis function is supported on a finite interval and it guarantees the localization of wavelets. For a gene expression profile $x(n)$, the DWT coefficients can be obtained as

$$X(a,b) = \sum_{n \in Z} x(n).\frac{1}{\sqrt{a}}\psi\left[\frac{n-b}{a}\right], \qquad (1)$$

where a is the dilation or scale, b the translation, and $\psi[.]$ represents the discrete wavelet. For dyadic wavelet transform, $a = 2^{-j}$, $b = k \times 2^{-j}$, $\psi_{a,b}[n] = 2^{j/2} \times \psi[2^j n - k], k \in Z, j \in N$. Performing the DWT of a signal $x(n)$ is passing it through low pass filters (scaling functions) and high pass filters simultaneously. The frequency of the signal is halved after passing the signal through a filter. So, by Nyquist's rule, half of the samples can be discarded. This is achieved by down-sampling or decimation by a factor 2. Hence, after the filtering and subsequent down-sampling operation, the number of coefficients will be equal to half the length of the original input for each filter. The DWT coefficients obtained from the output of the high-pass and low-pass filters are termed as detail and approximate coefficients, respectively. The filtering operations in the DWT result in a change in the signal resolution, whereas the sub-sampling (down sampling/up sampling) causes change in the scale. Thus, DWT helps in analyzing the signal at different frequency bands with different resolutions. Since gene expression data mainly exhibit low frequency variation, in the proposed method, approximate DWT coefficients are employed.

C. Classification Based on Kernel SVM

The classifier classifies the data into different groups generally, depending on the significant characteristics of the group members. The quality of a classifier depends on its ability to provide the compactness among the member within a cluster and the separation between the members of different clusters in terms of feature characteristics. The task of recognizer is to identify the class label of a test sample utilizing the classified data. In a feature based scheme, classification is performed utilizing the extracted features of the data, instead of directly employing the data themselves.

In the proposed method, the support vector machine (SVM) is used to classify the test gene. Considering a training dataset which consists of expression profiles of N genes $\boldsymbol{x}_i$, where each M dimensional expression profile $\boldsymbol{x}_i = x_i(n)$, $n = 1, \ldots, M$ is associated with a teacher value or class label. Given a discriminant function $f(\boldsymbol{x}) = f(\boldsymbol{w}, \boldsymbol{x})$, the objective is to find an M dimensional decision vector $\boldsymbol{w} = [w_1 \;\; w_2 \;\ldots\; w_M]^T$ so that $f(\boldsymbol{x}_i)$ can best match with teacher value y_i, with all the training dataset taken into consideration. Considering 2 class problem with teacher values +1 and –1, in the basic SVM, all the training vectors $\boldsymbol{x}_i$ satisfy the following inequalities:

$$\begin{aligned} \boldsymbol{w}^T \boldsymbol{x}_i + b \geq +1, \text{ for all positive } \boldsymbol{x}_i \\ \boldsymbol{w}^T \boldsymbol{x}_i + b \leq -1, \text{ for all negative } \boldsymbol{x}_i \end{aligned} \qquad (2)$$

An error term is defined as $\varepsilon_i \equiv \boldsymbol{w}^T \boldsymbol{x}_i + b - y_i$. The main objective here is to create a maximum margin to separate the two opposite classes. Maximization of the separation margin $2/\|\mathbf{w}\|$ can be achieved by minimizing $\|\mathbf{w}\|$. Apart from this, considering a set of slack variables $\{\xi_i\}_{i=1}^N \geq 0$, the optimization formula can be written as

$$\min_{\mathbf{w}} \left\{ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \right\}, \text{ subject to } y_i \varepsilon_i + \xi_i \geq 0 \qquad (3)$$

The above quadratic programming optimization problem is solvable by using convex optimization techniques, more specifically, by using a Lagrangian one can obtain the following Wolfe dual-optimization formulation in terms of empirical vector $\boldsymbol{a}$

$$\max_a \quad L(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{y} - \frac{1}{2}\boldsymbol{a}^T \boldsymbol{K}\boldsymbol{a}, \tag{4}$$

subject to $\sum_{i=1}^{N} a_i = 0,\ \boldsymbol{w} = \sum_{i=1}^{N} a_i \boldsymbol{x}_i\ , 0 \le a_i y_i \le C,\ i = 1,...,N$

where kernel matrix **K** is given by

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x}_1,\mathbf{x}_1) & K(\mathbf{x}_1,\mathbf{x}_2) & \cdots & K(\mathbf{x}_1,\mathbf{x}_N) \\ K(\mathbf{x}_2,\mathbf{x}_1) & K(\mathbf{x}_2,\mathbf{x}_2) & \cdots & K(\mathbf{x}_2,\mathbf{x}_N) \\ \vdots & \vdots & \cdots & \vdots \\ K(\mathbf{x}_N,\mathbf{x}_1) & K(\mathbf{x}_N,\mathbf{x}_2) & \cdots & K(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix} \tag{5}$$

and it is jointly determined by the kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ and the training vectors. After $\boldsymbol{a}$ is learned, the decision boundary is characterized by $f(\boldsymbol{x}) = 0$, where $f(\boldsymbol{x})$ is the discriminant function defined as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{6}$$

It is to be mentioned that the (i, j)-th element of the kernel matrix can be defined as the inner product of the i-th and j-th training vectors. A nonlinear kernel function can also be adopted as the inner product. In the proposed scheme, most widely used radial basis function (RBF) is used as the kernel, which is defined as

$$K(x, y) = \exp\left\{-\frac{\| x - y \|^2}{2\sigma^2}\right\}. \tag{7}$$

It is shown that the kernel approach hinges upon the mapping from the original space to a new representative vector space. Moreover, the number of basis function for the kernel vector space is usually (much) greater than the dimension of the original feature space. This plays a major role facilitating the design of highly discriminant classifiers. This is the major reason why the kernel based approach is usually much more effective for supervised classification.

III. Simulation Results

A. Microarray Gene Expression Data:

Yeast cell cycle microarray experiment was conducted by Spellman *et. al.* [1] to obtain temporal gene expression (α-factor synchronized) data through the measurements of expression levels of 6178 genes from 0 to 119 minute in 7 minute interval. In the study, 104 cell-cycle regulated genes were known using traditional method, among them in most of the cell-cycle studies, 90 genes were chosen, moving out the rest due to either the Spellman's failure to identify them or the missing data point that they had. Among these 90 genes,

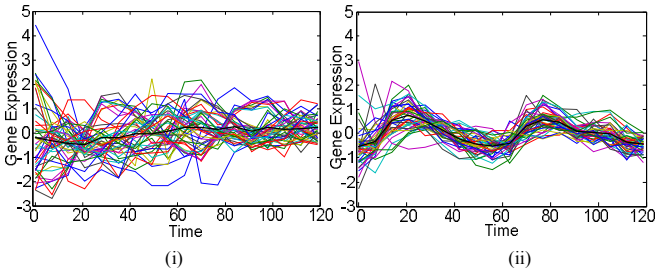


Fig. 3(a). Time varying gene expression profiles with centroid (black bold line). (i) 46 non- G1 phase regulated genes and (ii) 44 G1 phase regulated genes.

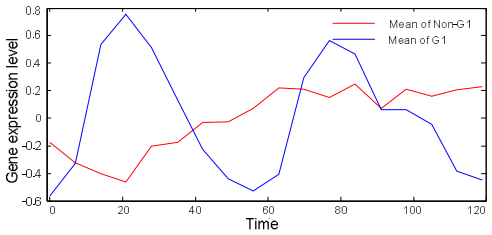


Fig. 3(b). Mean expression levels of G1 and non-G1 groups.

44 are known to G1 cell-cycle regulated and 46 to Non-G1 cell-cycle regulated (i.e. S, S/G2, G2/M and M/G1 phases). In Fig. 3(a), temporal gene expression profiles for genes related to G1 and Non-G1 cell cycle regulation are shown and in Fig. 3(b), the mean expression levels of these two groups are plotted. From the mean curves it is evident that there exists significant amount of between class separability. However, classifiers that exploit only time-course gene expressions often result in larger error rate than it is expected. We intended to propose a method that was able to extract both temporal and spectral information from the time-course gene expression.

B. Quality of the Proposed Feature for the Classification of Time Varying Gene Expression Profiles:

In case of wavelet transform, different mother wavelets have been tested and the results obtained using the 'db4' wavelet is reported in this paper. Since in the gene expression, low frequency contents dominate, only the approximate coefficients are taken into consideration. We also restrict the analysis here only up to level one decomposition. In Fig. 4(a), DWT coefficients corresponding to the genes belonging to each group are shown, where it is observed that wavelet decomposition is more effective in increasing the within class compactness in each individual group. Plots of mean coefficient values for the two groups are shown in Fig. 4(b), which clearly makes the claim stronger by depicting huge difference at each point between the coefficient values of two groups, in comparison to time domain representation shown in Fig. 3. This fact of high between class separation clearly motivates to utilize DWT coefficients as features instead of time domain expression. In the proposed method, first 11 approximate DWT coefficients are used as feature to classify the genes employing support

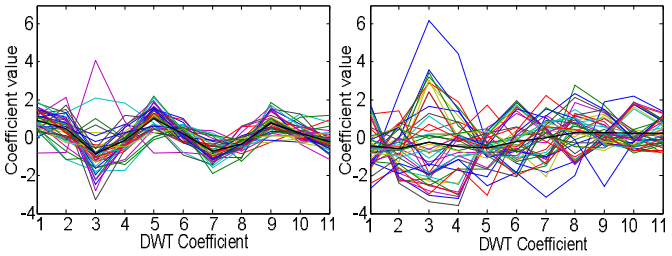


Fig. 4(a). Approximate DWT coefficients corresponding to gene expression profiles with centroid (black bold line). (i) 46 non- G1 phase regulated genes and (ii) 44 G1 phase regulated genes.

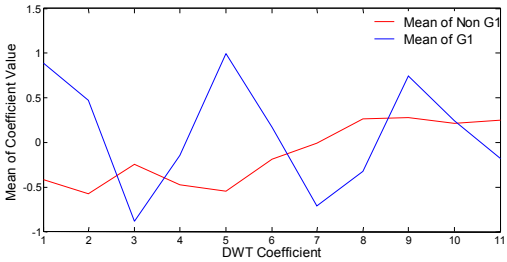


Fig. 4(b). Mean of DWT coefficient values for G1 and non-G1 groups.

vector machine (SVM) through mapping the data feature into kernel space with radial basis function (RBF) kernel. Average error rate of classification is found as 6.67%. Apart from SVM, the classification performance is also tested by using some other widely used classifiers, such as linear discriminant analysis (LDA) and Euclidian distance classifier (EDC). In Table 1, the classification performance obtained by using the proposed DWT feature is compared with that obtained by using directly the microarray (MA) temporal gene expression data. Effect of all three types of classifiers mentioned above is also demonstrated. It is observed from Table 1 that in comparison to microarray temporal data, the proposed DWT features provide better classification performance irrespective of the type of classifiers. Especially, the SVM classifier with RBF kernel provides the best classification performance.

C. Performance Comparison with Some Existing Methods Using Time Varying Gene Expression Profiles :

The classification performance of the proposed method is compared with that of some existing methods reported in [5]-[10]. For the purpose of performance evaluation, leave one out cross validation technique is used considering 90 cell cycle regulated genes with confirmed class labels [1]. In Table 2, the error rate (in %) obtained by using different methods are reported. It is clearly observed that the proposed scheme offers a very low error rate of 6.67%.

D. Classification of Genes Without Having Biologically Proven Cell-cycle Information:

Spellman *et. al.* [1] identified 800 cell cycle regulated genes by clustering cell-cycle genes using peak time of gene expression method. Among those 800 genes, we selected 380 genes for the purpose of testing considering a train set

TABLE 1
CLASSIFICATION ERROR RATES FOR TEMPORAL AND WAVELET FEATURES

Classifier	MA	DWT
SVM_RBF	16.67	6.67
LDA	16.67	14.44
EDC	15.56	13.33

TABLE 2
OVERALL CLASSIFICATION ERROR OBTAINED IN DIFFERENT METHODS

Classification Methods	Overall Error
KIRF [7]	7.8
Logistic [5]	10
B-spline [10]	11.1
PCA on kernel [9]	20
LWV [6]	10
FPCA [8]	7.69
Proposed Method	6.67

TABLE 3
OVERALL CLASSIFICATION ERROR OBTAINED FOR GENES WITHOUT HAVING BIOLOGICALLY PROVEN CELL-CYCLE INFORMATION

Classifier	MA	DWT
SVM_RBF	8.95	6.05
LDA	15.26	9.21
EDC	11.58	12.1

consisting of 90 known genes. Classification error is measured with respect to the class labels declared by Spellman and shown in Table 3. It is found that the proposed DWT feature with kernel SVM classifier provides the best performance. It is to be mentioned that Spellman's class labels are not biologically tested except for 90 genes. Thus, such a small error of 6.05% is acceptable, especially in case of classification of genes without having biologically proven cell cycle phase information. Hence the proposed scheme opens up new pathways in medical research for pattern recognition and classification of genomic data.

IV. CONCLUSION

Unlike most of the gene classification methods, instead of performing some sort of functional data analysis on the time domain microarray data, we propose to employ wavelet transform of the data and use approximate DWT coefficients as features. It is observed that for microarray gene expression data, DWT approximate coefficients exhibit better feature quality, especially high intra-class compactness and inter-class separability. In the proposed method, SVM classifier with RBF kernel is employed, which provides better classification performance in comparison to the distance based or LDA classifiers. The classification performance is tested not only on genes having biological ground truth but also on some genes available in Spellman's database without having the biologically confirmed phase information. In the leave one out cross validation error analysis, it is found that the proposed method outperforms all the existing methods in terms of overall classification accuracy and offers error pattern which is consistent and logical.

REFERENCES

[1] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the Yeast Saccharomyces Cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.

[2] W. Zhao, E. Serpedin, and E. R. Dougherty, "Identifying genes involved in cyclic processes by combining gene expression analysis and prior knowledge," *EUPASHIP Journal on Bioinformatics and Systems Biology*, (*online*) vol. 2009.

[3] J. Kim and H. Kim, "Clustering of change patterns using Fourier coefficients", Bioinformatics, Vol. 24 no. 2 2008, pages 184- 191.

[4] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65-73, 1998.

[5] X. Leng and H.-G. Muller, "Classification using functional data analysis for temporal gene expression data," *Bioinformatics*, vol. 22, no. 1, pp. 68-76, 2006.

[6] M. M. R. Hidalgo and M. D. R. Medina, "Local wavelet-vaguelette-based functional classification of gene expression data," *Biometrical Journal*, vol. 54, no. 1, pp. 1-19, 2012.

[7] J. Cao and J. Wang, "Functional data classification for temporal gene expression data with kernel-induced random forests," *in proc.* Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), May, 2010, pp. 1-5.

[8] J. J. Song, W. Deng, H.-J. Lee, and D. Kwon, "Optimal classification for time-course gene expression data using functional data analysis," *Computational Biology and Chemistry*, vol. 32, pp. 426-432, 2008.

[9] F. Ferraty and P. Vieu, *Nonparameric Functional Data Analysis,* Springer, New York, 2006.

[10] J. Rice and C. O. Wu, "Nonparametric mixed effects models for unequally sampled noisy curves," *Biometrics*, vol. 57, pp. 253-259, 2001.

[11] J. Quin, D. P. Lewis, and W. S. Noble, "Kernel hierarchical gene clustering from microarray expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2097-2104, 2003.