Similarity Matches of Gene Expression Data Based on Wavelet Transform

Mong-Shu Lee, Mu-Yen Chen, and Li-Yu Liu

Department of Computer Science & Engineering,
National Taiwan Ocean University,
Keelung, Taiwan R.O.C.
{mslee,chenmy,M93570030}@mail.ntou.edu.tw

Abstract. This study presents a similarity-determining method for measuring regulatory relationships between pairs of genes from microarray time series data. The proposed similarity metrics are based on a new method to measure structural similarity to compare the quality of images. We make use of the Dual-Tree Wavelet Transform (DTWT) since it provides approximate shift invariance and maintain the structures between pairs of regulation related time series expression data. Despite the simplicity of the presented method, experimental results demonstrate that it enhances the similarity index when tested on known transcriptional regulatory genes.

Keywords: Wavelet transform, Time series gene expression.

1 Introduction

Time series data, such as microarray data, have become increasingly important in numerous applications. Microarray series data provides us with a possible means for identifying transcriptional regulatory relationships among various genes. To identify such regulation among genes is challenging because these gene time series data result from complex activation or repressed exertion of proteins. Several methods are available for extracting regulatory information from the time series microarray data including simple correlation analysis [5], edge detection [7], the event method [13], and the spectral component correlation method [15]. Among these approaches, correlation-based clustering is perhaps the most popular one for this purpose in this occasion. This method utilizes the common Pearson correlation coefficient to measure the similarity between two expression series profiles and to determine whether or not two genes exhibit a regulatory relationship. Four cases are considered in the evaluation of a pair of similar time series expression data.

(1) **Amplitude scaling:** two time series gene expressions have similar waveform but with different expression strengths.
(2) **Vertical shift:** two time series gene expressions have the same waveform but the difference between their expression data is constant.
(3) **Time delay** (horizontal shift)**:** A time delay exists between two time series gene expressions.
(4) **Missing value** (noisy)**:** Some points are missing from the time series data because of the noisy nature of microarray data.

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 540–549, 2009.

Generally, the similarity in cases (1) and (2) can typically be solved by using the Pearson correlation coefficient (and the necessary normalization of each sequence according to its mean). However, the time delay problem caused by the regulatory gene on the target gene significantly degrades the performance of the Pearson correlation-based approach.

Over the last decade or so, the discrete wavelet transform (DWT) has been successfully adopted to various problems of signal and image processing, including data compression [20], image segmentation [17], and ECG signal classification [9]. The wavelet transform is fast, local in the time and the frequency domain, and provides multi-resolution analysis of real-world signals and images. However, the DWT also has some disadvantages that limit its range of applications. A major problem of the common DWT is its lack of shift invariance, which is such that, on small shifts, the input signal can abruptly vary in the distribution of energy between wavelet coefficients on various scales. Some other wavelet transforms have been studied recently to solve these problems, such as the over-complete wavelet transform which discards all down-sampling in DWT to ensure shift invariance. Unfortunately, this method has a very large computational cost that is often not desirable in applications. Several authors [6, 19] have proposed that in a formulation in which two dyadic wavelet bases form a Hilbert transform pair, the DWT can provide the answer to some of the aforementioned limitations. As an alternative, The Kingsburg's dual-tree wavelet transform (DTWT) [11, 12] achieves approximate shift invariance and has been applied to motion estimation [18], texture synthesis [10] and image denoising [24].

Wavelets have recently been used in the similarity analysis of time series because they can extract compact feature vectors and support similarity searches on different scales [3]. Chan and Fu [2] proposed an efficient time series matching strategy based on wavelets. The Haar wavelet transform is first applied and the first few coefficients of the transform sequences are indexed in an R-tree for similarity searching. Wu et al. [23] comprehensively compared DFT (discrete Fourier transform) with DWT transformations, but only in the context of time series databases. Aghili et al. [1] examined the effectiveness of the integration of DFT/DWT for sequence similarity of biological sequence databases.

Recently, Wang et al. [22] have developed a measure of structure similarity (SSIM) for evaluating image quality. The SSIM metrics models perception implicitly by taking into accounts high-level HVS (human visual system) characteristics. The simple SSIM algorithm provides excellently predicting the quality of various distorted images. The proposed approach to comparing similar time series data is motivated by the fact that the DTWT provides shift invariance, enabling the extracting the global shape of the data waveform, and therefore, such measures are to catch the structural similarity between time series expression data. The goal of this study is to extend the current SSIM approach to the dual-tree wavelet transform domain, and base it on a similarity metrics, creating the dual-tree wavelet transform SSIM. This work reveals that the DTWT-SSIM metric can be used for matching gene expression time series data. The regulation-related gene data are modelled by the familiar scaling and shifting transformations, indicating that the introduced DTWT-SSIM index is stable under these transformations. Our experimental results show that the proposed similarity measure outperforms the traditional Pearson correlation coefficient on Spellman's yeast data set.

In Section 2, we briefly give some background information about DWT and DTWT. In section 3, we present our proposed method for the DTWT based similarity measure. We then describe the sensitivity of the DTWT-SSIM metric under the general linear transformation. Experimental tests on a database of gene expression data, and comparison with the Pearson correlation are reported in Section 4. This demonstrates that our results are similar to the spectral component correlation method [15]. Finally, we draw the conclusions of our work in Section 5.

2 Dual-Tree Wavelet Transform

As shown in Fig. 1, in the one-dimensional DTWT, two real wavelet trees are used, each capable of perfect reconstruction. One tree generates the real part of the transform and the other one is used to generate the complex part. In Fig. 1, $h_0(n)$ and $h_1(n)$ are the low-pass and high-pass filters, respectively, of a Quadrature Mirror Filter (QMF) pair in the analysis branch. In the complex part, $\{g_0(n), g_1(n)\}$ is another QMF pair in the analysis branch. All filter pairs considered here are orthogonal and real-valued. Each tree yields a valid set of real DWT detail coefficients u_i and v_i, and altogether form the complex coefficients $d_i = u_i + jv_i$. Similarly, Sa_i and Sb_i is the pair of scaling coefficients of the DWT, as shown in Fig. 1.

A three-level decomposition of DTWT and DWT is applied to the test signal $T(n)$ and its shifted version $T(n-3)$, shown in Fig. 2(a) and (b), respectively, to demonstrate the shift invariance property of DTWT. Fig. 2(c) and (e) show the reconstruction signals $T(n)$ from the wavelet coefficients on the third level of the DWT and DTWT, respectively. Fig. 2(d) and (f) show the counterparts of the shifted signal $T(n-3)$. Comparing Figs. 2(a), (c), and (e) with Figs. 2(b), (d), and (f), they indicate that the shape of the DTWT-reconstructed signal remains mostly unchanged. However, the shape of the DWT-reconstructed signal varies significantly. These results clearly illustrate the characteristics of the shift invariance of the dual-tree wavelet transform. This property helps to simplify some applications.

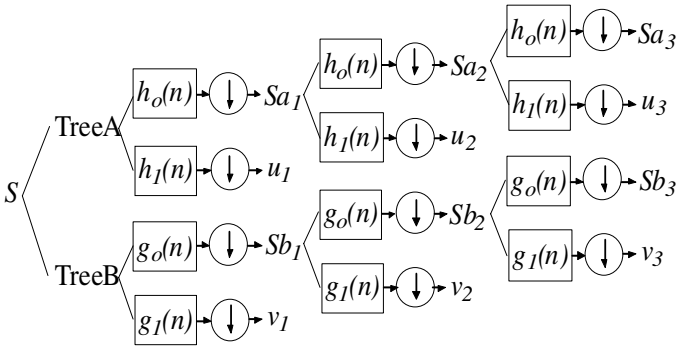

Fig. 1. Kingsbury's Dual-Tree Wavelet Transform with three levels of decomposition

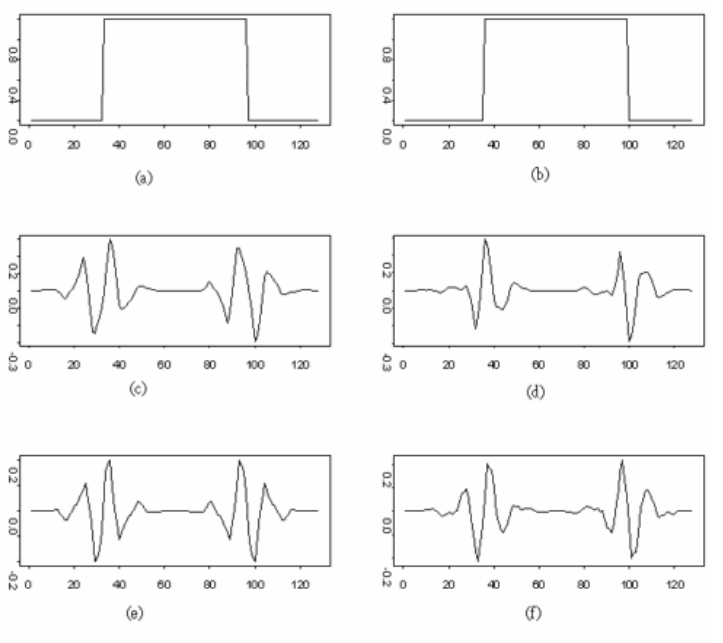

Fig. 2. (a) Signal T(n). (b) Shifted version of (a), T(n-3). (c), (d) are the reconstructed signals using the level 3 DWT coefficients of (a) and (b), respectively. (e), (f) are the reconstructed signals using the level 3 DTWT coefficients of (a) and (b), respectively.

3 DTWT-SSIM Measure

3.1 DTWT-SSIM Index

The proposed application of the DTWT to evaluate the similarity among time series data is inspired by the success of the spatial domain structural similarity (SSIM) index algorithm in image processing [22]. The use of the SSIM index to quantify image quality has been studied. The principle of the structural approach is that the human visual system is highly adapted and can extract structural information (about the objects) from a visual scene. Hence, a metric of structure similarity is a good approximation of a similar shape in time series data. In the spatial domain, the SSIM index that quantizes the luminance, contrast and structure changes between two image patches $x = \{ x_i \mid i = 1, \ldots, M \}$ and $y = \{ y_i \mid i = 1, \ldots, M \}$, and is defined as [22]

$$S(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1}$$

where C_1 and C_2 are two small positive constants; $\mu_x = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\sigma_x^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \mu_x)^2$, and $\sigma_{xy} = \frac{1}{M}\sum_{i=1}^{M}(x_i - \mu_x)(y_i - \mu_y)$.

μ_x and σ_x can be treated roughly as estimates of the luminance and contrast of x, while σ_{xy} represents the tendency of x and y to vary together. The maximum SSIM index value equals one if and only if x and y are identical.

A major shortcoming of the spatial domain SSIM algorithm is that it is very sensitive to translation, and the scaling of signals. The DTWT is approximately shift-invariant. Accordingly, the similarity between the global shapes of related time series data can be extracted by comparing their DTWT coefficients. Therefore, an attempt is made to extend the current SSIM approach to the dual tree wavelet transform domain and make it insensitive to "non-structure" regulatory distortions that are caused by the activation or repression of the gene series data.

Suppose that in the dual tree wavelet transform domain, $d_x = \{d_{x,i} \mid i = 1, 2, ..., N\}$ and $d_y = \{d_{y,i} \mid i = 1, 2, ..., N\}$ are two sets of the DTWT wavelet coefficients extracted from one fixed decomposition level of the expression series data x and y. Now, the spatial domain SSIM index in Eq. (1) is naturally extended to a DTWT domain SSIM as follows.

$$DTWT - SSIM(x, y) = \frac{(2\mu_{d_x}\mu_{d_y} + K_1)(2\sigma_{d_x d_y} + K_2)}{(\mu_{d_x}^2 + \mu_{d_y}^2 + K_1)(\sigma_{d_x}^2 + \sigma_{d_y}^2 + K_2)}$$

$$= \frac{\left(2\mu_{|d_x|}\mu_{|d_y|} + K_1\right)\left(\left(2\sum_{i=1}^{N}(|d_{x,i}| - \mu_{|d_x|})(|d_{y,i}| - \mu_{|d_y|})\right) + K_2\right)}{\left((\mu_{|d_x|})^2 + (\mu_{|d_y|})^2 + K_1\right)\left(\left(\sum_{i=1}^{N}(|d_{x,i}| - \mu_{|d_x|})^2 + \sum_{i=1}^{N}(|d_{y,i}| - \mu_{|d_y|})^2\right) + K_2\right)}$$

$$= \frac{\left(2\sum_{i=1}^{N}(|d_{x,i}|)(|d_{y,i}|)\right) + K_2}{\left(\sum_{i=1}^{N}(|d_{x,i}|)^2 + \sum_{i=1}^{N}(|d_{y,i}|)^2\right) + K_2}. \quad (2)$$

The third equality in Eq. (2) derives from the fact that the dual-tree wavelet coefficients of x and y are zero mean ($\mu_{|d_x|} = \mu_{|d_y|} = 0$), because the DTWT coefficients are normalized after the time series gene data taking DTWT. Herein $|d_x| = |d_{x,i}|$ denotes the magnitude (absolute value) of the complex numbers $d_x = d_{x,i}$, and K_1, K_2 are two small positive constants to avoid instability when the denominator is very close to zero. (We have $K_1 = K_2 = 0.3$ in the experiment).

3.2 Sensitivity Measure

The linear transformation is a convenient way to model the regulation-related gene expression that was described in the Introduction section. The general linear transformation is commonly written in the vector notations with coordinates in the $\mathbb{R}^n$. Now, the scaling and shifting (including vertical and horizontal) relationships that follow from regulation is described in terms of matrices and the following coordinate system as follow.

Let $x = [x_1, x_2, ..., x_n]$ and $y = [y_1, y_2, ..., y_n]$ be two gene expression data, we define $y = Ax + B$ by

$$[y_1, y_2, ..., y_n]^T = A[x_1, x_2, ..., x_n]^T + B^T$$

where matrix $A = [a_{ij}]_{i,j=1}^n$ and vector B specify the desired relation. For example, by defining $A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ and $B = [b_1, b_2, \quad \cdots, \quad b_n]$, this transformation can carry out vertical shifting.

Similarly, the scaling operation is $A = \begin{bmatrix} r & 0 & \cdots & 0 \\ 0 & r & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & r \end{bmatrix}$, $B = [0, 0, \quad \cdots, \quad 0]$.

The condition number $\kappa(A)$ denotes the sensitivity of a specified linear transformation problem. Define the condition number $\kappa(A)$ as $\kappa(A) = \| A \|_\infty \| A^{-1} \|_\infty$, where A is a $n \times n$ matrix and $\| A \|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} | a_{ij} |$.

For a non-singular matrix, $\kappa(A) = \| A \|_\infty \| A^{-1} \|_\infty \ge \| A \cdot A^{-1} \|_\infty = \| I \|_\infty = 1$. Generally, matrices with a small condition number, $\kappa(A) \cong 1$, are said to be well- conditioned. Clearly, the scaling and shifting transformation matrices are well-conditioned. Furthermore, the composition matrix of these well-conditioned transformations still satisfies $\kappa(A) \cong 1$. Let A_1 and A_2 be two such transformations. Applying $\kappa(A_1 A_2) \le \kappa(A_1)\kappa(A_2)$, we establish that the composition of two such transformations also satisfies $\kappa(A_1 A_2) \cong 1$. Fig. 3 and Table 1 present an example comparison of the stability of DTWT-SSIM index and Pearson coefficient under shifting and scaling transformations. Figure 3 shows the original waveform SIN and some distorted SIN waveforms with various scaling and shifting factors. The similarity index between the original SIN and the distorted SIN waveforms is then evaluated using the

proposed DTWT-SSIM and Pearson-correlated metrics. The results presented in Table 1 reveal that except in the scaling case, the DTWT-SSIM unlike the Pearson metric, which decreases sharply, is more stable than the Pearson metric, due to its steady decrease as distortion increases.

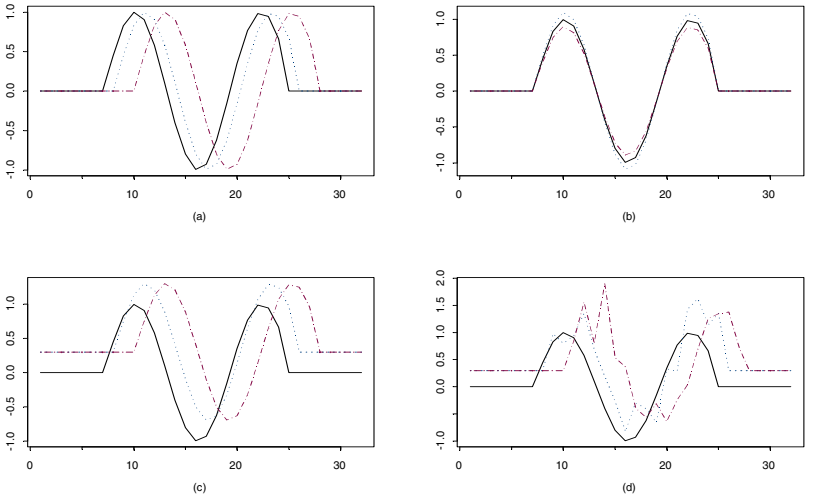

Fig. 3. Original signal SIN (the solid line) and distorted SIN signals with various scaling and shifting factors (the dashed lines). (a) The horizontal shift factors are 1 and 3 units, respectively. (b) The scaling factors are 0.9 and 1.1 respectively. (c) H. shift factor 1 unit + V. shift 0.3 units and H. shift factor 3 units + V. shift 0.3 units. (d) H. shift factor 1 unit + V. shift 0.3 units + noise and H. shift factor 3 units + V. shift0.3 units + noise. (H: Horizontal, V: Vertical)

4 Test Results

A time series expression data similarity comparison experiment was performed using the regulatory gene pairs from [4] and [21], to demonstrate the efficiency of SSIM in the DTWT domain. The gene pairs are extracted by a biologist from the Cho and Spellman alpha and cdc28 datasets. Filkov et al. [8] formed a subset of 888 known transcriptional regulation pairs, comprising 647 activations and 241 inhibitions. The data set is available from the web site at http://www.cs.sunysb.edu/~skiena/gene/jizu/. The alpha data set used in this experiment, contained 343 activations and 96 inhibitions. After all the missing data (noise) were replaced by zeros, the known regulation subsets were analyzed using the proposed algorithm.

The Q-shift version of the DTWT, with three levels of decomposition, was applied to the gene pair to be compared, to evaluate the DTWT-SSIM measure and thus determine gene similarity. The amount of energy is well-known to increase toward the low frequency sub-bands after decomposing the original data into several sub-bands with general wavelet transforms. Therefore, the DTWT-SSIM index was calculated in Eq. (2) using only the lowest sub-band and sequence of normalized wavelet coefficients.

The traditional Pearson correlation and DTWT-SSIM analysis were performed on each pair of 343 known regulations. The proposed DTWT-SSIM method was able to detect many regulatory pairs that were missed by the traditional correlation method due to small correlation value. Numerous visually dissimilar gene pairs have a high DTWT-SSIM index. Table 2 plots the distribution of the two similarity index among the 343 known regulatory pairs. The result demonstrates that less than 11% (36/343) had a Pearson coefficient greater than 0.5 between the activator and activated. However, the DTWT-SSIM index increases the similarity between the known activating relationships by up to 57% (198/343), and the ratio is very close to the result of the spectral component correlation method [15].

Table 1. Similarity comparisons between the original SIN and the distorted SIN waveforms in Fig. 3 using DTWT-SSIM and Pearson metrics

Various scaling and shifting factors in Fig. 3		Pearson coefficient	DTWT-SSIM index
Fig. 3(a)	H. shift 1 unit	0.8743	0.974
	H. shift 3 units	0.1302	0.7262
Fig. 3(b)	Scaling factor: 0.9	1	0.9945
	Scaling factor: 1.1	1	0.9955
Fig. 3(c)	H. shift 1 unit +V. shift 0.3 units	0.8743	0.974
	H. shift 3 units +V. shift 0.3 units	0.1302	0.7263
Fig. 3(d)	H. shift 1 unit +V. shift 0.3 units+ noise	0.8897	0.952
	H. shift 3 units +V. shift 0.3 units+ noise	0.2086	0.5755

Table 2. The cumulative distribution of Pearson and DTWT-SSIM similarity measures among the 343 pairs

Similarity index range	Pearson	DTWT-SSIM
> 0	173/343	336/343
> 0.25	97/343	291/343
> 0.33	72/343	265/343
> 0.5	36/343	198/343
> 0.75	9/343	107/343

The number of false dismissals that occurred in the experiment is considered to determine the effectiveness of these two similarity metrics. If the margin of DTWT-SSIM and the Pearson metrics of the pair expression data exceed 0.5, then the Pearson coefficient is regarded as a false dismissal. For instance, the DTWT-SSIM index of the gene pair is highly correlated with each other but the Pearson metric is negative or lowly correlated. Similarly, if the margin of the Pearson and DTWT-SSIM metrics of

the pair expression data exceed 0.5, then the DTWT-SSIM index is regarded as a false dismissal. 177 out of 343 pairs are false dismissals, based on the Pearson coefficient, while only two out of 343 pairs are false dismissals, based on the DTWT-SSIM.

5 Conclusion

This study presented a new similarity metric, called the DTWT-SSIM index, which not only can be easily implemented but also can enhance the similarity between activation pairs of gene expression data. The traditional Pearson correlation coefficient does not perform well with gene expression time series because of time shift and noise problems. In our dual-tree wavelet transform-based approach, the shortcoming of the space domain SSIM method was avoided by exploiting the almost shift-invariant property of DTWT. This effectively solves the time shift problem. The proposed DTWT-SSIM index was demonstrated to be more stable than the Pearson correlation coefficient when the signal waveform underwent scaling and shifting. Therefore, the DTWT-SSIM measure captures the shape similarity between the time series regulatory pairs. The concept is also useful for other important image processing tasks, including image matching and recognition [16].

References

[1] Aghili, S.A., Agrawal, D., Abbadi, A.: Sequence similarity search using discrete Fourier and wavelet transformation techniques. International Journal on Artificial Intelligence Tools 14(5), 733–754 (2005)
[2] Chan, K.P., Fu, A.: Efficient time series matching by wavelets. In: ICDE, pp. 126–133 (1999)
[3] Chiann, C., Morettin, P.: A wavelet analysis for time series. Journal of Nonparametric Statistics 10(1), 1–46 (1999)
[4] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
[5] Eisen, M.B., Spellman, P.T., Brown, P.O.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 96(19), 10943–10943 (1999)
[6] Fernandes, F., Selesnick, I.W., Spaendonck, V., Burrus, C.S.: Complex wavelet transforms with allpass filters. Signal Processing 83, 1689–1706 (2003)
[7] Filkov, V., Skiena, S., Zhi, J.: Identifying gene regulatory networks from experiomental data. In: Proceeding of RECOMB, pp. 124–131 (2001)
[8] Filkov, V., Skiena, S., Zhi, J.: Analysis techniques for microarray time-series data. Journal of Computational Biology 9(2), 317–330 (2002)
[9] Froese, T., Hadjiloucas, S., Galvao, R.K.H.: Comparison of extrasystolic ECG signal classifiers using discrete wavelet transforms. Pattern Recognition Letters 27(5), 393–407 (2006)
[10] Hatipoglu, S., Mitra, S., Kingsbury, N.: Image texture description using complex wavelet transform. In: Proc. IEEE Int. Conf. Image Processing, pp. 530–533 (2000)

[11] Kingsbury, N.: Image Processing with Complex Wavelets. Phil. Trans. R. Soc. London. A 357, 2543–2560 (1999)
[12] Kingsbury, N.: Complex wavelets for shift invariant analysis and filtering of signals. Appl. Comput. Harmon. Anal. 10(3), 234–253 (2001)
[13] Kwon, A.T., Hoos, H.H., Ng, R.: Inference of transcriptional regulation relationships from gene expression data. Bioinformatics 19(8), 905–912 (2003)
[14] Kwon, O., Chellappa, R.: Region adaptive subband image coding. IEEE Transactions on Image Processing 7(5), 632–648 (1998)
[15] Liew, A.W.C., Hong, Y., Mengsu, Y.: Pattern recognition techniques for the emerging field of bioinformatics: A review. Pattern Recognition 38, 2055–2073 (2005)
[16] Lee, M.-S., Liu, L.-Y., Lin, F.-S.: Image Similarity Comparison Using Dual-Tree Wavelet Transform. In: Chang, L.-W., Lie, W.-N. (eds.) PSIVT 2006. LNCS, vol. 4319, pp. 189–197. Springer, Heidelberg (2006)
[17] Liang, K.H., Tjahjadi, T.: Adaptive scale fixing for multiscale texture segmentation. IEEE Transactions on Image Processing 15(1), 249–256 (2006)
[18] Magarey, J., Kingsbury, N.G.: Motion estimation using a complex-valued wavelet transform. IEEE Transactions on Image Processing 46, 1069 (1998)
[19] Selesnick, I.: The design of approximate Hilbert transform pairs of wavelet bases. IEEE Trans. on Signal Processing 50, 1144–1152 (2002)
[20] Shapiro, J.M.: Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans. Signal Proc. 41(12), 3445–3462 (1993)
[21] Spellman, P., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9, 3273–3297 (1998)
[22] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing 13, 600–612 (2004)
[23] Wu, Y., Agrawal, D., Abbadi, A.: A comparison of DFT and DWT based similarity search in time series database. CIKM, 488–495 (2000)
[24] Ye, Z., Lu, C.: A complex wavelet domain Markov model for image denoising. In: Proc. IEEE Int. Conf. Image Processing, pp. 365–368 (2003)