SIMILARITY ANALYSIS OF TIME SERIES GENE EXPRESSION USING DUAL-TREE WAVELET TRANSFORM

Mong-Shu Lee, Li-Yu Liu, and Mu-Yen Chen
Department of Computer Science
National Taiwan Ocean University
Keelung, Taiwan R.O.C.

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 fact that Dual-Tree Wavelet Transform (DTWT) can provide 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.

Index Terms— Time series, Wavelet transforms

1. INTRODUCTION

Time series data, such as microarray data, are increasingly important in numerous applications. Microarray series data provides us 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 the complex activation or repression exertion of proteins. Several methods are available for extracting regulatory information from time series microarray data, including simple correlation analysis [5], edge detection [7], and the event method [12]. Among these approaches, correlation-based clustering is perhaps the most popular one for this purpose. 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.

Generally, the similarity in cases (1) and (2) can be solved easily 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. 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. Several authors [6, 16] 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) [10, 11] achieves approximate shift invariance and has some applications in motion estimation [15] and texture synthesis [9].

Wavelets have been recently 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. [19] comprehensively compared DFT and 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 searching of biological sequence databases.

Recently, Wang et al. [18] 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 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 of 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 based it on a similarity metric, 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.

2. WAVELET TRANSFORM AND SIMILARITY AMONG TIME SERIES DATA

2.1. Dual tree wavelets transform (DTWT)

The DTWT is made by computing two parallel wavelet tree, tree A and tree B which act upon shifted samples of the input so that tree B picks the samples which tree A decimates. This leads to approximate shift invariance rather than the conventional DWT coefficients which are shift sensitive. The DTWT expansion of a signal $f(x)$ is given by

$$f(x) = \sum_k c_\phi(i_0,k)\phi_{i_0,k}(x) + \sum_{i\geq i_0}\sum_k d_\psi(i,k)\psi_{i,k}(x),$$

where $c_\phi(i_0,k)$ and $d_\psi(i,k)$ are the scaling and wavelet coefficients of the DTWT, using dual-tree scaling functions $\phi_{i_0,k}$ and wavelet functions $\psi_{i,k}$, respectively. For simplicity of notation, the wavelet coefficients $d_\psi(i,k)$ of a signal $f(x)$ are denoted as d_x.

2.2. 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 [18]. 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.

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.

Assume that the expression time series data of two genes x and y are represented by two vectors $x = [x_1, x_2, ..., x_n]$ and $y = [y_1, y_2, ..., y_n]$, where n is the number of sampled data points measured along the time axis.

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 is naturally extended to a DTWT domain SSIM as follows [18].

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

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

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

The third equality in Eq. (1) 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 according to their average after the time series gene data taking DTWT. Herein $|d_{x,i}|$ denotes the magnitude (absolute value) of the complex numbers $d_{x,i}$, and K_2 is a small positive constant to avoid instability when the denominator is very close to zero.

2.3. Sensitivity Metric

The linear transformation is a convenient way to model the regulation-related gene expression that was described in the

Introduction section. 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.

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 and vector B specify the desired relation. For example, by defining $A = I_{n\times n}$ (identity matrix) and $B = [b_1, b_2, \quad \cdots, \quad b_n]$, this transformation can carry out vertical shifting. Similarly, the scaling operation is $A = r \bullet I_{n\times n}$ (r : scaling factor) and $B = [0, 0, \quad \cdots, \quad 0]$.

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 $\kappa(A) \cong 1$.

Fig. 1 and Table 1 present example comparison of the stability of DTWT-SSIM index and Pearson coefficient under shifting and scaling transformations. Figure 1 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 is more stable than the Pearson metric, because the DTWT-SSIM index steadily decreases as the distortion increases, unlike the Pearson metric, which decreases sharply.

3. TEST RESULTS

A time series expression data similarity comparison experiment was performed using the regulatory gene pairs from [4] and [17], to demonstrate the efficiency of SSIM measure task 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 alpha data set used in this experiment, contained 343 activations. After all the missing data (noise) were replaced by zeros, the known regulation subsets were analyzed using the proposed algorithm.

The traditional Pearson correlation and DTWT-SSIM analysis were performed on each pair of 343 known regulations. The result demonstrates that less than 11% (36/343) had a Pearson coefficient > 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). Numerous visually dissimilar gene pairs have a high DTWT-SSIM index. For instance, the Pearson correlation coefficient of genes YAL040C and YBR111C is -0.3885, whose time series expressions are shown in Fig. 2(a), but with a DTWT-SSIM index of 0.9796. Figure 2(b) presents the magnitude of lowest sub-band DTWT coefficients of genes YAL040C and YBR111C after the three levels of decomposition have been applied. Genes YAL040C and YBR111C in Fig. 2(b) are easily seen by eye to exhibit a regulatory relation in perspective of the dual-tree wavelet transform domain.

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 low correlated. Similarly, if the margin of the Pearson and DTWT-SSIM metrics of the pair expression data exceeds 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.

4. CONCLUSION

This study presented a new similarity metric, called the DTWT-SSIM index, which not only can be easily implemented but also enhances 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 task, including image matching and recognition [14].

5. REFERENCES

[1] Aghili SA, Agrawal D, and Abbadi A, Sequence similarity search using discrete Fourier and wavelet transformation techniques. INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS 14 (5): 733-754 OCT 2005.

[2] Chan KP and Fu A, Efficient time series matching by wavelets, ICDE: 126-133 1999.

[3] Chiann C and Morettin P, A wavelet analysis for time series, JOURNAL OF NONPARAMETRIC STATISTICS 10 (1): 1-46 1999.

[4] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW, A genome-wide transcriptional analysis of the mitotic cell cycle, MOLECULAR CELL 2 : 65-73, 1998.

[5] Eisen MB, Spellman PT, Brown PO, 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 SEP 14 1999.

[6] Fernandes F, Selesnick I W, Spaendonck V, and Burrus C S, Complex wavelet transforms with allpass filters, SIGNAL PROCESSING 83 : 1689-1706, 2003.

[7] Filkov V, Skiena, S and Zhi, J, Identifying gene regulatory networks from experiomental data, Proceeding of RECOMB,124-131, 2001

[8] Filkov V, Skiena S, and Zhi J, Analysis techniques for microarray time-series data, JOURNAL OF COMPUTATIONAL BIOLOGY 9 (2): 317-330 2002.

[9] Hatipoglu S, Mitra S, and Kingsbury N, Image texture description using complex wavelet transform, Proc. IEEE Int. Conf. Image Processing, 530-533, 2000

[10] Kingsbury N, Image Processing with Complex Wavelets, Phil. Trans. R. Soc. London. A, V. 357, 2543-2560, Sept. 1999.

[11] Kingsbury N, Complex wavelets for shift invariant analysis and filtering of signals, Appl. Comput. Harmon. Anal., vol. 10, no 3, pp. 234-253, May 2001.

[12] Kwon AT, Hoos HH, and Ng R, Inference of transcriptional regulation relationships from gene expression data, BIOINFORMATICS 19 (8): 905-912 MAY 22 2003.

[13] Kwon O and Chellappa R, Region adaptive subband image coding, IEEE Transactions on Image Processing. Volume 7, Issue 5, May 1998 pp. 632 – 648.

[14] Mong-Shu Lee, Li-Yu Liu, and Fu-Sen Lin, Image Similarity Comparison Using Dual-Tree Wavelet Transform, Lecture Notes in Computer Science, Vol. 4319, pp. 189-197, 2006.

[15] Magarey J. and N. G. Kingsbury, Motion estimation using a complex-valued wavelet transform, IEEE TRANSACTIONS ON SIGNAL PROCESSING 46: 1069 1998.

[16] Selesnick I, The design of approximate Hilbert transform pairs of wavelet bases, IEEE Trans. on Signal Processing, vol. 50, pp.1144-1152, Mar 2002.

[17] Spellman P, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, 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.

[18] Wang Z, Bovik A, Sheikh H, and Simoncelli E, Image quality assessment: From error Visibility to structural similarity, IEEE Trans, Image Processing, vol. 13, pp. 600-612, Apr. 2004.

[19] Wu Y, Agrawal D, and Abbadi A, A comparison of DFT and DWT based similarity search in time series database. CIKM : 488-495, 2000.

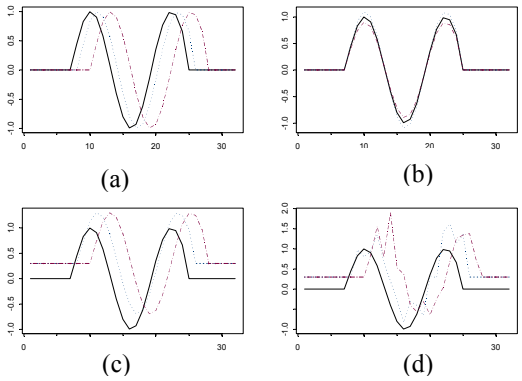

Fig. 1. Original signal SIN (the solid line) and distorted SIN signals with various scaling and shifting factors (the dashed lines). (a) The horizontal shift factors are 1 and 3 units, respectively. (b) The scaling factors are 0.9 and 1.1 respectively. (c) H. shift factor 1 unit + V. shift 0.3 units,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

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

various scaling and shifting factors in Fig. 1		Pearson coefficient	DTWT-SSIM index
Fig. 1(a)	H. shift 1 unit	0.8743	0.974
	H. shift 3 units	0.1302	0.7262
Fig. 1(b)	scaling factor: 0.9	1	0.9945
	scaling factor: 1.1	1	0.9955
Fig. 1(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. 1(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

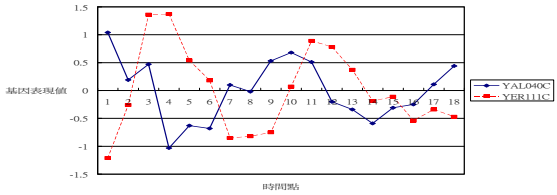

Fig. 2(a). Gene expression data: YAL040C and YER111C.

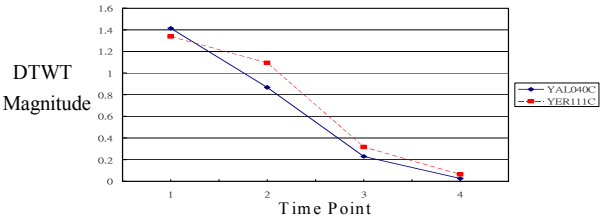

Fig. 2(b). DTWT magnitude of the gene YAL040C and YER111C.