Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 39382, 7 pages
doi:10.1155/2007/39382

Research Article

The Wavelet-Based Cluster Analysis for Temporal Gene Expression Data

J. Z. Song,[1] K. M. Duan,[2] T. Ware,[3] and M. Surette[2]

[1] *Department of Animal and Avian Science, 2413 Animal Science Center, University of Maryland, College Park, MD 20742, USA*
[2] *Department of Microbiology and Infectious Diseases, and Department of Biochemistry and Molecular Biology, Health Sciences Centre, University of Calgary, Calgary, AB, Canada T2N 4N1*
[3] *Department of Mathematics, University of Calgary, Calgary, AB, Canada T2N 4N1*

Received 4 December 2005; Revised 1 October 2006; Accepted 4 March 2007

Recommended by Ahmed H. Tewfik

A variety of high-throughput methods have made it possible to generate detailed temporal expression data for a single gene or large numbers of genes. Common methods for analysis of these large data sets can be problematic. One challenge is the comparison of temporal expression data obtained from different growth conditions where the patterns of expression may be shifted in time. We propose the use of wavelet analysis to transform the data obtained under different growth conditions to permit comparison of expression patterns from experiments that have time shifts or delays. We demonstrate this approach using detailed temporal data for a single bacterial gene obtained under 72 different growth conditions. This general strategy can be applied in the analysis of data sets of thousands of genes under different conditions.

1. INTRODUCTION

High-throughput gene expression techniques, such as oligonucleotide and cDNA microarrays, SAGE (series analysis gene expression), and promoter arrays [1–5], make it possible to obtain large amounts of time series gene expression data in different organisms under various conditions. These large datasets prove to be invaluable for determining coordinately regulated genes and the underlying regulatory networks among genes. Temporal gene expression patterns have been used to define cell cycle regulated genes and metabolic and genetic networks [6–10]. However, how to extract expression patterns in temporal gene expression data represents a challenging analytical problem particularly when comparing data obtained under different growth conditions.

Because high-throughput gene expression technologies involve thousands of genes (or variables), reducing the dimensionality of the data can be a crucial issue for identifying coordinately regulated gene or inferring gene regulation networks. The current solutions include clustering coregulated genes from thousands of genes by similar expression profiles via unsupervised analysis [11–13], and Bayesian networks modeling [14]. Each method has its own merits and shortcomings. In temporal gene expression analysis, a main challenge is to extract the continuous representation of all genes through the time course of the experiment. Aligning gene expression time series profiles based on dynamic time warping [15], hidden Markov model [16], local clustering [17] and fitting time series data with cubic splines [18–20] have been used. However, a significant challenge remains in the comparisons of high-throughput temporal expression profiles obtained from same genes in different experimental conditions where patterns may be shifted in time. The current analysis methods do not specifically address the issue of time delays between experiments or conditions.

Many mathematical and statistical methods have been developed for identifying underlying patterns in complex data with varieties of applications, such as signal classification in speech processing, electrocardiography and sleep research. These methods cluster points in multidimensional space, and are routinely used in gene expression analysis. For example they have been used to identify genes whose expression correlated with the cell cycle [21–23]. These methods are readily applicable to many datasets. However, these strategies have limitations when comparisons of temporal data between different conditions are being carried out. Over the past few years, the wavelet has become an essential tool in

genome analysis [24–27]. In this study, we propose the use of wavelet transformation as a method to characterize structure at multiple positions and length scales. Wavelet transforms are capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the temporal gene expression signals, the wavelet transformed data can be further analyzed by cluster analysis. We demonstrate this approach with temporal expression profiles for a single gene under 72 growth conditions. Clustering of the data after wavelet transformation overcomes the problem of temporal shifts in expression patterns observed under different experimental conditions.

2. MATERIALS AND METHODS

2.1. Gene expression data

Temporal gene expression profiles were obtained using promoter fusion technique. Briefly, the promoters of interest are clones into a promoterless *luxCDABE* operon on a plasmid vector pMS402 [28]. Promoter activity correlates with light production generated by the *luxCDABE* gene products. Therefore, the activity of the promoter fused upstream *luxCDABE* is directly measured as light production after the fusion construct is introduced into the bacterium. The promoter regions of the *Pseudomonas aeruginosa rpoS* gene was amplified from *P. aeruginosa* PAO1 chromosomal DNA by PCR using oligonucleotide primers [28]. The PCR amplified promoter region were then cloned into the *Xho*I-*Bam*HI sites of pMS402 upstream of the promoterless *luxCDABE* genes and transformed into PAO1 by electroporation. PCR, DNA manipulation and transformation were performed following general procedures. Overnight cultures of the reporter strain were diluted 1 : 200 in a 96-well microtiter plate and the promoter activity (CPS) and optical density at 620 nm (OD_{620}) were measured every 30 minutes for 24 hours in a victor^2 multilabel counter. The details of the 72 growth conditions will be described elsewhere.

2.2. Expression data wavelet transformation and clustering analysis

To overcome the gene expression profile shift issue (time delay) among different conditions, we first used continuous wavelet analysis to transform all expression data by wavelet transform; it decomposes temporal gene expression data in both time and frequency domains. In wavelet transform we take a real/complex valued continuous time function with two main properties, (1) it will integrate to zero; (2) it is square integrable. This function is called the mother wavelet. The CWT or continuous wavelet transform of a function $f(t)$ with respect to a wavelet $\psi(t)$ is defined as

$$W(a,b) = \int_{-\infty}^{\infty} f(t)\Psi_{a,b}(t)dt,$$
$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\Psi\frac{t-b}{a}. \quad (1)$$

Here, a and b are real. $W(a,b)$ is the transform coefficient of $f(t)$ for given a, b. Thus the wavelet transform is a function of two variables. For a given b, $\psi_{a,b}(t)$ is a shift of $\psi_{a,0}(t)$ by an amount b along time axis. The variable b represents time shift or translation. Since a determines the amount of time scaling or dilation, it is referred to as scale or dilation variable. If $a > 1$, there is stretching of $\psi(t)$ along the time axis whereas if $0 < a < 1$ there is a contraction of $\psi(t)$. Each wavelet coefficient $W(a,b)$ is a measure of the correlation of the input waveform with a translated and dilated version of the mother wavelet. By investigating the wavelet transform over different bases, we adopted dB2 as the mother function (see Figure 1). The output of the transform shows the correlation between the signal and the wavelet as a function of time across a range of scales. To avoid negative coefficients and in order to display differences clearly, we define

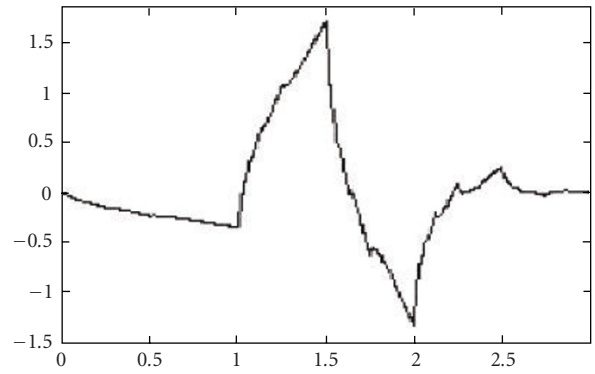

Figure 1: Mother wavelet (dB2).

$$S(a) = W^2(a,b). \quad (2)$$

Based on the squared coefficient $S(a)$, we clustered the 72 conditions with the average linkage method [29]. The distance between two clusters is defined by

$$D_{KL} = \frac{1}{N_K N_L}\sum_{i\in K}\sum_{j\in L} d(x_i, x_j). \quad (3)$$

If $d(x,y) = |x-y|^2$, then

$$D_{KL} = ||\overline{x}_K - \overline{x}_L||^2 + \frac{W_K}{N_K} + \frac{W_L}{N_L}. \quad (4)$$

The combinational formula is

$$D_{JM} = \frac{(N_K D_{JK} + N_L D_{JL})}{N_M}. \quad (5)$$

In average linkage the distance between two clusters is the average distance between pairs of observations, one in each cluster. Average linkage tends to join clusters with small variances, and is slightly biased toward producing clusters with the same variance. All calculation was done by SAS and Matlab.

3. RESULT AND ANALYSIS

3.1. The variation of gene expression profile

A large data set was generated from a unique gene expression experiment where activity of the promoter of the *rpoS*

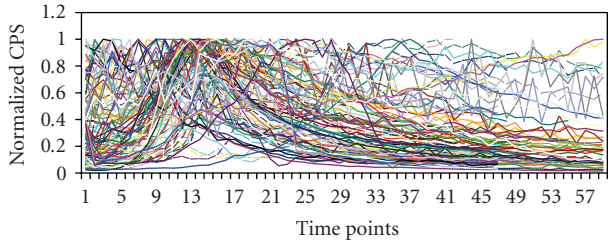

FIGURE 2: The *rpoS* gene expression profiles in 72 conditions and 48 time points. Because the strength of expression of the *rpoS* promoter varies among conditions, the expression levels were normalized for each condition with its maximum so that the expression level is the range between 0 and 1.

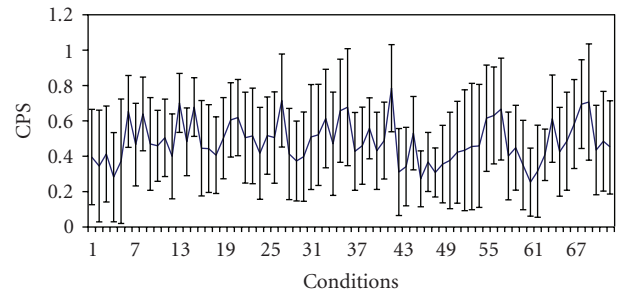

FIGURE 3: The fluctuation of standard deviation of *rpoS* promoter activity in 72 different conditions and 48 time points. The blue line is mean and the error bar is standard deviation of gene in each condition.

gene in *P. aeruginosa* was measured under 72 growth conditions. For each condition, measurements were obtained at 48 time points. Figure 1 shows the expression profile variation of this gene in different experimental conditions. Because the strength of expression of the *rpoS* gene varies among conditions, that is, the expression pattern may be similar although the magnitude of expression may vary, we normalized each expression profile with its maximum, so all expression level is in the range between 0 and 1. As expected, the gene expression profile varies significantly in different experiments and conditions. As shown in Figure 2, the time point of maximum expression of the *rpoS* shifts among conditions, that is, with clear expression profile shift or time delay phenomena. To further evaluate the variation of the *rpoS* promoter activity over 72 conditions, we determined the mean and standard deviation of the gene in each condition. The fluctuation of the mean and standard deviation of expression levels of the *rpoS*, as shown in Figure 3, highlights the variation of expression level and expression strength among conditions. These results clearly show the expression profiles and levels are condition-specific, that is, the regulation of the *rpoS* gene varies in different conditions.

3.2. The wavelet transformation of gene expression profile

Wavelet transformation is an analysis method that uses both time and the frequency domains. It allows a time series to be viewed in multiple resolutions, and each resolution reflects a different frequency. The wavelet technique takes averages and differences of a signal and breaks the signal down into spectrum. In the gene expression analysis, we assume that any gene expression level is a comprehensive result of gene effects and condition effects, that is, the expression profile shift or time delay is caused by the conditions which dictate the activation order and expression strength of the *rpoS* gene. The profile shifts or time delays certainly make comparison of expression patterns among conditions problematic. Overcoming this time delay, the wavelet transform addresses it by using dB2 (Figure 1) mother function that can be scaled. If the signal and wavelet are in a good match, then the correlation between the signal and the wavelet is high, resulting

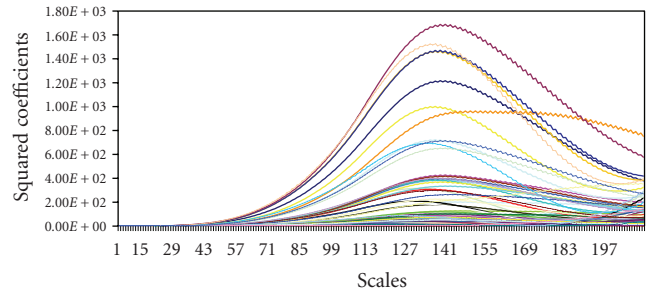

FIGURE 4: The power plot of the wavelet transformation of the *rpoS* gene promoter activity profiles obtained under 72 conditions. The mother wavelet id dB2 and the coefficients of wavelet transformation were squared.

in a large coefficient. The coefficients of wavelet transformation indicate correlation intensities between wavelet function and expression profile if the expression signal level is between 0 and 1. When the wavelet is highly compressed it extracts the localized high-frequency details of the expression signal. When the wavelet is fully diluted, the length of the wavelet is more comparable to the length of the gene expression signal and therefore it extracts the low frequency trends of the signal. In order to overcome the issue in temporal gene expression data analysis we take an approach using wavelet transformation. The transformation results of the gene *rpoS* over 72 conditions, as shown in Figure 4, demonstrate the squared coefficients with a bell-shaped curve, the curves of the different conditions vary in skew and kurtosis which represent the difference of expression profiles. If expression profiles are similar, the bell-shaped curve will be very similar and close; otherwise, they will disperse. The wavelet analysis is able to overcome the profile shift problem, meanwhile, it is worth noting that the analysis loses time series information.

3.3. Clustering analysis and evaluation

To evaluate the behavior of gene expression under different culture conditions, expression profiles are typically compared using cluster analysis. This provides a comparison of

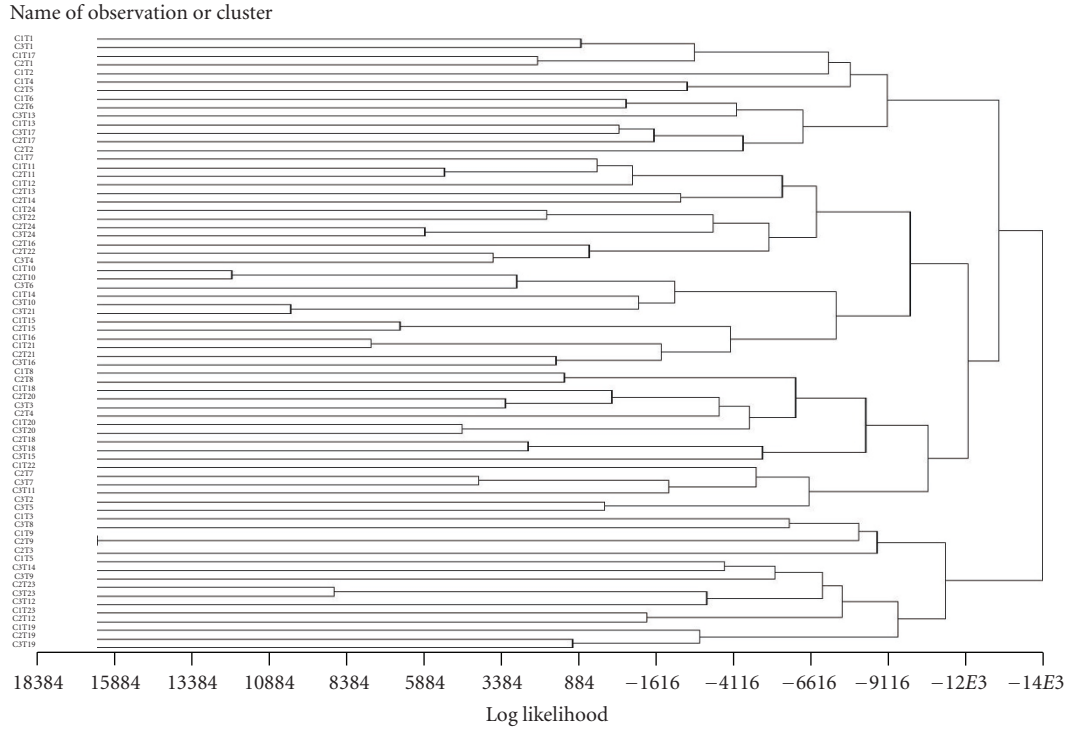

Figure 5: The cluster tree of 72 conditions of the *rpoS* gene expression before wavelet transformation based on the 48 time points measurements.

patterns of expression such that those with similar patterns of expression will fall close together on the hierarchical tree while those with dissimilar patterns will be far apart. To evaluate the effect of wavelet transformation, we clustered the data before and after transformation using average linkage method. The hierarchical cluster trees are shown in Figures 5 and 6. Note that the numerical values used in the two figures differ and consequently the distance measures are not directly comparable.

We would predict that genes with similar expression profiles before wavelet transformation would cluster together in both Figures 5 and 6. Wavelet transformation would not make the expression patterns dissimilar. To illustrate this we have plotted the expression data for two conditions (C1T23 and C2T23) that cluster closely in Figure 7. We can see that the activity profiles of the *rpoS* promoter are very similar in these two conditions (Figure 7(a)) and likewise the power plots of their wavelet transformation are also similar (Figure 7(b)). As expected they also clustered together in Figure 5.

To illustrate the effect of the wavelet transformation, we highlight the expression of two conditions (C1T7 and C2T7) that cluster close together after wavelet transformation (as in Figure 5) but not before it (as in Figure 4). We would predict that these will have similar expression patterns but with a time shift between the experimental conditions. This is clearly illustrated in Figure 8(a). This temporal shift is sufficient to prevent close clustering of these conditions in Figure 4. By contrast, the profiles appear very similar after wavelet transformation (Figure 8(b)) and the two conditions cluster close together in Figure 5. In this experiment, the growth medium used in C1T7 and C2T2 was the same and the expression profile would be expected to match however experimental variables leading to different initial conditions. The results indicate that wavelet transformation can extract expression pattern information and overcome difficulties that arise because of temporal delays in patterns of expressions between conditions or experiments.

4. DISCUSSION

To deeply understand gene temporal expression behavior and interactions in cells is a fundamental task in functional genomics. While methods for obtaining high-throughput temporal gene expression data are readily available, methods and strategies for analysis of these complex data sets are still emerging. Because the unique feature of temporal gene expression data is autocorrelation between successive points, the immediate goals are to extract and to compare the fundamental patterns of gene expression inherent in the data. Most of the current methods are based on certain distances between expressed genes or variables (conditions), such as hierarchical clustering, self-organizing maps, relevance network, principal components analysis and machine learning. Application of clustering analysis directly to the expression data ignores some basic features of temporal expression data and more over can be complicated by temporal shifts or time delays between experiments. These temporal shifts arise not

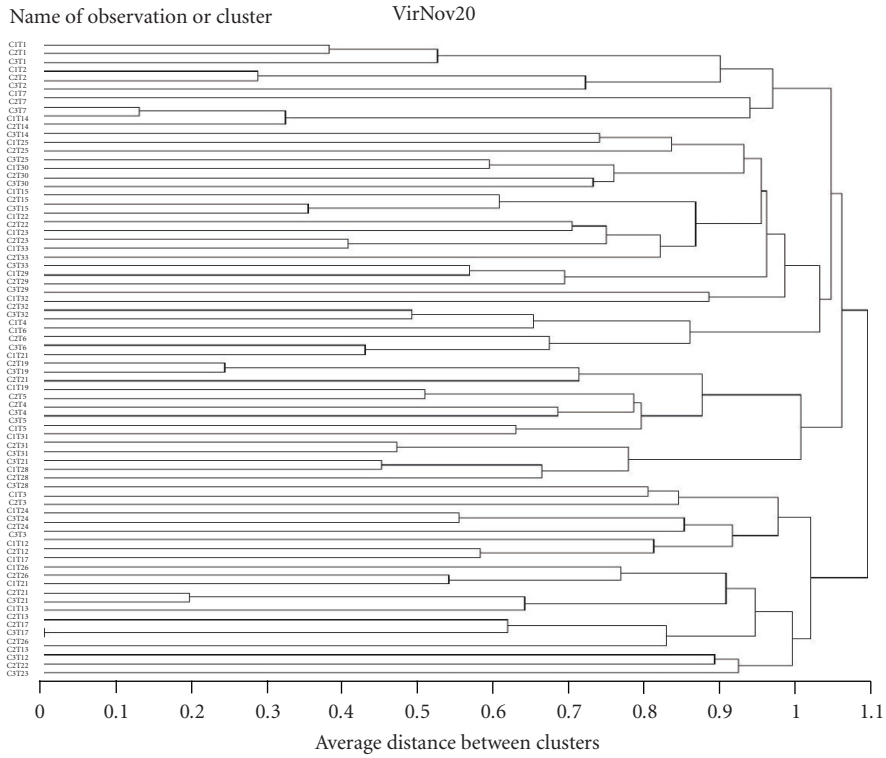

Figure 6: The cluster tree of 72 conditions of the *rpoS* gene expression after wavelet transformation based on the 48 time points measurements.

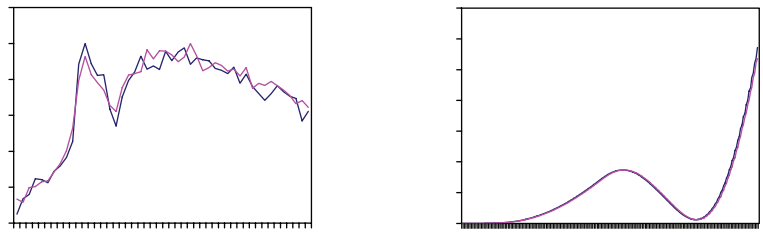

Figure 7: (a) The expression profiles of the *rpoS* in conditions C1T23 and C2T23 and (b) the power of the wavelet transform.

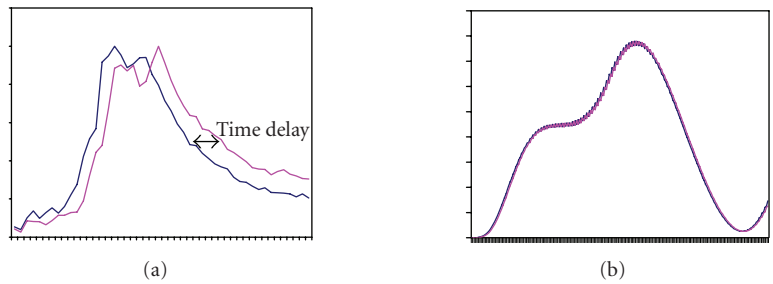

Figure 8: (a) The expression profiles of the *rpoS* in conditions C1T7 and C2T7 and (b) the power of the wavelet transform.

because of intrinsic features of the expression pattern but because of differences in initial conditions between experiments. These are often unavoidable experimental variables. Dynamic time warping is a discrete method similar to sequence alignment algorithms [5] that can be used to align time series data. It involves many degrees of freedom and the time points can "stop" or go "backwards" in the alignment. Overfitting can also be a problem with this method. The cubic spline is a powerful technique for data fitting, interpolation, extrapolation, and visualization [20], and permits the principled estimation of unobserved time-points and dataset alignment. Each temporal gene expression profile is modeled as a cubic spline (piecewise polynomial) that is estimated from the observed expression data. It constrains the spline coefficients of genes in the same cluster to have same or similar expression patterns. The splines are piecewise-smooth polynomials that can be used to represent functions over large intervals, where it would be impractical to use a single approximating polynomial. As for the clustering analysis with the cubic splines, especially in large scale of temporal gene expression data, further research and comparison are needed.

In this paper, we firstly transformed temporal gene expression data with continuous wavelet analysis and then we did hierarchical clustering analysis. Average linkage method was used because it proceeds by first finding pairs of expression profiles that are most similar, joining them, calculating the (sometimes weighted) average between the members of the joined cluster, recalculating the pairwise distance, and treating the average profile as one profile, and repeating the procedure until all profiles are joined. Average linkage clustering can be conducted using all-pairwise-sample average of differences or using cluster average differences. The latter is also known as centroid clustering, but centroids can be calculated using methods other than simple averages.

It is worth noting that wavelet analysis and the Fourier transformation (FT) are two widely used methods in signal processing. In its original form, the FT assumes that the expression signal exists for all time. This for practical purposes is not a realistic assumption in temporal gene expression and does not give any information about how the expression signal changes with respect to time. This is not a problem when the gene expression signal being analyzed is stationary, that is when the statistical properties of the expression signal are not changing with time. All gene expression signals, however, are nonstationary. It is especially necessary to identify and locate the changing frequency characteristics of the gene expression signals. An alternative FT, which is called the short-time Fourier transform (STFT), is a time-dependent or windowed-Fourier transformation. It attempts to analyze nonstationary signals by dividing the whole signal into shorter data frames, but one of the limitations of the STFT is that the timeframe for analysis is fixed. Wavelet transformation is a measure of similarity between the basis functions (wavelets) and gene expression profiles, and the calculated CWT coefficients refer to the closeness of the gene expression profile to the wavelet at the current scale. The flexible approach uses a scalable window. The advantages of the method are a compressed window for analyzing high-frequency details and a diluted window for uncovering low-frequency trends within the signal. Wavelets are also well localized in frequency, although not as well as sinusoids. Since wavelet analysis incorporates the concept of scale into the wavelet equation it is suited to resolve the transient nature of gene expression data.

Then choosing appropriate scales and the number of scales are imminent issues. Scale is the inverse of frequency. Once the mother wavelet is chosen, the computation will start from high frequencies and proceed towards low frequencies. This first value of scale will correspond to the most compressed wavelet. As the value of scale is increased, the wavelet will dilate. Smaller scales (high frequencies) have better scale resolution which corresponds to poorer frequency resolution. Similarly, large scales have better frequency resolution. From the results presented here, it is apparent that wavelets are better suited to the analysis of transient gene expression signals, since they are well localized in time, whereas sinusoids extend over all time. We also need to emphasize that although the wavelet analysis overcomes the time delay or profile shift, the transformation will lose temporal information if we need it, so the analysis is application dependent.

In summary, the paper presents an alternative way to extract expression patterns in temporal gene expression data with continuous wavelet analysis. It has been demonstrated that the application of wavelet transformation to gene temporal expression data is feasible. We anticipate that the wavelet analysis and transformation could be used in large scale temporal gene expression research and single cell experiments. It is of particular value in comparison of temporal expression profiles obtained under different conditions or from different experiments. The pattern recognition is of important value on monitoring simultaneously the expression patterns of thousands of genes during cellular differentiation and responses.

ACKNOWLEDGMENTS

The authors thank members of the Surette lab for helpful discussions. This work was supported by the Canadian Institutes of Health Research, Genome Canada through the University of Saskatchewan. M.G.S. is an Alberta Heritage Foundation for Medical Research Senior Scholar and Canada Research Chair in Microbial Gene Expression.

REFERENCES

[1] S. Kalir, J. McClure, K. Pabbaraju, et al., "Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria," *Science*, vol. 292, no. 5524, pp. 2080–2083, 2001.

[2] A. T. Weeraratna, "Serial analysis of gene expression (SAGE): advances, analysis and applications to pigment cell research," *Pigment Cell Research*, vol. 16, no. 3, pp. 183–189, 2003.

[3] A. Schulze and J. Downward, "Navigating gene expression using microarrays—a technology review," *Nature Cell Biology*, vol. 3, no. 8, pp. E190–E195, 2001.

[4] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, pp. 129–153, 2002.
[5] E. M. Southern, "DNA microarrays: history and overview," *Methods in Molecular Biology*, vol. 170, pp. 1–15, 2001.
[6] A. H. Y. Tong, G. Lesage, G. D. Bader, et al., "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, no. 5659, pp. 808–813, 2004.
[7] N. Fedoroff and W. Fontana, "Genetic networks: small numbers of big molecules," *Science*, vol. 297, no. 5584, pp. 1129–1131, 2002.
[8] R. Bundschuh, F. Hayot, and C. Jayaprakash, "Fluctuations and slow variables in genetic networks," *Biophysical Journal*, vol. 84, no. 3, pp. 1606–1615, 2003.
[9] P. T. Spellman, G. Sherlock, M. Q. Zhang, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.
[10] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.
[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.
[12] N. Banerjee and M. Q. Zhang, "Functional genomics as applied to mapping transcription regulatory networks," *Current Opinion in Microbiology*, vol. 5, no. 3, pp. 313–317, 2002.
[13] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," *FEBS Letters*, vol. 451, no. 2, pp. 142–146, 1999.
[14] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.
[15] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495–508, 2001.
[16] A. Schliep, A. Schönhuth, and C. Steinhoff, "Using hidden Markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, supplement 1, pp. i255–i263, 2003.
[17] J. Qian, M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein, "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1053–1066, 2001.
[18] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola, "Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 18, pp. 10146–10151, 2003.
[19] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.
[20] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "Continuous representations of time-series gene expression data," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 341–356, 2003.
[21] S. Hampson, D. Kibler, and P. Baldi, "Distribution patterns of over-represented *k*-mers in non-coding yeast DNA," *Bioinformatics*, vol. 18, no. 4, pp. 513–528, 2002.
[22] B. Futcher, "Transcriptional regulatory networks and the yeast cell cycle," *Current Opinion in Cell Biology*, vol. 14, no. 6, pp. 676–683, 2002.
[23] R. J. Cho, M. J. Campbell, E. A. Winzeler, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.
[24] P. Liò, "Wavelets in bioinformatics and computational biology: state of art and perspectives," *Bioinformatics*, vol. 19, no. 1, pp. 2–9, 2003.
[25] J. Z. Song, T. Ware, S.-L. Liu, and M. Surette, "Comparative genomics via wavelet analysis for closely related bacteria," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 5–12, 2004.
[26] J. Z. Song, A. Ware, and S.-L. Liu, "Wavelet to predict bacterial *ori* and *ter*: a tendency towards a physical balance," *BMC Genomics*, vol. 4, no. 1, p. 17, 2003.
[27] P. Liò and M. Vannucci, "Finding pathogenicity islands and gene transfer events in genome data," *Bioinformatics*, vol. 16, no. 10, pp. 932–940, 2000.
[28] K. M. Duan, C. Dammel, J. Stein, H. Rabin, and M. Surette, "Modulation of *Pseudomonas aeruginosa* gene expression by host microflora through interspecies communication," *Molecular Microbiology*, vol. 50, no. 5, pp. 1477–1491, 2003.
[29] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansaa Science Bulletin*, vol. 38, pp. 1409–1438, 1958.