The Wavelet Transformation for Temporal Gene Expression Analysis

JZ. Song1* (songj@ucalgary.ca)
KM. Duan1* (kduan@ucalgary.ca)
M. Surette1,2 (surette@ucalgary.ca)

Department of Microbiology and Infectious Diseases1, and Department of Biochemistry and Molecular Biology2, Health Sciences Centre, University of Calgary, Calgary, AB, Canada, T2N 4N1

*: Both authors contributed equally to this work. Corresponding author: M G. SURETTE

Abstract

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. 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 for can be applied in the analysis of data sets of thousands of genes during cellular differentiation and response.

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 pattern 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 co-regulated genes from thousands of genes by similar expression profiles via unsupervised analysis [11-13], and Bayesian networks modeling[14]. Each method has 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 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. A big 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.

To extract the fundamental patterns of gene expression in the temporal data, many mathematical and statistical methods have been developed for identifying underlying patterns in complex data in varieties of applications, such as signal classification in speech processing, electrocardiography and sleep research. The methods cluster points in multidimensionalspace, 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. In this study, we propose the use of wavelet transformation as a method to characterize structure at multiple positions and length scales. Over the past few years, the wavelet has become an essential tool in genome analysis[24-27]. Wavelet transform is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the temporal gene expression signal. Wavelet transformed data can be further analyzed by cluster analysis. Here we demonstrate this approach with temporal expression profiles for a single gene under 72 growth conditions. Clustering of 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 \textit{Pseudomonas aeruginosa} rpoS gene was amplified from \textit{Pseudomonas aeruginosa} PAO1 chromosomal DNA by PCR using oligonucleotide primers [28]. The PCR amplified promoter region were then cloned into the \textit{XhoI-BamHI} sites of pMS402 upstream of the promoterless \textit{lucCDABE} 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 (K. Duan and M.G. Surette, manuscript in preparation)

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

\[ W(\lambda, t) = \int_{-\infty}^{\infty} \Psi_{\lambda,t}(u) x(u) \, du \]

\[ \Psi_{\lambda,t}(u) = \frac{1}{\sqrt{\lambda}} \Psi \left( \frac{u - t}{\lambda} \right) \]

Here, \( W(\lambda, t) \) is the coefficients of the continuous wavelet transformation (CWT) of \( x(u) \), \( \Psi(u) \) is the wavelet, \( \lambda \) is the scale associated with transformation and \( t \) is its location. The CWT compresses or dilates and translates a DB2 "mother" wavelet and correlates it with the gene expression signal at all time and scales. The output of the transform shows the correlation between the signal and the wavelet as a function of time. To avoid negative coefficients and clearly display difference, we make

\[ S(\lambda) = W^2(\lambda, t) \]

\[ R(\lambda) = \sum W^2(\lambda, t) \]

based on the two squared coefficient and sum of squared coefficients, we clustered the 72 conditions with the average linkage method [29]. The distance between two clusters is defined by average linkage. 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

The large data set was generated from a unique gene expression experiment where activity of the promoter of the rpoS gene in \textit{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
standard deviation of expression levels of the rpoS, as shown in the Figure 2, highlights the variation of expression level and expression strength among conditions. These results clearly show the expression profiles and levels are conditional-specific, i.e., the regulation of the rpoS gene varies conditions.

### 3.2. The Wavelet Transformation of Gene Expression Profile

Wavelet transform 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, i.e., the expression profile shift or time delay is caused by the conditions which regulate the activation order and expression strength of the rpoS gene. The profile shifts or time delays certainly make comparison of expression patterns among conditions in problem. Overcoming this time delay, the wavelet transform addresses it by using DB2 mother function that can be scaled. If the signal and wavelet are a good match then the correlation between the signal and the wavelet is high, resulting 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 3, 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 the analysis lose 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 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 the Figure 4 and Figure 5. 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 Figure 4 and 5. Wavelet transformation would not make the expression patterns dissimilar. To illustrate this we have plotted the expression data for two conditions (C2T23 and C3T23) that cluster closely in Figure 4. We can see that the activity profiles of the rpoS promoter are very similar in these two conditions (Figure 6A) and likewise the power plots of their wavelet transformation are also similar (Figure 6B). As expected they also clustered together in the 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 (Figure 5) but not before (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 7A. 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 7B) and the two conditions cluster close together in Figure 5. The results indicate that wavelet transformation can extract expression pattern information and overcome difficulties that arise because of temporal delays in patterns of expression between conditions or experiments.

4. Discussion

To understand how genes behave individually and how the interaction in cells is a fundamental task in functional genomics. While methods for obtaining high throughput temporal gene expression data are readily available, methods for analysis of these complex data sets are still emerging. Because the unique feature of temporal gene expression data is autocorrelation between successive points, one goal is to extract and 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. These can ignore some basic features of the temporal data type, that is time delays or time shifts between experiments. Aligning gene expression time series based on dynamic time warping is a discrete method which is similar to sequence alignment[15], it involves many degrees of freedom, time point can “stop” or “go backward”, the alignment score doesn’t have statistically significant, the drawbacks may lead some overfitting issue. The cubic spline is a powerful technique for data fitting, interpolation, extrapolation, and visualization[20], it 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 smooth piecewise 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 signal processing, wavelet analysis and Fourier transformation (FT) are two widely used methods. In its original form, 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 the statistical properties of the expression signal are not changing with time. All gene expression signals, however, are non-stationary. It is especially necessary to identify and locate the changing frequency characteristics of the gene expression signals. An alternative FT, which is called short-time Fourier transform (STFT), is a time-dependent or windowed-Fourier transformation. It attempts to analyze non-stationary 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 more flexible approach which uses a scalable window. The advantages of the method are a compressed window for analyzing high frequency detail and a diluted window for uncovering low frequency trends within the signal. From the results presented here, it is apparent that wavelets are better suited to analysis transient gene expression signals, since they are well localized in time, whereas sinusoids extend over all timeframe. Wavelets are also well localized in frequency, although not as well as sinusoids. The property of time and frequency localization is known as compact support, since wavelet analysis incorporates the concept of scale into the wavelet equation it is suited to resolve transient of gene expression data.

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

5. Acknowledgement

The authors thank members of the Surette lab for helpful discussions. This work was supported by the CIHR, the CBDN and 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.

6. Reference


