Gene expression

Characterizing the dynamic connectivity between genes by variable parameter regression and Kalman filtering based on temporal gene expression data

Qinghua Cui, Bing Liu, Tianzi Jiang* and Songde Ma

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, People’s Republic of China

Received on July 28, 2004; revised on October 12, 2004; accepted on November 29, 2004

Advance Access publication December 7, 2004

ABSTRACT

Motivation: One popular method for analyzing functional connectivity between genes is to cluster genes with similar expression profiles. The most popular metrics measuring the similarity (or dissimilarity) among genes include Pearson’s correlation, linear regression coefficient and Euclidean distance. As these metrics only give some constant values, they can only depict a stationary connectivity between genes. However, the functional connectivity between genes usually changes with time. Here, we introduce a novel insight for characterizing the relationship between genes and find out a proper mathematical model, variable parameter regression and Kalman filtering to model it.

Results: We applied our algorithm to some simulated data and two pairs of real gene expression data. The changes of connectivity in simulated data are closely identical with the truth and the results of two pairs of gene expression data show that our method has successfully demonstrated the dynamic connectivity between genes.

Contact: jiangtz@nlpr.ia.ac.cn

INTRODUCTION

With the ability to simultaneously measure the activity of thousands of genes under different conditions (Iyer et al., 1999; Eisen et al., 1998; Cho et al., 1998; Spellman et al., 1998; Bozdech et al., 2003), DNA microarray technology has attracted tremendous interest in both the scientific community and industry during the past several years. This has led to a dramatic increase in microarray data in both the scientific community and industry during the past several years. This has led to a dramatic increase in microarray data in both the scientific community and industry during the past several years. This has led to a dramatic increase in microarray data in both the scientific community and industry during the past several years. This has led to a dramatic increase in microarray data in both the scientific community and industry during the past several years.

One of the applications of microarray technology is to characterize the functional connectivity between genes. A basic assumption of this application is that genes with similar expression profiles have similar functions in cells. The most popular metrics used to evaluate the similarity (or dissimilarity) between gene expression profiles may be Pearson’s correlation (Eisen et al., 1998). Linear regression coefficient and Euclidean distance are two metrics very similar to Pearson’s correlation.

One of the main limitations of these metrics is that their values are constant and stationary. However, for many gene time-series expression profiles, the connectivity between genes is variable and dynamic. Hence, constant and stationary metrics cannot always characterize the variable and dynamic connectivity between genes. So far, there has been no study on this dynamic relationship. We believe that variable parameter regression is an appropriate tool for characterizing this time-dependent correlation relationship. It happened that Buchel and Friston (1998) used variable parameter regression and Kalman filtering to characterize the dynamic relationship between two fMRI signals. We believe that they can also be used to model the dynamic relationship between genes, although our problem is very different from that studied by Buchel and Friston (1998). This idea was tested on some simulated data and real gene expression data. All the results demonstrate that this method can detect successfully the changes of connectivity between genes (or other signals).

METHODS

Materials

In this paper, we apply our algorithm to a simulated dataset and some real data. As shown in Figure 1, we generated two simulated signals $x$ and $y$ with 286 points along the time line. Signal $x$ has six half-sin curves and the content between any two half-sin curves is Gaussian noise. Signal $y$ is similar to signal $x$. The main difference between $x$ and $y$ is that the half-sin curves in signal $y$ are added uniformly distributed noise. We also selected four similar gene expression profiles from the dataset of Cho et al. (1998) and grouped them into two pairs randomly. One pair is YNL309w and YML060w, as shown in Figure 2. Figure 3 shows another pair, YDL164c and YLR338c. Cho et al. collected cells at 17 time points at 10 min intervals, covering nearly two full cell cycles. The time course was divided into five phases: early G1, late G1, S, G2 and M based on the size of the buds. In order to weaken the effect of system error, we first normalized the raw dataset of Cho et al. such that the mean is 0 and the variance is 1.

Variable parameter regression

Variable parameter regression can be described as follows:

$$y_t = x_t \beta_t + u_t, \quad t = 1, \ldots, T, \quad (1)$$

$$u_t \sim N(0, \sigma^2) \quad (2)$$

where $y_t$ is the expression value of gene $y$ at time $t$, $x_t$ is the expression value of gene $x$ at time $t$ and $\beta_t$ is an unknown coefficient that corresponds to estimates of connectivity at time $t$; $u_t$ obeys Gaussian distribution with zero mean and $\sigma$ standard deviation. As described in Buchel and Friston (1998), the dynamic evolution of $\beta$ over time is assumed to follow the following equations:

$$\beta_t = \beta_{t-1} + p_t, \quad t = 2, \ldots, T, \quad (3)$$

$$p_t \sim N(0, \sigma^2 Q) \quad (4)$$

*To whom correspondence should be addressed.
where $\sigma^2 Q$ is the stationary covariance matrix of the innovation $p_t$. From Equation (4), we can see that if $Q = 0$, then parameter $\hat{\beta}_t$ does not change along time and the variable parameter regression reduces to the stationary coefficient linear regression problem. We can see that Equation (3) is in fact a random walk model for $\hat{\beta}_t$. The innovations $n_t$ and $p_t$ are uncorrelated.

**Parameter estimation using Kalman filtering**

Given two gene expression profiles $x_1, \ldots, x_T$ and $y_1, \ldots, y_T$, we are interested in the corresponding regression coefficients $\beta_1, \ldots, \beta_T$. In this paper, we use Kalman filtering to estimate these regression coefficients. Kalman filtering is a recursive solution to the optimal linear filtering problem. We define $\hat{\beta}_t^\text{e}$ to be the prior estimate of regression coefficient at time $t$ given knowledge of the process prior to time $t$, and $\hat{\beta}_t$ to be the posteriori estimate of regression coefficient at time $t$ given the expression value of $y$ at time $t$. Let $P_t^\text{e}$ be the prior estimate error variance and $P_t$ be the posteriori estimate error variance. We define $K_t$ to be the gain that minimizes the posteriori error variance. Then the first step of Kalman filtering is to obtain the prediction that updates $\hat{\beta}_{t-1}$ and its error variance for the passage of time $t - 1$ to $t$:

$$\hat{\beta}_t^\text{e} = \hat{\beta}_{t-1}$$

$$P_t^\text{e} = P_{t-1} + Q. \tag{5}$$

Equations (5) and (6) are also called time update equations. The time update equations are responsible for projecting forward the current state and error variance estimates to obtain a prior estimate for the next time step. The second step of Kalman filtering is the filter step, which revises this estimate of $\hat{\beta}_t$ using the new information contained in the measurement $y_t$:

$$K_t = P_t^\text{e} \frac{x_t}{P_t^\text{e} x_t^2 + 1} \tag{7}$$

$$\hat{\beta}_t = \hat{\beta}_t^\text{e} + K_t (y_t - x_t \hat{\beta}_t^\text{e}) \tag{8}$$

$$P_t = (1 - K_t x_t) P_t^\text{e}. \tag{9}$$

Equations (7)–(9) are also called measurement update equations. The measurement update equations are responsible for incorporating a new measurement into a prior estimate to obtain an improved posteriori estimate. Time update equations and measurement update equations update each other recursively. $x_t^\text{T}$ is the transpose of $x_t$. And because $x_t$ is a scalar, $x_t^\text{T}$ equals $x_t$ here. We know that estimates from previous time are less reliable than those from later ones. We then use the third step, Kalman smoothing, to circumvent this problem. This step can add the information that arrived after time $t$ to the estimate of $\hat{\beta}_t$. Let $\hat{\beta}_{t+}^\text{s}$ be the smoothed estimate of $\hat{\beta}_t$, and then the third step can be depicted as follows:

$$\hat{\beta}_{t+}^\text{s} = \hat{\beta}_t + \frac{P_t}{P_t + Q} (\beta_{t+1}^\text{e} - \hat{\beta}_t). \tag{10}$$

The initial value $\beta_{1+}^\text{s}$ is set to $\hat{\beta}_1$. Then Equation (10) is also a recursive process and $\hat{\beta}_{T+}^\text{s}$ can be solved by this process. We take $\hat{\beta}_{T+}^\text{s}$ as the final estimate of $\beta_T$. From the process of Kalman filtering, we can see that the regression coefficients are determined not only by $y/x$ but also by its historical
The result of the simulated data. The experimental result of the simulated data x and y. The dynamic changes of regression coefficient $\beta$ reflect the dynamic connectivity strength between x and y. High-regression coefficients indicate high connectivity and low regression coefficients indicate low connectivity between two signals.

RESULTS

We first applied our algorithm to the simulated data. From the simulated data shown in Figure 1, we can see that signals x and y are strongly correlated (connectivity or strong connectivity) at corresponding sine regions and weakly correlated (non-connectivity or weak connectivity) at the random noise regions. The result on the simulated data is shown in Figure 4. From Figure 4, we can see that the regression coefficients change dynamically along the time axis. This regression coefficient curve is perfectly consistent with the curves of signal x and signal y. The peaks of this curve correspond to the sine regions and the valleys of this curve correspond to the Gaussian noise regions. This means that the sine regions are more correlated than the noise regions. From this result, we can see that our algorithm depicts dynamic connectivity between genes with time.

Subsequently, we applied our algorithm to two pairs of genes selected from the dataset of Cho et al. Figure 5 shows the result of genes YNL309w and YML060w. From Figure 5, we can see that the regression coefficients (connectivity) between YNL309w and YML060w have two peaks, which means that the two genes profiles are more time-dependent near these peaks and then have strong connectivity near these peaks. According to Cho et al.’s information, we mapped the time points of these two peaks back to the cell cycles and then deduced that these peaks were near G and S phases. This means that YNL309w and YML060w are more correlated near G and S phases and interact with each other during G and S phases with a high probability. YNL309w takes part in the processes of DNA ligation and base-excision repair, which are also strongly related to G1 and S phases. The result of genes YDL164c and YLR383w is shown in Figure 6. Two peaks located near G and S phases means YDL164c and YLR383w have strong connectivity during G and S phases. YDL164c takes part in the processes of DNA ligation, DNA recombination, base-excision repair, lagging strand elongation and nucleotide-excision repair. YLR383w takes part in the processes of DNA repair and cell proliferation. These processes take place mainly during G and S phases. These results mean that our algorithm successfully shows the dynamic connectivity between these pairs of genes.

CONCLUSIONS AND DISCUSSIONS

Main contribution of the current work is that we introduce a novel insight for characterizing the relationship between genes and suggest a proper mathematical tool to model it. The results have demonstrated...
that this technique successfully assesses the dynamic connectivity between two signals on both simulated data and real data. Connectivity can be regarded as the influence that one signal (or gene) exerts over another. Dynamic connectivity depicts the changes of connectivity strength. Moreover, some detailed information about these genes supports our results well.

Apart from these advantages of our method, there are still some limitations. First, an implicit assumption of gene expression data analysis is that genes with similar expression profiles have similar functions in cells. However, this assumption is not always right (Zhou et al., 2002). Almost all gene expression data analysis methods have this limitation. Second, we assumed that the connectivity between two genes is linear, which may not be true. Third, the time points of gene expression profiles are too small; therefore, more time points are needed in order to get better results. Then, if we are interested in some particular genes, we can use real-time quantitative PCR (RTQ–PCR) to analyze these genes for more time points. The result of RT–PCR data for more time points will demonstrate the dynamic connectivity better.

ACKNOWLEDGEMENTS

We thank Meng Liang and Chaozhe Zhu for some valuable discussions. We are grateful to Dr Elizabeth E. Budy for some language corrections. This work was partially supported by the Natural Science Foundation of China, Grant nos 30425004 and 60121302, and the National Key Basic Research and Development Program (973) Grant no. 2004CB318107.

REFERENCES