Measuring Correlation between Microarray Time-series Data Using Dominant Spectral Component

Lap Kun Yeung[1], Hong Yan[1, 2], Alan Wee-Chung Liew[1], Lap Keung Szeto[1], Michael Yang[3] and Richard Kong[3]

[1] Department of Computer Engineering and Information Technology
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[2] School of Electrical and Information Engineering
The University of Sydney, Sydney, NSW 2006, Australia
[3] Department of Biology and Chemistry
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

itwcliew@cityu.edu.hk

Abstract

Microarray time-series data provides us a possible means for identification of transcriptional regulation relationships among genes. Currently, the most widely used method in determining whether or not two genes have a potential regulatory relationship is to measure their expressional similarity using Pearson's correlation coefficient. Although this traditional correlation method has been successfully applied to find functionally correlated genes, it does have many limitations. In this paper, we propose a new metric for more reliable measurement of correlation between gene expression data. In our method, time-series expression profiles are decomposed into spectral components and correlations between them are computed in a component-wise sense. This technique has been applied to known gene regulations of yeast and is able to identify many of those missed by the traditional correlation method.

Keywords: Microarray, gene expression, correlation, gene regulation

1 Introduction

Recent studies (DeRisi *et al* 1997, Spellman *et al* 1998) have demonstrated the use of DNA microarray technology for measuring gene expression comprehensively in a biological system across time. Within the resulting genome-wide expression data produced by this technology is an immense amount of biological information waiting to be discovered and organized. In particular, such time-series data provides a possible means for inference of transcriptional regulations among various genes, and finding of such regulations is one of the major tasks in microarray data analysis.

There are a few approaches for extracting regulatory information from time-series microarray data including simple correlation analysis (Eisen *et al* 1998), edge detection method (Filkov *et al* 2000), event method (Kwon *et al* 2003) and Bayesian networks modeling (Friedman *et al* 2000). Among various approaches, correlation-based clustering is perhaps the most popular one for this purpose. This method determines whether or not two genes have a regulatory relationship by checking the global similarity between their expression profiles. However, it does not take into account the fact that there is often a time delay before the regulator gene product can exert its influence on the target gene. In fact, such time delay can significantly degrade the performance of the method. The edge detection method and the event-based method are specifically designed to overcome various shortcomings of the correlation-based analysis. While these new methods have been shown to be more robust under certain conditions, they, as well as the Bayesian networks approach, have not fully utilized the temporal properties of time-series data.

In this study, we propose an alternative approach based on the autoregressive modeling technique (Marple 1987) to measure correlation between time-series expression data and the results obtained can be used to infer potential regulatory relationships. This technique summarizes the essential features of an expression pattern by means of an estimated frequency spectrum. Specifically, the pattern is decomposed into a set of damped sinusoids of different frequencies so that each sinusoid can be considered separately during the analysis. Hence, this method allows us to have the flexibility of ignoring irrelevant frequency components that may otherwise be too overwhelming

2 Methods

Before discussing our method in details, it is necessary to state the assumptions, which will be made in our analysis, for transcriptional regulation. There are two types of regulation at the level of transcription – activation and inhibition. In the activation process, the product of gene A affects the transcription process of gene B such that the production rate for gene B increases. On the other hand, the inhibition process involves gene A's product decreasing the production of gene B. Following such definition, we would expect to observe in data a rise in gene A followed by a corresponding rise (activation) or fall (inhibition) in gene B with a certain amount of time

delay, and a fall in gene A followed by a delayed fall (activation) or rise (inhibition) in gene B.

If the expression of gene A varying periodically at a constant frequency, the expression of gene B should be varying more or less at the same frequency. This frequency of variation, however, may not be easily seen from the two time-series expression profiles due to noise and other factors. For example, if gene B is under influences of both gene A and gene C simultaneously, and the expression profiles of these influencing genes are varying at different frequencies, then the relationship between gene A and gene B may not be easily seen from their time-series profiles. This is particularly true for correlation-based similarity comparison. Therefore, we are interested in detecting regulatory relationship between two genes through the spectral properties of their expression measurements. In fact, we would like to identify the "influence-by-gene-A" evidences from the spectrum of gene B's expression pattern in the above "two-regulating-one" situation.

The fundamental idea behind our proposed technique is to decompose a time-series expression sequence $x[n]$ into a set of discrete-time damped sinusoids of various frequencies. In other words, we model the sequence as,

$$x[n] = \sum_{i=1}^{M} x_i[n] = \sum_{i=1}^{M} \alpha_i \exp(\sigma_i n)\cos(\omega_i n + \phi_i) \qquad (1)$$

The parameters in this model, α_i, σ_i, ω_i, and φ_i (i = 1, 2, 3 … M), are the amplitude, damping factor, normalized frequency and phase angle respectively of component i. They can be determined based on the autoregressive model commonly used in signal processing (Marple 1987, Yan 2002) and completely define the gene expression spectrum. Now, correlation of $x[n]$ with another sequence $y[n]$ can be reformulated as a sum of scaled sub-correlations,

$$x[n] \circ y[n] = \sum_i \sum_j \sqrt{\frac{E_{x_i} E_{y_j}}{E_x E_y}}\, x_i[n] \circ y_j[n] \qquad (2)$$

where ○ represents correlation operation and each term with letter E represents either total energy of a sequence or energy of a particular component. This equation explains how a correlation of two sequences can be separated into a set of scaled sub-correlations, which may provide more significant insights into the regulatory relationship. For instance, considering the previously mentioned "two-regulating-one" situation, correlation between the expression profiles of gene A and gene B may not be strong enough to suggest their relationship due to the presence of spectral components in gene B inherited from gene C. However, we would expect that spectral components of gene B inherited from gene A will exhibit stronger correlations to gene A's expression profile. Therefore, these scaled sub-correlations can be used instead as a more reliable measurement for the relationship between these genes. Later in the study, we will use the maximum scaled sub-correlation with phase alignment as a new metric for measuring gene-to-gene relationship and its corresponding non-scaled value will be called component-wise correlation coefficient.

3 Results and Discussion

Our analysis is conducted on the publicly available Spellman's alpha-synchronized yeast cell-cycle dataset. The full Spellman's alpha-synchronized dataset consists of 18-point temporal mRNA level measurements sampled at every 7-minute time interval for all 6,178 ORFs in yeast. The test samples in this experiment were synchronized by alpha-factor method such that all the cells would be at the same stage in their cell cycle and the reported data are log ratios of the test sample expression by control sample expression level measurements. In the following discussion, we will focus on a subset of data selected by Filkov. This set of data contains 439 pairs of known transcriptional regulations of which 343 pairs are activations and 96 pairs are inhibitions. Altogether, 288 genes are involved in these regulations.

The traditional correlation method is only capable of identify less than 20% of known regulatory pairs that exhibit strong correlations in the Spellman's dataset. Obviously this is not a satisfactory result and suggests that the simple correlation metric is ineffective. Our main objective here is thus to look for any spectral-domain feature which can help to detect potential regulatory candidates. In particular, we would like to determine whether or not expression patterns of a regulated pair have fully or partially identical frequency contents. To achieve such objective, spectrum signatures obtained by the AR modeling technique for all 439 regulated pairs are compared and we are able to obtain several interesting findings based on our spectral analysis method.

First of all, there are many regulation pairs having strong oscillatory but time-shifted expression that we can easily identified by using only the spectral magnitude information while ignoring the phases. An example is given in figure 1(a) in which the expression profiles and spectrums for an activation pair involving genes *YAL040C* and *YER111C* are shown. These two patterns strongly oscillate at around 0.76 radians per second but still have a relatively low correlation value of -0.3885 because of the time-lag between them as well as other unmatched components. The spectral magnitude plot (the middle graph of figure 1(a)) shows that their dominant components are closely matched with each other. One thing should be mentioned here is that the component for gene *YAL040C* at frequency of 3.1416 radians per second is not considered as the dominant one due to its large decaying factor. Another example, which is an inhibition regulation, is shown in figure 1(b). Careful examination of phases of the dominant components in these examples suggests that the activatee has a phase-lag of 143.6^o relative to the activator's phase angle, whereas the inhibitee has a phase-lead of 95.4^o relative to the inhibitor's phase angle. These two properties are clearly demonstrated in figure 1(c) and figure 1(d) respectively.

Secondly, we are often required to neglect certain irrelevant components that may otherwise corrupt the correlation between two expression patterns. Actually, a large number of known regulations having weak correlations are caused by such noise-like components. Let us once again consider the example shown in figure 1(a). The components at 0.7248 radians per second for gene

YAL040C and 0.8066 radians per second for gene *YER111C*, as shown in table 1, clearly dominate over the others and we may therefore correlate only these two components. As a matter of fact, the component-wise correlation with phase alignment in this case is 0.7665. Comparing to the original one of -0.3885, this new value strongly suggests the similarity between them.

Thirdly, for those regulations involving a single gene being simultaneously regulated by two or more genes with different expressional frequencies, we may be able to identify them by simply checking the existence of regulators' frequencies from the expression profile of the gene being regulated. Figure 2 shows two known activation regulations with a common gene *YPR120C* as an activatee. The figure reveals that the first regulation has its expression profiles correlated at frequency of around 1.48 radians per second, whereas the second regulation has its profiles correlated at around 0.76 radians per second. These examples suggest the "two-regulating-one" situation mentioned earlier.

a	*YAL040C*				*YER111C*			
i	σ_i	ω_i	α_i	φ_i	σ_i	ω_i	α_i	φ_i
1	-1.60	0.00	0.87	3.14	-0.02	0.81	0.52	-2.53
2	-0.00	0.72	0.32	-0.02	-0.26	1.21	0.29	-2.85
3	-0.07	1.59	0.17	-2.39	-0.36	1.58	0.13	-2.69
4	-0.13	3.03	0.21	0.68	-0.09	2.65	0.16	-0.04
5	-3.13	3.14	1.20	0.00	-	-	-	-

Table 1a: Estimated frequency components for gene *YAL040C* and *YER111C*.

b	*YBR049C*				*YGR254W*			
i	σ_i	ω_i	α_i	φ_i	σ_i	ω_i	α_i	φ_i
1	-1.00	0.00	0.45	0.00	-1.50	0.00	0.07	3.14
2	-0.04	0.64	0.24	2.64	-0.10	0.63	0.14	-1.97
3	-0.10	2.02	0.05	1.25	-0.37	1.66	0.08	2.25
4	-0.07	3.14	0.24	3.14	-0.17	2.18	0.18	0.29
5	-0.38	3.14	0.02	-3.14	-0.03	3.14	0.19	0.00

Table 1b: Estimated frequency components for gene *YBR049C* and *YGR254W*.

From all these observations on known regulations, we believe that component-wise correlation analysis should be able to identify those missed by the traditional correlation method. Consequently, a procedure is developed to correlate expression pairs in a component-wise sense and is applied to all 439 known regulations. The results indicate that 223 out of 343 activations and 55 out of 96 inhibitions have their component-wise correlations score greater than 0.5. Table 2 summarizes what we have obtained by the traditional and component-wise correlation methods. The statistics in the table imply that a large number of visually dissimilar expression pairs do have very similar dominant frequency components. For example, among those 307 pairs having traditional correlation coefficients of less than 0.5, 196 of them have greater than 0.5 component-wise correlation coefficients. Furthermore, 60 out of this 196 pairs have their component-wise correlation coefficients even greater than 0.9 and the expression patterns in each of these pairs strongly oscillate at almost identical frequencies.

We have also found eight "two-regulating-one" activation sets having the following properties: i) each set contains a common activate; and ii) the activatee has two different correlation frequencies to its regulators. These activation sets are summarized in table 3. Inhibition sets with similar characteristics have also been found but will not be included here due to the poor quality of the data.

a	*Traditional Correlation < 0.5*	*Traditional Correlation > 0.5*	*Total*
Component-wise Correlation < 0.5	111	9	120
Component-wise Correlation > 0.5	196	27	223
Total	307	36	343

Table 2a: Results for different correlation methods applied to all 343 activation pairs.

b	*Traditional Correlation < -0.5*	*Traditional Correlation > -0.5*	*Total*
Component-wise Correlation < 0.5	1	40	41
Component-wise Correlation > 0.5	4	51	55
Total	5	91	96

Table 2b: Results for different correlation methods applied to all 96 inhibition pairs.

The results we obtained suggest that relationship between two genes may be more clearly revealed by considering the spectral components of their expression profiles than simply looking at their time-domain format.

We should point out here that the usefulness of the spectral decomposition method proposed in this paper is not limited to expression data from cyclic genes. Autoregressive modeling is widely used for processing of signals from electric, acoustic, mechanical and nuclear spin systems (Marple 1987, Yan 2002), and most of these signals are non-periodic in nature. It is a general method for data processing and analysis, similar to the Fourier and wavelet transforms. Each spectral component corresponds to a mode of system response in electric, acoustic or mechanical systems, or a chemical component in nuclear magnetic resonance. In fact, it will be an interesting problem to investigate the biological interpretation of these spectral components, and each component may come from a specific gene or regulatory pathway.

Activator	*Activatee*	*Traditional Correlation*	*Component-wise Correlation*	*Activator Frequency*	*Activatee Frequency*
YKL 109W	*YGL 167C*	0.2877	0.9237	0.6505	0.6842
YLR 433C	*YGL 167C*	0.1980	0.5287	1.3502	1.6359
YHR 079C	*YJL 034W*	-0.6594	0.9894	0.7063	0.7205
YPL 085W	*YJL 034W*	0.2717	0.6120	1.7581	1.9513
YKL 109W	*YLL 041C*	0.3792	0.9917	0.5339	0.5230
YBL 021C	*YLL 041C*	0.2586	0.9564	1.3748	1.3725
YGL 237C	*YLL 041C*	-0.4687	0.8484	0.6456	0.5230
YOR 358W	*YLL 041C*	0.3800	0.8008	1.2639	1.3725
YLR 182W	*YLR 286C*	-0.1208	0.8984	1.1082	1.0378
YLR 071C	*YLR 286C*	0.0349	0.6662	0.3324	0.4353
YLR 131C	*YLR 286C*	-0.2762	0.6535	0.5338	0.4353
YEL 009C	*YOR 202W*	0.0554	0.9276	1.2653	1.1670
YRL 082C	*YOR 202W*	0.6075	0.8912	0.3199	0.3517
YEL 009C	*YPR 035W*	-0.3737	0.9541	1.2653	1.2241
YFL 021W	*YPR 035W*	-0.2153	0.9002	0.4095	0.3662
YGR 274C	*YPR 120C*	0.4075	0.8541	1.5266	1.4566
YAL 040C	*YPR 120C*	-0.4331	0.7288	0.7248	0.8120
YLR 256W	*YPR 191W*	-0.1491	0.9173	0.7762	0.7295
YGL 237C	*YPR 191W*	-0.7333	0.8821	0.6456	0.7295
YBL 021C	*YPR 191W*	-0.2231	0.7569	1.3748	1.4294
YOR 358W	*YPR 191W*	0.0937	0.7209	0.6227	0.7295
YKL 109W	*YPR 191W*	0.2663	0.6254	0.5339	0.7295

Table 3: Selected activation regulation sets. Each set contains a common activatee which has two different correlated frequency components.

4 Conclusion

In summary, we have demonstrated the use of autoregressive modeling technique to extract the spectral characteristics of time-series expression data and use them to infer potential regulation relationships among genes. This method allows us to analyze the temporal aspect of time-series microarray data. In particular, it can reveal the hidden spectral component-wise relationship between two expression profiles. Furthermore, the phase information obtained from our technique provides a possible way of separating between potential activation and inhibition regulations.

5 Acknowledgment

This work is supported by a CityU interdisciplinary grant

6 References

DeRisi, J., Iyer, R. & Brown, P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* **278**, pp. 680-686.

Spellman, P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell* **9**, pp. 3273-3297.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U.S.A.* **95**, pp. 14863-14868.

Filkov, V., Skiena, S. & Zhi, J. (2002) Analysis techniques for microarray time-series data, *J. Comput. Biol.* **9**, pp. 317-330.

Kwon, A., Hoos, H. & Ng, R. (2003) Inference of transcriptional regulation relationships from gene expression data, *Bioinformatics* **19**, pp. 905-912.

Friedman, N., Linial, M., Nachman, I. & Péer, D. (2000) Using Bayesian networks to analyze expression data, *J. Comput. Biol.* **7**, pp. 601-620.

Marple, S. (1987) *Digital Spectral Analysis with Applications*, (Prentice Hall Inc., Englewood Cliffs, New Jersey), pp. 206-260.

Yan, H. (2002) *Signal Processing for Magnetic Resonance Imaging and Spectroscopy*, (Marcel Dekker, New York), pp. 475-507.

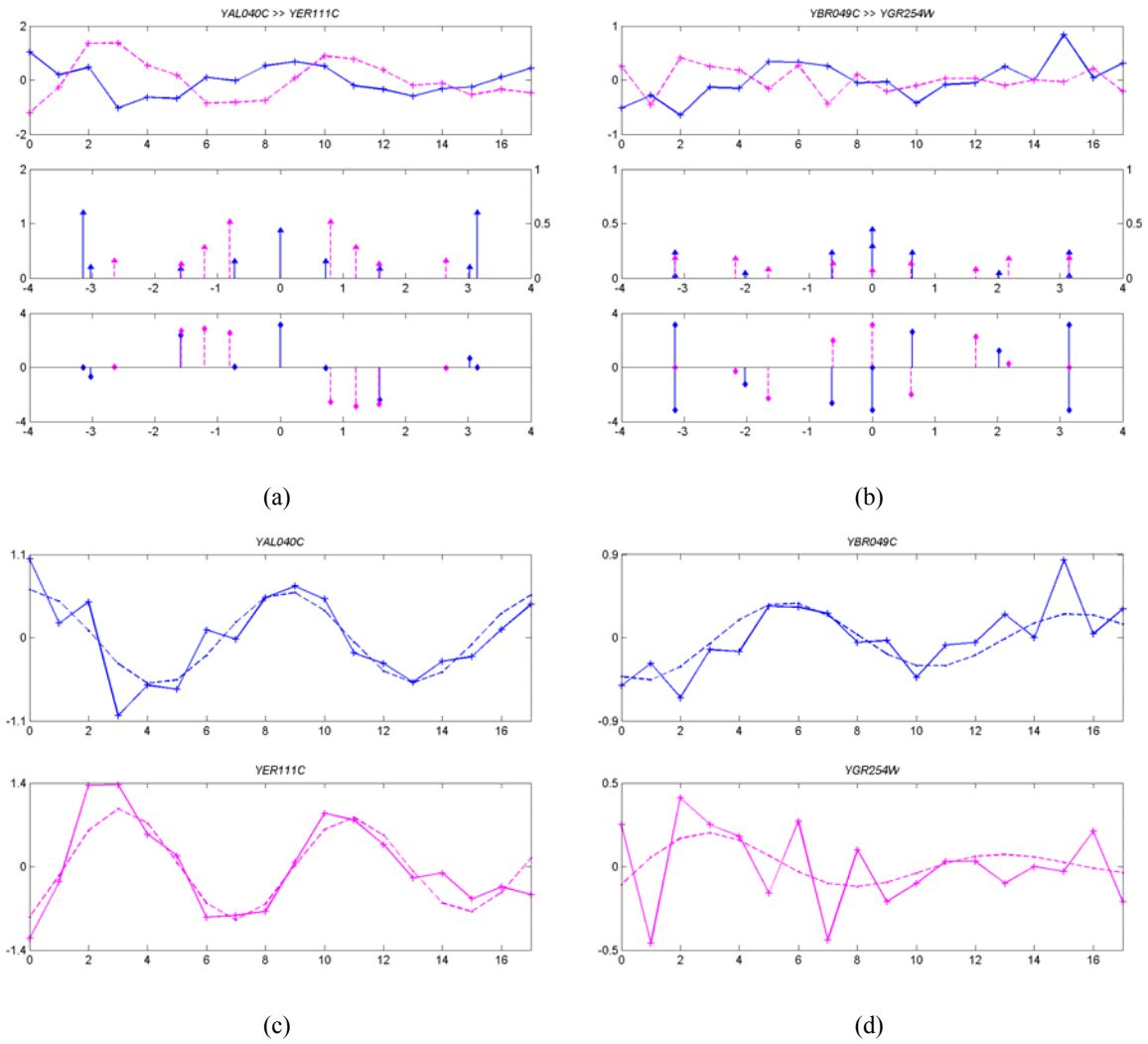

Figure 1: Selected gene regulations of yeast. (a) Known activation regulation involving genes *YAL040C* and *YER111C*. The dominant frequencies for these two profiles are 0.7248 radians per second and 0.8066 radians per second respectively. (b) Known inhibition regulation involving genes *YBR049C* and *YGR254W*. The dominant frequencies are 0.6395 radians per second for the first gene and 0.6271 radians per second for the second gene. (c) Dominant sinusoids for the activation pair. (d) Dominant sinusoids for the inhibition pair.

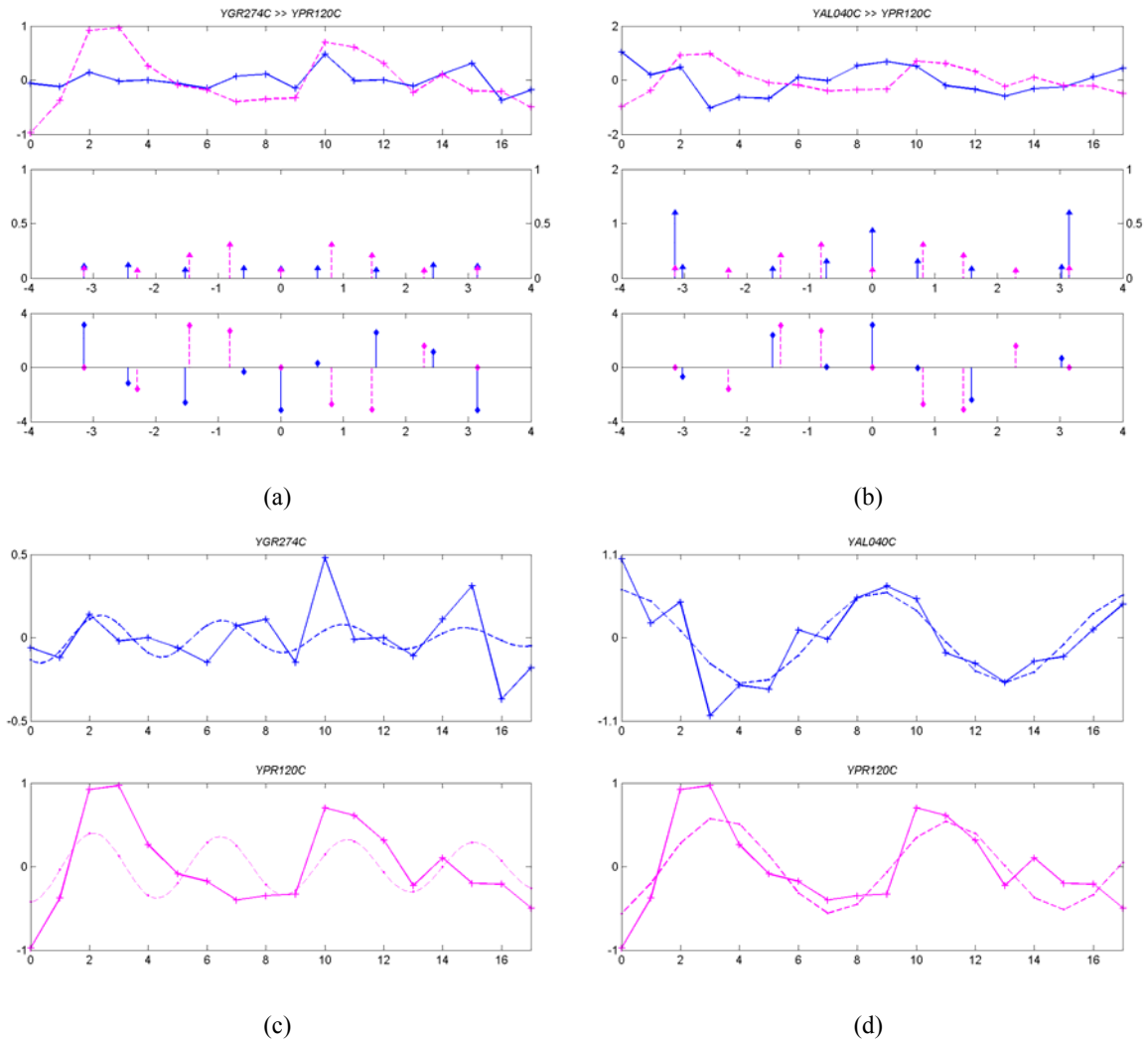

Figure 2: Two activation regulations with gene *YPR120C* as an activatee. (a) Activation regulation with gene *YGR274C* as an activator. (b) Activation regulation with gene *YAL040C* as an activator. (c) Correlated frequency components for the first pair. (d) Correlated frequency components for the second pair.