2 Related Models and Analyses

Only a few studies are available for the phase shift of gene expression profiles. There is a study that phase shifts of circadian gene expression were modeled as a mixture of two von Mises distributions corresponding to two gene clusters, tissue-dependent phase cluster and tissue-independent synchronized phase cluster [10].

In biology, phase advance and phase delay phenomena were thought to be related with different stress [11]. In this study, three kinds of phase shift experiments were examined induced by the timing of phase-reset stimuli by drug. First experiment is called CT6, and the master clock gene *per1* is considered to keep its phase unchanged as in the control condition. Second experiment is called CT14, and the clock gene *per1* shows phase delay against the control condition. The last experiment is called CT22, and the *per1* shows phase advance in comparison with the control condition. Not all circadian-related genes are synchronized like *per1* in phase, and phase-difference distributions seem to have unidentified complex structure. A part of the mechanism will be elucidated in Data section.

About the modeling and approximation for circadian data analyses, many conventional studies use cosine fitting with minimum square method. For example, Fast Fourier Transform (FFT) was used for Arabidopsis circadian rhythm [12]. Another trigonometrical function analysis was used for sleep analysis [13]. Cluster analysis based on the cosine correlation was applied for mouse circadian clock [14]. Lomb-Scargle periodograms were also applied [15][16][17].

In this study, we focused on the phase-shift phenomena caused by drug stimuli for SCN cultured cells. A few similar study exists [10][18], in which only known clock genes were observed in mice. In contrast, we observed all oscillating genes including known clock genes to explore the characteristics of phase-difference distribution among three phase-shift experiments.

3 Materials and Methods

3.1 Phase detection

Each time series of control and three conditions was first normalized assuming a normal distribution whose mean is zero and the variance is 1 for each experiment. Fast Fourier transform was formulated for each normalized time series. Because the variance of power spectra of each gene determines whether typical oscillations exist or not, we identified only those genes as oscillating whose spectral variance are significantly large. In our case, around 300 oscillating genes, about 1 percent of the whole genes, were extracted, considering the previous reports that the number of oscillatory genes is from several percents to ten percents of expressed genes in each organ [14][19][20][21].

Random period fitting was formulated based on the following formula.

$$y = a\sin(\frac{2\pi x}{p} - \theta) + b$$

y is the expression time series of oscillating genes. x is time. a and b are constant parameters. θ is phase parameter. p is period variable sampled from a normal distribution whose mean is 27 and variance is 1.

The phase-difference distributions were generated by calculating the phase difference between control phase and experimental phase for each gene assuming that the control oscillations keep their periodicity until the same time with experimental conditions. Also the relationship between the random periods and phase difference was explored.

3.2 Data

We used rat cultured cells sampled from SCN, and measured gene expression profiles with Affimetrix microarray (Genechip Rat Genome 230 2.0). The oscillation period was set to about 27 h because our previous report explored and found the circadian period around 27h.

4 Results and Discussion

4.1 Random Period Model

We follow the random period model for the approximation of each gene time-series [18], because genes apparently fluctuate in various scale for each cycle, to adjust to light conditions or other environmental factors. However, our model is much simpler than the model by Liu *et al.* [18] in order to check the phase-difference distribution under drug stimuli. The phase difference was calculated between the control data and three phase-shifted experimental data.

Figure 1 shows the distributions of phase difference, consisting of 300 oscillatory genes between the phase under control and three different conditions: a) In the experimental condition called CT6 phase-reset stimulus was supplied at the time of 18 hours from the time of exchanging culture medium for the cultured cells; b) In the experimental condition called CT14, the stimulus was supplied at the time of 27 hours; and c) In the experimental condition called CT22, the stimulus was added at 36 hours.

4.2 Existence of Two Major Periodic Groups

The experimental condition CT6 has been considered that the phase-reset stimuli does not cause phase shift. However, Figure 1a shows that there are many phase-shifted oscillatory genes. CT14 has been considered as the phase delay condition. Figure 1b shows, however, the result including dual phase differences, around 27 degree and 207 degree. CT22 has been considered as the phase advanced condition. Figure 1c shows the result including dual phase differences, around 99 degree and 279 degree.

These results indicate that the phase shift characteristics depend on each gene and phase-reset timing, even though the conditions are roughly called "phase stable" or "phase-advanced/delay" [3]. The results of CT14 and CT22 imply the existence of dual phase-fluctuation structure which may be achieved the role difference in circadian clock system [1]. Note that the dual deviations are consistent and shifted about 60 degree between CT14 and CT22.

The phase-difference distributions were different from the report of the past study [10], which extrapolated the mixture of only two von Mises distributions for known circadian genes in mouse. Unlike the past study [10], "unchanged phase cluster" was not identified in our study. Moreover, CT22 data exhibited multiple phase differences, although the other two (CT6 and CT14) showed only two major phase differences. There

are two reasons why the distributions were so different from the past study. First is the difference of biological target. The past study examined the phase difference among tissues, but our study examined different timings of phase-reset stimuli. The other reason is the coverage of genes. Since we examined all genes that seem to be oscillating, the number was well over 300.

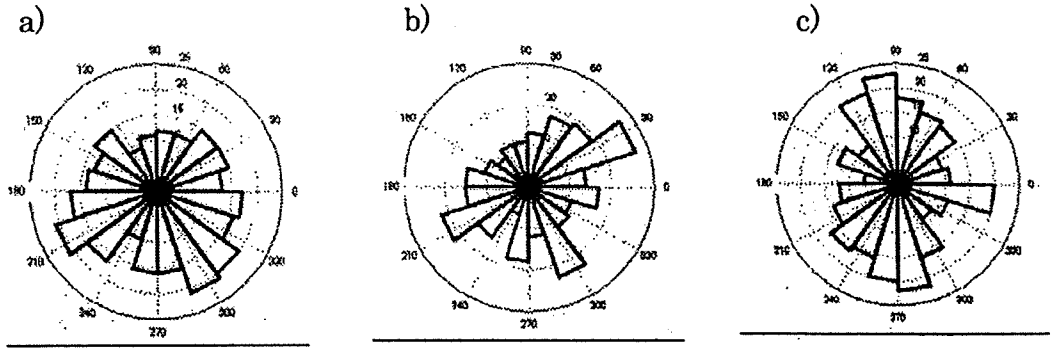


Figure 1: The phase-difference distribution of a) CT6 b) CT14 and c) CT22 among oscillating 300 genes

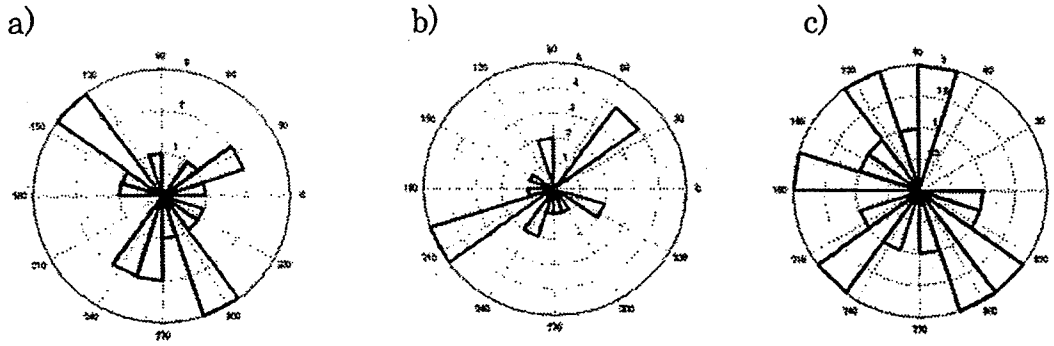


Figure 2: Phase-difference distribution of known clock genes under the condition of a) CT6 b) CT14 and c) CT22

Figure 2 shows the phase-difference distribution among known clock genes [1]. As in Figure 1, the results in Figure 2 show dual phase differences. The clock related genes, *Per1, Per3* and *NPAS2*, were included in the 135 degree, and *CK1ε/δ* are in the 297 degree in Figure 2a. *Clock, Per1* and *Rorβ* and *CK1σ* are included in the 207 degree, and *Cry2, Nr1d1* and *CK1ε/δ* are included in the 45 degree in Figure 2b. Both Figure 2a and Figure 2b show dual shifts, and the phase angles of clock related genes are different by 180 degree. A negative-feedback regulation [1] may exist in its background to adjust the cycle at transcription level and enzyme level (*CK1* in this case). Between Figure 2a and 2b, the dual structure is shifted by 90 degree. Figure 3c is quite different from the other two: clock related genes were fluctuated strongly, and various phase shifts were observed . These results indicate that turbulence of phase syntony depends on the timing of stimuli.

In summary, genetic interactions among oscillating genes were kept under the control, CT6 and CT14 conditions. The only change was the phase differences. On the other hand, the condition CT22 affected the synchronization mechanism and generated strong fluctu-

ation of the oscillatory system. We can hypothesize the existence of unknown regulations that cause the difference between CT22 and the other conditions.

4.3 Dispersion of Clock Genes

Two representative clock genes, *per1* and *CK1*, were synchronized in all phase shifts regardless of different phase-reset stimuli (Figure 2a and 2b). We consider that observations only on such genes have led to the false assumption of CT6, CT14, and CT22 as phase-stable, advance, and delay, respectively. There were also genes scattering in phase-difference distribution (Figure 2c). Biological reason for this large variance of CT22 is unknown. One of our assumption for this dispersion phenomenon is the timing of folskolin stimuli CT22 is close to the border between the phase advance and phase delay, and the timing closed to the border might cause the fluctuation of phase shift mechanism [3].

5 Conclusion

We extracted over 300 oscillatory genes, including known clock-related genes, from the expression data of over 30,000 genes in mouse. By fitting their oscillation to sine curve, their distribution of phase differences was obtained. The distribution had a novel complex structure. Two large gene clusters showed a phase difference of 180 degree in all three experiments with stimuli, indicating the hierarchical role in circadian system. Two experiments showed a clear 90 degree shift, which was almost consistent with the time of stimuli (the time difference of stimuli is 8 hours in 27 hour-cycle and the phase shift is 90 degree.) The last experiment (CT22), however, showed scattered phase differences. It suggested the possibility that there is a particular timing of stimuli which causes large fluctuations of phase synchronization in circadian system.

Acknowledgments

This work was also supported, in part, by Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] CH. Ko, JS. Takahashi, Molecular components of the mammalian circadian clock, *Human Molecular Genetics*, 2006, 15, 271-277.

[2] JC. Dunlap, JJ. Loros, PJ. Decoursey, Chronobiology: Biological Timekeeping, Sinauer Associates, 2003.

[3] S. Kawaguchi, A. Shinozaki , M. Obinata , K. Saigo , Y. Sakaki , H. Tei, Establishment of cell lines derived from the rat suprachiasmatic nucleus, *Biochemical and Biophysical Research Communications*, 355, 2007, 555–561.

[4] JS. Takahashi, FW. Turek, RY. Moore, Handbook of Behavioral Neurobiology: Circadian Clocks, Springer, 2001.

[5] S. Honma, W. Nakamura, T. Shirakawa, K. Honma, Diversity in the circadian periods of single neurons of the rat suprachiasmatic nucleus depends on nuclear structure and intrinsic period, *Neuroscience Letters*, 358, 2004, 173–176.

[6] ED. Herzog, JS. Takahashi, GD. Block, *Clock* controls circadian period in isolated suprachiasmatic nucleus neurons, *Nature Neuroscience*, 1, 1998, 708-713.

[7] S. Honma, T. Shirakawa, Y. Katsuno, M. Namihira, K. Honma, Circadian periods of single suprachiasmatic neurons in rats, *Neuroscience Letters*, 250, 1998, 157–160.

[8] C. Liu, DR. Weaver, SH. Strogatz, SM. Reppert, Cellular Construction of a Circadian Clock: Period Determination in the Suprachiasmatic Nuclei, *Cell*, 91, 1997, 855–860.

[9] K. Yagita, H. Okamura, Forskolin induces circadian gene expression of *rPer1*, *rPer2* and *dbp* in mammalian rat-1 fibroblasts, *Journal of Federation of Ecuropean Biochemical Societies Letters*, 465 2000, 79-82.

[10] D. Liu, SD. Peddada, L. Li, CR. Weinberg, Phase analysis of circadian-related genes in two tissues, *BMC Bioinformatics*, 2006, 7:87.

[11] A.J. Davidson, M. T. Sellix, J. Daniel, S. Yamazaki, M. Menaker and G. D. Block, Chronic jet-lag increases mortality in aged mice, *Current Biology*, 16, 2006, 914-916.

[12] D. Alabadi, M. Yanovsky, P. Mas, S. Harmer, S. Kay, Critical role for CCA1 and LHY in maintaining circadian rhythmicity in Arabidopsis. *Current Biology*, 12, 2002, 757-761.

[13] EJV. Someren, E. Nagtegaal, Improving melatonin circadian phase estimates, Sleeping Medicine, 8, 2007, 590-601.

[14] S. Panda, MP. Antoch, BH. Miller, AI. Su, AB. Schook, M. Straume, PG. Schultz, SA. Kay, JS. Takahashi, JB. Hogenesch, Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock, *Cell*, 109, 2002 307-320.

[15] EJV. Someren, DF. Swaab, CC. Colenda, W. Cohen, WV. McCall, PB. Rosenquist, Bright light therapy: improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods. *Chronobiol Int.*,16, 1999, 505-518.

[16] M. Schimmel, Emphasizing difficulties in the detection of rhythms with Lomb-Scargle periodograms, *Biol Rhythm Res.*, 32, 2001, 341-345.

[17] EF. Glynn, J. Chen, AR. Mushegian, Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms, *Bioinformatics*, 22, 2006, 310-316.

[18] D Liu, DM. Umbach, SD. Peddada, L Li, PW. Crockett, CR. Weinberg, A random-periods model for expression of cell-cycle genes, *Proc Natl. Acad. Sci.* 2004, 101 7240-7245

[19] RA. Akhtar, AB. Reddy, ES. Maywood, JD. Clayton, VM. King, AG. Smith, TW. Gant, MH. Hastings, CP. Kyriacou, Circadian Cycling of the Mouse Liver Transcriptome, as Revealedby cDNA Microarray, Is Driven by the Sprachiasmatic Nucleus, *Current Bioloby,* 12, 2002, 540-550.

[20] KF. Storch, O Lipan, I Leykin, N. Viswanathan, FC. Davis, WH. Wong, HJ. Weitz, Extensive and divergent circadian gene expression in liver and heart, *Nature*, 417 2002, 78-83

[21] HR. Ueda, W Chen, A Adachi, H Wakamatsu, S Hayashi, T Takasugi, M Nagano, K Nakahama, Y Suzuki, S Sugano, M Lino, Y Shigeyoshi, S. Hashimoto, A transcription factor response element for gene expression during circadian night, *Nature*, 408 2002 534-539

The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

An Algebraic-Numeric Algorithm for the Model Selection in Network Motifs in *Escherichia coli*

Masahiko Nakatsui[1] Hiroshi Yoshida[2] Katsuhisa Horimoto[1]

[1]Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial Science and Technology (AIST),
Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan
[2]Faculty of Mathematics, Organization for the Promotion of Advanced Research,
Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581 Japan

Abstract Recently, we have proposed a novel algorithm to select a model that is the most consistent with the time series of observed data. In the algorithm, first, a system of differential equations that express the kinetics for a biological phenomenon and a sum of exponentials that are fitted to the observed data are transformed into the corresponding system of algebraic equations, by the Laplace transformation. Then, the two systems of algebraic equations are compared by an algebraic-numeric approach. One of the merits of our algorithm estimates the model's consistency with the observed data and the determined kinetic constants. Furthermore, our algorithm allows a kinetic model with cyclic relationships between variables that cannot be handled by the usual approaches. In this paper, we examined the performance of our proposed algorithm by using three kinds of highly significant network motifs in *Escherichia coli*; feed-forward loop, single input module, dense overlapping regulons, which are found by Shen-Orr, *et al*[14].

1 Introduction

One of the most remarkable features in biological network analysis is that the network structure itself is unknown, in contrast that the network model is almost always given in the engineering field. This situation indicates that the construction of network model is the first step to clarify the molecular mechanism underlying the biological phenomena. Indeed, the aim of the experimental studies is frequently the discovery of new molecules related with the biological phenomena, and the following aim is to reveal the relationships (interactions) between the newly found molecules. The knowledge about the molecules and their relationships by experimental studies have been reported in many literatures, and they have been compiled at the web (for example, [19]).

The approach for constructing a biological network model by systematic extraction of enormous knowledge from the literatures and the following superimposition of them is recognized as one of the most promising approaches [5]. Since each relation identified by experimental studies is regarded as strong evidence for the existence of edges in the network model, biological network models have been constructed for various biological phenomena. On the other hand, it is well-known that the relationships between the molecules in a living cell change dynamically, depending on the cellular environment.

Thus, the molecular relationships in the literature represent the responses to the different conditions in the experimental studies, and in the network model generated from the biological knowledge, the consistency of the model with the data observed by experimental studies must be considered carefully. Actually, several distinctive models of the relationship between molecules for a biological phenomenon can be obtained from the large amount of information in the literature [3, 6]. In these cases, a model that is consistent with the data observed under particular conditions should be selected from the candidate models.

The consistency of a model with the observed data is investigated intensively by statistical and algebraic approaches. In statistics, the issue of the consistency of a model with the observed data is also well known , as the test for causal hypotheses by using the observed data. The origin of the test for causal hypotheses is attributed to path analysis [17]. Unfortunately, the importance of this cornerstone research has been ignored for a long time, but the natural extension of the path analysis has been established as the well-known structural equation model (SEM) [9]. Indeed, the SEM has been utilized recently in various fields, in accordance with increased computer performance. However, the SEM without any latent variables, which is the natural form for applying the SEM to the biological networks, frequently faces difficulty in the numerical calculation of the maximum likelihood for the observed data. To overcome the difficulty of this calculation, the d-sep test [15] has been developed, based on the concept of d-separation in a directed acyclic graph [12]. Notice that the graph consistency with the data in the d-sep test can consider only the directed acyclic graph (DAG), without any cyclic relationships. In algebraic approach, there exists the identifiability problem in the compartmental models for tracer kinetics [1, 7, 6]. In the compartmental models, the unknown constants are estimated from tracer data in the accessible pools. The identifiability problem addresses the issue of whether the unknown constants can be determined uniquely or non-uniquely from the tracer data. This issue has usually been solved through the transformation of differential equations into algebraic equations, by the Laplace transformation. Although a systematic algorithm for the identifiability problem was proposed [4], its application is limited to the unrealistic context of an error–free model structure and noise–free tracer data. Thus, it still seems to be difficult to solve the identifiality problem for actually observed data, in spite of the mathematical studies.

Recently, we have proposed a new method for selecting models, by estimating the consistency of a kinetic model with the time series of observed data [18]. First, the kinetics for describing a biological phenomenon is expressed by a system of differential equations, assumed that the relationships between the variables are linear. Simultaneously, the time series of the data are numerically fitted as a sum of exponentials. Next, the differential equations with the kinetic constants and the sum of exponentials fitted to the observed data are both transformed into the corresponding system of algebraic equations, by the Laplace transformation. Finally, the two systems of algebraic equations are compared by an algebraic approach. Thus, our method estimates the model's consistency with the observed data and the determined kinetic constants. Indeed, we have successfully illustrated that our method can select the actual botanic models [10], in which a kinetic model with cyclic relationships between variables that cannot be handled by the usual approaches is included, with the corresponding data generated by the differential equations for the relationships. Although we have examined the performance of our method for selecting the

model with a cyclic loop in the previous paper, it remains to be investigated in terms of the model variation, especially the typical forms in the relationships between biological molecules.

Fortunately, the gene regulatory network identified by experimental studies is composed of the limited number of network motifs[14]; each motif has simple forms of 2-layer relationship between the transcription factor and its regulating genes. Even in a complex regulatory network, therefore, the entire network can be factorized into small subnetworks by combination of network motifs. In this paper, we address the issue on the selection of the network motifs in *Escherichia coli* which are proposed by Shen-Orr, *et al.* As the same way as in the previous paper, the data are generated by the differential equations for the relationships, and the consistency of the models with the generated data is calculated by our algebraic-numeric method.

2 Methods

2.1 Overview of Model Selection Algorithm

The procedure for model selection can be summarized as follows:

(i) We fit the observed data as a sum of exponentials in 2.2.
(ii) We perform the Laplace-transformation of both the system of differential equations for the models and the sum of exponentials for the observed data in 2.3.
(iii) By using the least squares method (abbreviated as *LSM*), we calculate the consistency of the model with the observed data.

In what follows, the details of our method will be shown.

2.2 Observed Data Fitting by Genetic Algorithm (GA)

In this paper, we need Laplace-transformed observed data, because we perform the model selection over the Laplace domain. Let $Mo_i(t)$ denote the observed data corresponding to $M_i(t)$ derived theoretically. By genetic-algorithm based numerical fitting, $Mo_i(t)$ is expressed in terms of a sum of exponentials as follows:

$$\beta_b + \sum_{j=1}^{n} \beta_j \exp(-\alpha_j t), \tag{2.1}$$

where n is the number of distinct exponentials determined by $M_i(t)$, and β_b is zero in the case of the non-existence of a constant term within $M_i(t)$. $Mo_i(t)$ thus fitted is changed into the Laplace-transformed data as follows:

$$\frac{\beta_b}{s} + \sum_{j=1}^{n} \frac{\beta_j}{s+\alpha_j}, \tag{2.2}$$

where L denotes the Laplace transformation. In this problem, each set of parameter values α_i, β_i and β_b to be estimated is evaluated using the following procedure: Suppose that $Mo_i(t)$ is the calculated time-course at time t of i and that $Ms_i(t)$ represents sampling data at time t of i. The sum of the square values of the relative error between $Mo_i(t)$ and $Ms_i(t)$ gives the total relative error E_i;

$$E_i = \sum_{t=1}^{T} \left(\frac{Ms_i(t) - Mo_i(t)}{Ms_i(t)} \right)^2, \tag{2.3}$$

where T is the total number of sampling points.

The computational task is to determine a set of parameter values α_i, β_i and β_b that minimizes the objective function E_i. Instead of the use of `NMinimize` command of `Mathematica 5.2` in the previous study [18], here, we use the well-known genetic algorithm (GA). We applied RCGAs with a combination of *unimodal normal distribution crossover* (UNDX)[11] and *minimal generation gap* (MGG)[13] as a nonlinear numerical optimization method for estimating constants.

2.3 Laplace-transformation of Model Formula

Suppose that the model formulae are described over the time domain as follows:

$$\frac{dM_i(t)}{dt} = F_i(\vec{M}, \vec{k}), \tag{2.4}$$

where $\vec{M} = \{M_1, M_2, \ldots, M_n\}$ and $\vec{k} = \{k_1, k_2, \ldots, k_m\}$. Function $F_i(\vec{M}, \vec{k})$ can be determined according to the graph describing the model, and $\vec{k}$ denotes the kinetic constants between the chemicals. We transform this system of differential equations into a system of algebraic equations over the Laplace domain, and solve the equations in $L[M_i(t)](s)$ $(i = 1, 2, \ldots, n)$.

2.4 Calculation of Consistency Measure and Model Selection

To evaluate the consistency of the model with the observed data, we define *consistency measure*. If the model is completely consistent with the observed data and the data lack noise and inaccuracies, then $L[M_i(t)](s) = L[Mo_i(t)](s)$ $(i = 1, 2, \ldots, n)$ holds. This fact has led us to the following definitions of consistency measure:

Let *comp* denote the set of polynomials obtained by matching the coefficients of $L[M(t)](s)$ and $L[Mo(t)](s)$ over the Laplace domain, in which every element is zero in the case of $L[M_i(t)](s) = L[Mo_i(t)](s)$ $(i = 1, 2, \ldots, n)$; that is, when Formula $L[M_i(t)](s) = L[Mo_i(t)](s)$ is an identity in s.

The consistency measure (in short, *CM*) of the model is defined as the smallest sum-square value of the elements in *comp* under the following constraint:

$$k_1 \geq 0, k_2 \geq 0, \ldots, k_m \geq 0. \tag{2.5}$$

In order to obtain the smallest value, we have utilized the least squares method using the following equations:

$$\frac{\partial}{\partial k_1} g(\vec{k}) = \frac{\partial}{\partial k_2} g(\vec{k}) = \cdots = \frac{\partial}{\partial k_m} g(\vec{k}) = 0, \tag{2.6}$$

where $g(\vec{k})$ is the sum-square value of the elements in *comp*.

Then, we survey all of the possible candidates of the minimum by calculating *all* of the real positive roots of the system of algebraic equation (2.6). Several method and

tools exist to calculate all real roots of algebraic equations adjoined by a zero-dimensional ideal.

The consistency measure can be calculated by the following recursive procedure [18]:

Let $MinimumValue(q(\vec{l}))$ denote the *minimum value* of function q with variables: $\vec{l} = \{l_1, l_2, \ldots, l_m\}$ by the following procedure:

1. If the cardinality of $\vec{l}$, namely m, is zero, then the *minimum value* is infinity.
2. Otherwise, let v_0 denote the minimum value of q under Constraint (2.5) via homotopy method. Furthermore, let v_i $(i = 1, 2, \ldots, m)$ denote the value calculated by $MinimumValue(q(\vec{l_i}))$, where $\vec{l_i}$ is the vector: $\{l_1, l_2, \ldots, l_{i-1}, 0, l_{i+1}, \ldots, l_m\}$.
3. The *minimum value* is the smallest value among $v_0, v_1, \ldots, v_m$.

Using the consistency measures, *CM*, we performed model selection. We, first, calculated the consistency measures of the candidate models with the observed data. Then, we listed the smallest consistency measures and the corresponding values of kinetic constants of each candidate model for the two consistent measures. Last, we select simply one candidate model showing the smallest values by the consistent measures.

2.5 Case Study

Shen-Orr *et al.* found three highly significant motifs in the transcriptional regulation network of *Escherichia coli*.[14] We modified these three kinds of network motif to four nodes (M_1, M_2, M_3, and M_4). Fig. 1 shows the three network motif analyzed in this paper. One is a motif of a chain graph with feed-forward loop, the other one is a motif of single input module, and last one is a motif of dense overlapping regulons.

3 Results

3.1 Formulation

According to the models in Fig. 1, the kinetics can be expressed by two systems of differential equations as follows:

Model (a)

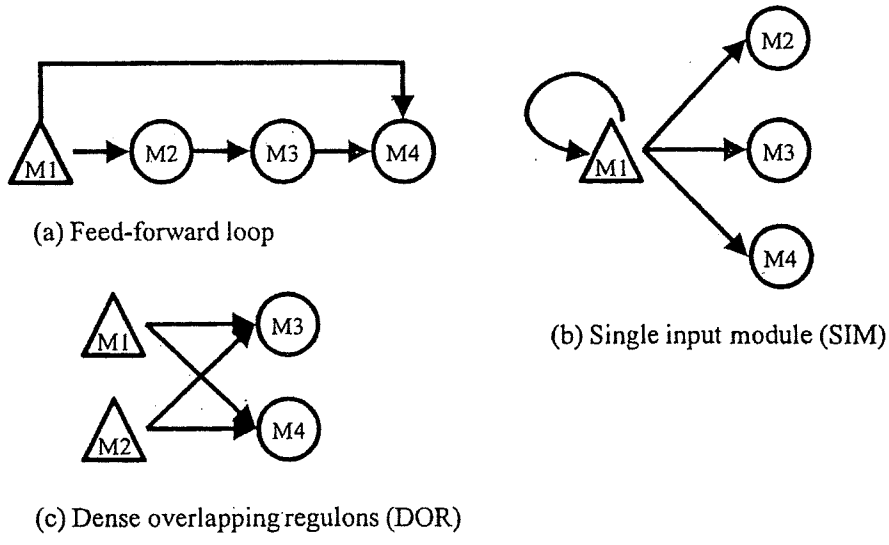
$$\begin{cases} \mathrm{d}/\mathrm{d}t\, M_1(t) &= -k_{12}\, M_1(t) - k_{14}\, M_1(t), \\ \mathrm{d}/\mathrm{d}t\, M_2(t) &= k_{12}\, M_1(t) - k_{23}\, M_2(t), \\ \mathrm{d}/\mathrm{d}t\, M_3(t) &= k_{23}\, M_2(t) - k_{34}\, M_3(t), \\ \mathrm{d}/\mathrm{d}t\, M_4(t) &= k_{14}\, M_1(t) + k_{34}\, M_3(t). \end{cases} \tag{3.1}$$

Model (b)

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, M_1(t) &= k_{11}\, M_1(t) - k_{12}\, M_1(t) - k_{13}\, M_1(t) - k_{14}\, M_2(t), \\ \mathrm{d}/\mathrm{d}t\, M_2(t) &= k_{12}\, M_1(t), \\ \mathrm{d}/\mathrm{d}t\, M_3(t) &= k_{13}\, M_1(t), \\ \mathrm{d}/\mathrm{d}t\, M_4(t) &= k_{14}\, M_3(t). \end{cases} \tag{3.2}$$

Model (c)

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, M_1(t) &= -k_{13}\, M_1(t) - k_{14}\, M_1(t), \\ \mathrm{d}/\mathrm{d}t\, M_2(t) &= -k_{23}\, M_2(t) - k_{24}\, M_2(t), \\ \mathrm{d}/\mathrm{d}t\, M_3(t) &= k_{13}\, M_1(t) + k_{23}\, M_2(t), \\ \mathrm{d}/\mathrm{d}t\, M_4(t) &= k_{14}\, M_1(t) + k_{24}\, M_2(t). \end{cases} \tag{3.3}$$



(a) Feed-forward loop

(b) Single input module (SIM)

(c) Dense overlapping regulons (DOR)

Figure 1: Three kinds of network motif which were proposed by Shen-Orr *et al*[14]. The nodes shown as M1 to M4 are transcription factors. (a) Feed-forward loop: a transcription factor M1 regulates M2, and both jointly regulate M4. (b) Single input module: A single transcription factor M1 regulates a set of regulons shown as M2 to M4. (c) Dense overlapping regulons: a set of regulons M3 and M4 were each regulated by combination of a set of regulator M1 and M2.

where $M_1(t)$, $M_2(t)$, $M_3(t)$ and $M_4(t)$ represents the expression level of transcription factor M_1, M_2, M_3 and M_4 at time t, respectively. Then the above differential equations are transformed into the corresponding systems of algebraic equations by the Laplace transformation.

3.2 Data Generation for Simulation

In order to evaluate our proposed algorithm for the model selection, we prepared three sets of artificial simulated time-series data which were considered to be experimental observations. The initial conditions for each molecules and the kinetic constants are set as follows: $M_1(0) = 10, M_2(0) = 7, M_3(0) = 3$, $M_4(0) = 1$, $k_{12} = 135/928$, $k_{23} = 1/29$, $k_{34} = 1/8$, $k_{14} = 13/928$ for feed-forward loop, $k_{11} = 1/11$, $k_{12} = 1/17$, $k_{13} = 1/21$, $k_{14} = 1/23$ for single input module, $k_{13} = 1/23$, $k_{14} = 1/25$, $k_{23} = 1/21$, $k_{24} = 1/27$ for dense overlapping regulons, respectively. By using kinetic constants, we sampled the data for examining the models. Since the digits of the constants are different in the above sets of equations, we sampled the data at 100 points when t is in the range from 0 to 10, at 100 points when t is from 10 to 30, and at 70 points when t is from 30 to 100. Furthermore, 5% of fluctuation is added for each data as the noise of data. Results of fitting by using RCGAs, three sets of generated data are fitted well to three different models.

3.3 Model Selection by Algebraic-Numeric Approach

To examine the performance of our method for three kinds of network motif, we selected one motif among the two motifs with the data generated from one model. In actual use of the present method, first, the data are observed by the experiments, and then

Table 1: Consistency measure with kinetic constants. The given values of kinetic constants are $k_{12} = 135/928(\sim 0.145)$, $k_{23} = 1/29(\sim 0.0345)$, $k_{34} = 1/8(\sim 0.125)$, $k_{14} = 13/928(\sim 0.0140)$ for feed-forward loop, $k_{11} = 1/11(\sim 0.0909)$, $k_{12} = 1/17(\sim 0.0588)$, $k_{13} = 1/21(\sim 0.0476)$, $k_{14} = 1/23(\sim 0.0435)$ for single input module, $k_{13} = 1/23(\sim 0.0434)$, $k_{14} = 1/25(\sim 0.0400)$, $k_{23} = 1/21(\sim 0.0476)$, $k_{24} = 1/27(\sim 0.0370)$ for dense overlapping regulons. The symbol '0*' indicates the exact value of zero.

data-generating model	examined model	smallest ssq	k_{11}	k_{12}	k_{13}	k_{14}	k_{23}	k_{34}
(a)	(a)	0.000907	-	0.149	-	0.00925	0.0345	0.125
(a)	(b)	0.537	0.0*	0.00666	0.0313	0.0844	-	-
(a)	(c)	0.531	-	-	0.0472	0.0755	0.0*	0.00285
(b)	(a)	1.52	-	0.0700	-	0.0*	0.0371	0.0*
(b)	(b)	0.00000177	0.0934	0.0593	0.0486	0.0446	-	-
(b)	(c)	0.0157	-	-	0.0295	0.0355	0.00557	0.0*
(c)	(a)	0.000689	-	0.0172	-	0.0672	0.111	0.00471
(c)	(b)	0.357	0.0572	0.0*	0.0728	0.0676	-	-
(c)	(c)	0.000184	-	-	0.0722	0.0108	0.0*	0.846

a model is selected among some candidates of models. Thus, we examine the performance of the present method by solving which models one set of data is consistent with.

Table 1 shows the consistency of the models with the three motifs by consistency measure, together with the estimated values of kinetic constants. As seen in the table, in the all cases, the consistency estimation was succeeded. The kinetic constants in network motifs are well estimated when the data are generated from network motif (a) and (b). Unfortunately, our method does not operate well about estimation of the values of kinetic constants, when the data are generated from model (c).

In summary, our method can identify the network motif from observed time-course data sets. Furthermore, our method also can estimate the value of kinetic constants well excluding dense overlapping regulons.

4 Discussion

We examined the performance of our method for selecting the model with three kinds of network motifs which are proposed as highly significant motifs in the transcriptional regulation network of *Escherichia coli*. We have perfectly succeeded in selecting the correct network motif by using consistency measure. This result shows that, by factorizing large-scale network to simple network motif, we could apply our proposed algorithm to analyze organizationally complex system.

Moreover, we have partly succeeded in estimating the kinetic constants in the network motifs. Note that the present performance is examined by one set of data generated from the given values of kinetic constants. At any rate, we should further test the performance of our method for the generated data by different kinetic constants as well as for actually observed data. Furthermore, we should test the performance of our method for various structures of motifs.

References

[1] Audoly, S., D'Angiò, L., Saccomani, M. P. and Cobelli, C.: Global identifiability of linear compartmental models — A computer algebra algorithm, *IEEE Trans. Biomed. Eng.*, Vol. 45 (1998), 36–47.

[2] Bernshtein, D.N.: The number of roots of a system of equations, *Functional Anal. Appl.*, Vol. 9(3) (1975), 183–185.

[3] Bisits, A. M., Smith, R., Mesiano, S., Yeo, G., Kwek, K., MacIntyre, D. and Chan, E. C.: Inflammatory aetiology of human myometrial activation tested using directed graphs, *PLoS Comput. Biol.*, Vol. 1 (2005), 132–136.

[4] Buchberger, B.: An Algorithmic Criterion for the Solvability of a System of Algebraic Equations, in Buchberger, B. and Winkler, F. eds., *Gröbner Bases and Applications*, London Mathematical Society Lecture Notes Series 251, Cambridge University Press, 1998, 535–545.

[5] Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F. and Inflammation and Host Response to Injury Large Scale Collab. Res. Program, : A network-based analysis of systemic inflammation in humans, *Nature*, Vol. 437 (2005), 1032–1037.

[6] Cobelli, C., Foster, D. and Toffolo, G.: *Tracer Kinetics in Biomedical Research: From Data to Model*, Kluwer Academic/Plenum Publishers, 2000.

[7] Cobelli, C. and Toffolo, G.: *Theoretical aspects and practical strategies for the identification of unidentifiable compartmental systems*, Pergamon Press, Oxford, 1987, chapter 8, 85–91.

[8] Hanzon, B. and Jibetean, D.: Global minimization of a multivariate polynomial using matrix methods, *Journal of Global Optimization*, Vol. 27 (2003), 1–23.

[9] Joreskog, K. G.: A general method for analysis of covariance structures, *Biometrika*, Vol. 57 (1970), 239–251.

[10] Meziane, D. and Shipley, B.: Direct and Indirect Relationships Between Specific Leaf Area, Leaf Nitrogen and Leaf Gas Exchange. Effects of Irradiance and Nutrient Supply, *Annals of Botany*, Vol. 88 (2001), 915–927.

[11] Ono, I. and Kobayashi, S.: A real-coded genetic algorithm for function optimization using unimodal distribution crossover, *Proc 7th ICGA*, (1997) 249–253.

[12] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, 1988.

[13] Satoh, H., Ono, I. and Kobayashi, S.: A new generation alternation model of genetic algorithm and its assessment, *J. of Japanese Society for Artificial Intelligence*, 15(2) (1997) 743-744.

[14] Shen-Orr, S. S., Milo, M., Mangan, S and Alon, U.: Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genetics*, Vol. 31 (2002), 64–68.

[15] Shipley, B.: A new inferential test for path models based on directed acyclic graphs, *Structural Equation Modeling*, Vol. 7 (2000), 206–218.

[16] Verschelde, J. and Haegemans, A.: Homotopies for solving polynomial systems within a bounded Domain, *Theor. Comp. Sci.*, Vol. 133(3) (1994), 165–185, (See also http://www.math.uic.edu/ jan/)

The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Time Series Segmentation for Gene Regulatory Process with Time-Window-Extension Technique

Zhi-Yong Zhang[1,2] Katsuhisa Horimoto[3] Zengrong Liu[2]

[1]Department of Mathematics, Shanghai University, Shanghai 200444, China
[2]Institute of systems biology, Shanghai University, Shanghai 200444, China
[3]National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

Abstract Many important Biological processes fall into different successive phases with piecewise time varying structures. To reveal the sequential regulatory relationship between different phases, time series segmentation is the first step toward elucidations of the underlying structure of GRN dynamics. In this paper, we aim to propose a new approach to solve this segmentation problem, called Time-Window-Extension Technique. Combined with clustering techniques, e.g. NMF method, we can produce the biological meaningful segmentation from time series expression profile, or identify the change points of nonstationary time series. Artificial data sets are also adopted to validate its effectiveness.

Keywords time series; cluster; NMF; segmentation; correlation matrix

1 Introduction

During the last few years, studying on Gene Regulatory Networks (GRNs) has drawn much attention due to recent rapid progress of high-throughput technologies which generate a vast amount of gene expression data. As a key control process of cells, GRNs are considered to be essential to regulate cellular processes and facilitate biological functions. A great number of papers have been published, and many computational methods and theoretical models have also been developed to infer the regulatory networks, e.g. Boolean networks, Bayesian networks, differential equations, data mining approaches etc.[8]. However, most of the above methods assume that the topologies of the Regulatory Networks are static[8], so the inferred networks are only the temporal profiling, which is actually not true for many biological processes.

Many important Biological processes, such as cell cycling, cellular differentiation during development, aging, and disease aetiology, are regulated not by a stationary GRN but a time-varying one [3, 7]. Furthermore, it has been recognized that the regulatory pathway does not always persist over all the time. In particular, an important experimental result [1] has confirmed that the topologies of GRNs change depending on the underlying condition. The present clues converge on the time-varying GRNs. However, due to the lack of data availability and status quo of methods, reconstruction of regulatory networks

with time-vary structures is still not a tractable problem from computational viewpoint [3]. Fortunately, it has been observed that many biological processes are actually phase-dependent, rather than complete time-varying. In other words, a GRN for many cases can be viewed as a piece-wise stationary structure. Therefore, instead of full time-varying GRN, we can reconstruct phase-specific GRNs, which requires much less data and can be inferred in a more reliable way.

At the same time, the huge amount of large-scale and genome-wide time series expression data provides a great opportunity to reveal the phase-specific GRNs, which are becoming increasing available in recent years. The time series analysis plays a crucial role in the study of disease progression [5], and cyclical biological processes, e.g., the cell cycle[1, 2], metabolic cycle[6], and even entire life cycles[7]. Recent efforts have considered inferring the direct regulatory relationship between different phases[4]. In this paper, we aim to identify the change points and reveal the relationship between different biological processes, especially the sequential biological processes based on time series analysis. Specifically, in this paper, we first identify where are the change points (or checkpoints) to separate the different phases of the biological processes. To solve this problem, we partition the time series expression profile to obtain the temporal segments in an automatical manner, based on the clues of changing of genes clusters. Then the "direct" regulatory relationship between these segments (or phases) is be inferred, which is believed to be essential for understanding of the underlying structure of regulatory network dynamics. The numerical example is also provided to verify the effectiveness of the proposed method.

2 Methods

Given time series gene expression data $X = [g^1, g^2, \cdots, g^n]$, each $g^i \in \mathbb{R}^l$ is a l-vector of gene i's expression profile $[g^i_1, g^i_2, \cdots, g^i_l]^\top$, which is from a time series of measurements over time points $\tau = \{t_1, t_2, \cdots, t_l\}$. The gene i's expression profile at the jth time point is denoted by g^i_j. For a time window $W^e_s = \{t_s, t_{s+1}, \cdots, t_e\}(t_s < t_e)$, which is a sequence of consecutive time points, the "windowed" time series data of gene i's expression profile is denoted by ${}^e_s g^i = [g^i_s, g^i_{s+1}, \cdots, g^i_e]^\top$, and the "windowed" time series data of the total n genes' expression profiles are denoted by ${}^e_s X = [{}^e_s g^1, {}^e_s g^2, \cdots, {}^e_s g^n]$.

Within the windows, we can cluster the genes based on their similarity of expression profiles. The concerted behavior of the genes in the clusters may be caused by the same regulatory factors, such as TFs. Around the checkpoint, i.e. the boundary of two successive phases, the association of the expression behavior of genes will change, which may be triggered by some underlying inputs, such as TFs, or result in new phase or regrouping of genes. Actually, we can identify these checkpoints or the boundaries of the phases by analysis of the regrouping of clusters.

2.1 Clustering over time windows

Given the windowed time series gene expression data ${}^e_s X \in \mathbb{R}^{m \times n}(m = e - s + 1)$, the NMF(non-negative matrix factorization) method[9, 11] is employed to find the gene clusters. The problem is formulated as follows:

$$ {}^e_s X \approx WH $$

where $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$ are non-negative matrices, and r is the predefined number of clusters. The gene assignment depends on the relative values in each column of H, that is to say, if h_{ki} is the maximum element of the column h_i, then gene i is assigned to the cluster k.

The NMF method does not converge to the same solution on each run, depending on the random initial conditions. For each run, the gene assignment can be represented by a connectivity matrix $C \in \mathbb{R}^{n\times n}$, with entry $c_{ij} = 1$ if genes i and j belong to the same cluster, and $c_{ij} = 0$ if not. In this paper, we then compute the average connectivity matrix over multiple runs, $\overline{C}$. We continue the iterative computations (or runs) until $\overline{C}$ appears to converge. The entries of $\overline{C}$ reflect the probability that genes i and j cluster together, ranging from 0 to 1 [11].

We then recover the final clustering solution with the spectral clustering method [10], which is the most consistent to the average connectivity matrix $\overline{C}$.

2.2 Segmentation Algorithm

Given two windowed time series data ${}^{e_1}_{s_1}X$ and ${}^{e_2}_{s_2}X$, let the average connectivity matrices be denoted by $\overline{C}^1$ and $\overline{C}^2$ respectively, which can also represent the clustering results. We introduce the correlation matrix as follows:

$$T = (t_{ij})_{n\times n} = \rho(\overline{C}^1_i, \overline{C}^2_j)$$

where $\rho(\cdot,\cdot)$ is the correlation coefficient between random variables of $\overline{C}^k_i = [\overline{C}^k_{i,1}, \cdots, \overline{C}^k_{i,i-1}, \overline{C}^k_{i,i+1}, \cdots, \overline{C}^k_{i,n}]^\top \in \mathbb{R}^{n-1}, k = 1,2$. Note that the diagonal elements $\overline{C}^k_{i,i} (i = 1, \cdots, n; k = 1,2)$ are omitted in the above definition due to $\overline{C}^k_{i,i} \equiv 1$. The element t_{ij} of the matrix T represents the correlation coefficient between the genes i's connection vector in one window and the genes j's connection vector in the next one. Specially, the element t_{ii} indicates the relationship of gene i's connectivity between different time windows, and thus provides a measure of the cluster-regrouping behavior of gene i.

The correlation matrix captures the topological change of networks denoted by the average connectivity matrix, and provides a new method to capture the regrouping of the clusters of genes over different time windows, which is more appropriate than the previous methods such as contingency matrix[6]. The diagonal elements of matrix T will be close to 1 if the genes possess the similar average connectivity matrix in two different windows, and the diagonal elements of matrix T will be close to 0 if the genes undergo the cluster-regrouping process. Here we propose a quantitative measure of the cluster-regrouping process as follows:

$$\mathscr{F}(\overline{C}^1, \overline{C}^2) = \frac{1}{n}\sum_{i=1}^{n} |t_{ii}|.$$

For two successive (or consecutive) time windows, the problem of segmentation is then to minimize $\mathscr{F}$ as the criterion function.

We develop a new approach to the segmentation problem by turning it to the problem of boundary determination, and we call it the time-window-extension technique, as illustrated in Figure 1. Given the time window W^e_s and its extension $W^{e'}_s, e' > e$. If they are

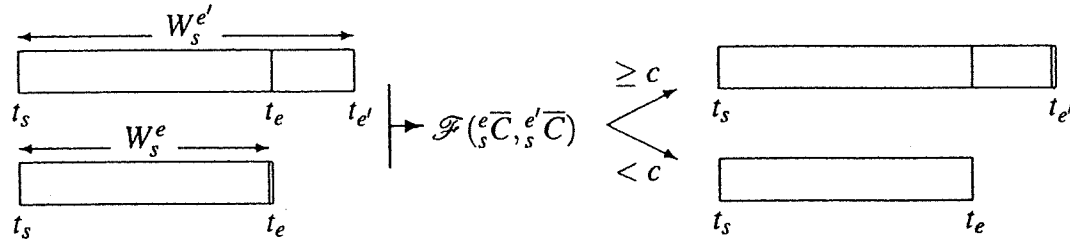


Figure 1: The extension procedure of the time window(thickline: checkpoint; single-line: extended boundary; double-line: putative boundary)

both parts of the same segment, then the clustering results will be similar, i.e. the diagonal elements of the correlation matrix T will be close to 1 such that $\mathscr{F}$ will be close to 1. On the other hand, if there is a boundary between e_1 and e_2, then the diagonal elements of the correlation matrix T will deviate from 1 such that $\mathscr{F}$ will decrease towards 0. Clearly, we can capture the change point by using $\mathscr{F}$ as the criterion function, thereby identifying the boundary of the segment by extending the window in a systematical manner (see figure 1).

The computational steps in detail can be described as follows:

1. Given the left boundary t_s and the postulated right boundary t_e. Note that the minimum time window length should be predefined such that, for example, $e - s \geq 2$.
2. Calculate the average connectivity matrix for ${}_s^eX$, denoted by ${}_s^e\overline{C}$.
3. Extend the right boundary to $t_{e'}$ and calculate the average connectivity matrix for ${}_s^{e'}X$, denoted by ${}_s^{e'}\overline{C}, e' > e$. Note that the minimum extension length should be predefined too.
4. Calculate the criterion measure $\mathscr{F}({}_s^e\overline{C}, {}_s^{e'}\overline{C})$.
5. If $\mathscr{F}$ is larger than the cutoff value c predefined, set $t_{e'}$ as the new postulated right boundary, and goto step a.
6. If $\mathscr{F}$ is less than the cutoff value c, the right boundary can be found between t_e and $t_{e'}$. Reduce the extension length, and goto step c.

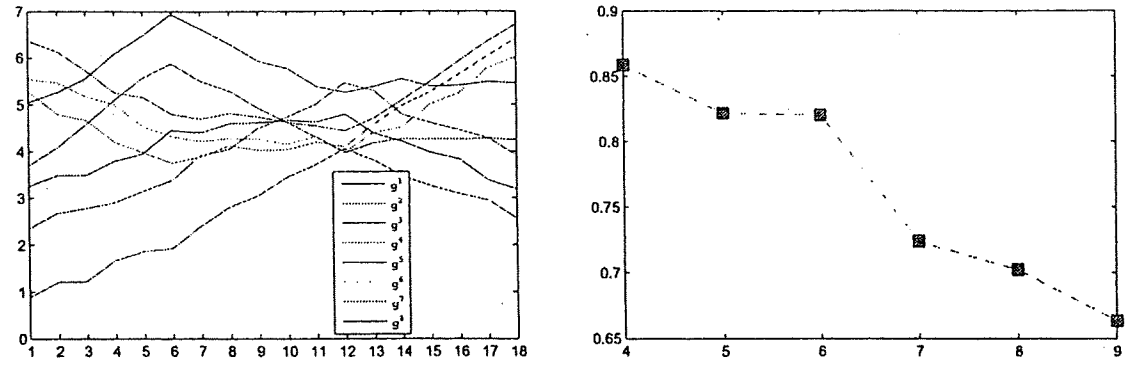
2.3 Inferring directed Cluster-Cluster Regulations using Graphical Gaussian Model

Based on the temporal segmentations (phases), we next infer directed cluster-cluster regulations between consecutive phases or reconstruct the gene regulatory network among clusters. In particular, we adopt Gaussian Graphical Model (GGM)[12] to infer the direct regulatory relationship of these clusters between different phases. The detail description will be given and discussed in another paper.

2.4 Numerical Simulation for An Artificial Case

We provide a case study where the time-window-extension technique proposed in the paper is applied to an artificial gene expression data set with 8 genes and 18 time points (3 phases).

Figure 2(a) shows the gene expression profiles in different time points generated from the artificial data set. Figure 2(b) shows the evolution of $\mathscr{F}$ during the first time window extension (on the purpose of identifying the first checkpoint), namely, the evolution



(a) (b)

Figure 2: Simulation result. (a) the artificial genes expression profiles; (b) the evolution of $\mathscr{F}$ during the window extension for the first and second phases, i.e. identify the first checkpoint.

of $\mathscr{F}({}_1^3\overline{C}, {}_1^n\overline{C}), n = 4, 5, \cdots, 9$, based on the proposed procedure. From figure 2(b), clearly the first segment extends to time point 6 with cutoff 0.75 for $\mathscr{F}$, which agrees with the observation from Figure 2(a). Based on our algorithm, all of the three phases were correctly identified.

3 Conclusion

In this paper, we developed a new computation procedure to solve this segmentation problem for nonstationary time series data. Based on clustering technique and a new criterion, we can produce the biological meaningful segmentation from time series expression profile by identifying the change points of nonstationary time series. The proposed method in this paper was employed to the artificial gene expression data set which were generated with unambiguous structure of clusters and clear-cut segmentation. The numerical simulation confirms the effectiveness of the method. As a future topic, we will test our method to the real gene expression profiles to further identify the phase-dependent structure of GRN.

Acknowledgement

The authors thank Prof. Luonan Chen for helpful discussions and suggestions.

References

[1] Luscombe, N. M. et al., Genomic analysis of regulatory network dynamics reveals large topological changes, Nature, 2004, Vol. 431, p308-312.

[2] Spellman, P.T. et al., Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization, Molecular Biology of the Cell, 1998, Vol. 9, p3273-3297.

[3] Rao, A. et al., Inferring Time-Varying Network Topologies from Gene Expression Data, EURASIP Journal on Bioinformatics and Systems Biology, Volume 2007, Article ID 51947.

[4] Aburatani, S., Saito, S., Toh, H., Horimoto, K., A graphical chain model for inferring regulatory system networks from gene expression profiles, Statistical Methodology, Volume 3, Issue 1, 2006, p17-28.

[5] Kleinberg, S. et al., Systems biology via Redescription and Ontologies: Untangling the Malaria Parasite Life Cycle, 2007, International Conference on Life System Modeling and Simulation, Shanghai, China.

[6] Tadepalli, S. et al., Simultaneously Segmenting Multiple Gene Expression Time Courses by Analyzing Cluster Dynamics, 2008, Asia Pacific Bioinformatics Conference, Kyoto, Japan.

[7] Li, X. et al., Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling, BMC Bioinformatics, 2006, 7 : 26

[8] Ma, P. C. H. Ma et al., Inference of Gene Regulatory Networks from Time Series Expression Data: A Data Mining Approach, 2006, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)

[9] Lee, D. D. et al., Algorithms for Non-negative Matrix Factorization, Advances in Neural Information Processing Systems,2001,13:556-562.

[10] Gong, Y. et al., machine learning for multimedia content analysis, 2007, springer.

[11] Brunet, J.-P., et al., Metagenes and molecular pattern discovery using matrix factorization, PNAS, 2004, vol. 101, no. 12, 4164-4169

[12] Aburatania, S. et al., A graphical chain model for inferring regulatory system networks from gene expression profiles, Statistical Methodology, 2006, 3, 17-28

The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Revealing Disease Related Interactions by Correlation Analysis

Zi-Kai Wu[1,2] Zhi-Yong Zhang[1,2] Lv-Wen Zhang[1,3]
Katsuhisa Horimoto[4]

[1]Institute of Systems Biology, Shanghai University, Shanghai 200444
[2]School of Communication and Information Engineering, Shanghai University, Shanghai 200444
[3]School of Computer Engineering and Science, Shanghai University, Shanghai 200444
[4]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Abstract The computational identification of disease related lesions is still a key open problem in biomedicine and systems biology. Dysregulated interactions may be an important reason that causes disease. In this paper, we aim to identify dysregulated interactions so as to elucidate the mechanism of disease in a systematic manner. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated due to disease process. The proposed method was applied to a human molecular interaction network and a prostate cancer microarray dataset to reveal dysregulated interactions. The enrichment analysis of cancerous genes and disease related GO terms in identified dysregulated interactions shows that the identified dysregulated interactions are disease related, which verifies the effectiveness of our method.

Keywords Network; Disease; Interaction; Correlation

1 Introduction

Life is a complex phenomenon, which cannot be clearly understood by merely studying individual components of cells. It is the interactions of those components or networks that ultimately hold responsibility of living organisms' forms and functions. Due to the recent rapid progress on biomedical science, the fundamental mechanisms on many diseases have be revealed at molecular level. For example, it has be elucidated that many cancers originate from some mutations on certain genes caused by chance or experimental factor because these mutations trigger downstream effect to the cellular system, i.e. on genes, proteins, partial pathway or entire pathway [1]. From the viewpoint of network biology, a disease can be viewed as a perturbation to the cellular system or biomolecular interaction network. In other words, the cellular system under disease state is a disturbed system which is rewired from the original undisturbed system (or control state) accordingly. As disease is considered to perturb the cellular system from the aspect of node and edge (connectivity), computational method of identifying disease related lesions can be grouped into two classes naturally, i.e. node-centric method and edge-centric method.

At present, computational identification methods are mainly node-centric. Take cancer research as an example. Until now, a number of methods have been proposed to