Lecture Notes in Operations Research 9

Series Editors:
Dr. Zhu Da and Xiang Sun Zhang

OPTIMIZATION AND SYSTEMS BIOLOGY

[illegible] ONEIOS
[illegible] November 3, 2008

[illegible] CORPORATION

The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
, pp. 109–114

Phase Shifts of Circadian Transcripts in Rat Suprachiasmatic Nucleus

Ryoko Morioka[1,2] Masanori Arita[1,2,3] Katsuhiko Sakamoto[4]
Soshi Kawaguchi[4] Hajime Tei[4] Katsuhisa Horimoto[4,*]

[1]Computational Biology Research Center, Advanced Industrial Science and Technology, Koto, Tokyo, Japan
[2]Plant Science Center, RIKEN, Yokohama, Kanagawa, Japan
[3]Department of Computational Biology, University of Tokyo, Kashiwa, Chiba, Japan
[4]Mitsubishi Kagaku Institute of Life Science, Machida, Tokyo, Japan

Abstract We analyzed the phase shifts of oscillated transcripts relative to the timing of phase-reset stimuli. Oscillations in gene expression profiles were extracted using the fast Fourier transform and fitted by sine curve with random periods. The phase differences among multiple phase-reset conditions were analyzed to elucidate the mechanism of the core circadian clock. For the expression profiles, cultured cells of rat suprachiasmatic nucleus were measured by Affimetrix GeneChip system. Folskolin stimulus was used as a phase-reset agent, causing irregular shift characteristic to oscillatory transcripts. The results suggest that the fluctuations of gene expressions in the core clock fall into two major categories and can be shifted by folskolin. Other clock related genes might adjust their oscillation by counter-steering each other in the feed-back regulation system.

1 Introduction

The master pacemaker of circadian rhythm in mammalians resides in suprachiasmatic nucleus (SCN), where a transcriptional-translational autoregulatory loop generates molecular oscillations of the "central clock" [1]. The variations of the oscillating expression profiles of clock genes tell the regulatory motion of the transcriptional-translational feedback loop of the clock system caused by external stimuli.

In rat circadian system, the free-running period is 24.5 hours. It is adjusted to 24 hours by entraining agents like light and temperature [2][3][4]. In *vivo*, the average of free-running period of SCN cells is 24 hours and the average period is kept by interaction among SCN cells [5]. In cultivated cells of SCN, on the other hand, the circadian period is 27 hours, and forskolin stimuli can reset the clock of the cultivated cells [2][6][7][8][9]. In this study, we analyzed the periodic fluctuation of SCN cultivated cells under different timing of forskolin stimuli.

*To whom correspondence should be addressed.

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.

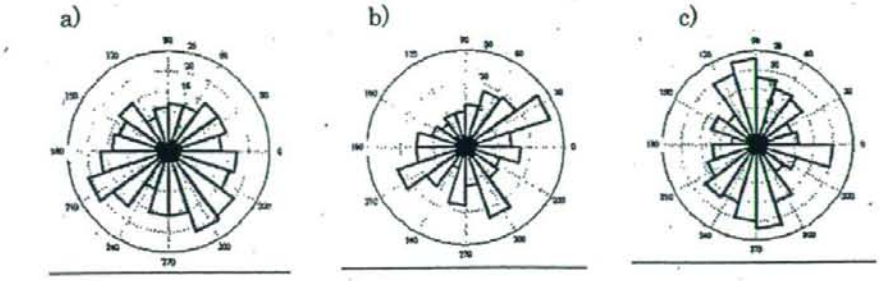


Figure 1: The phase-difference distribution of a) CT6 b) CT14 and c) CT22 among oscillating 300 genes

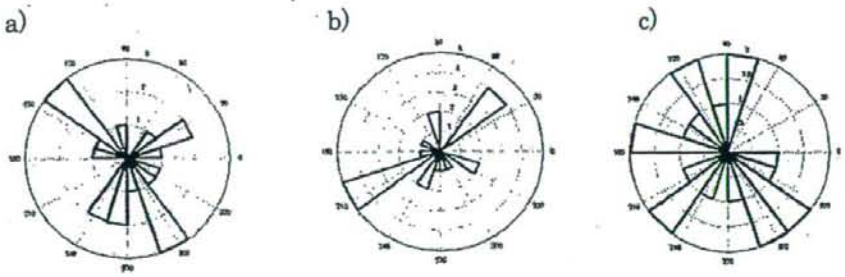


Figure 2: Phase-difference distribution of known clock genes under the condition of a) CT6 b) CT14 and c) CT22

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 forskolin 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.

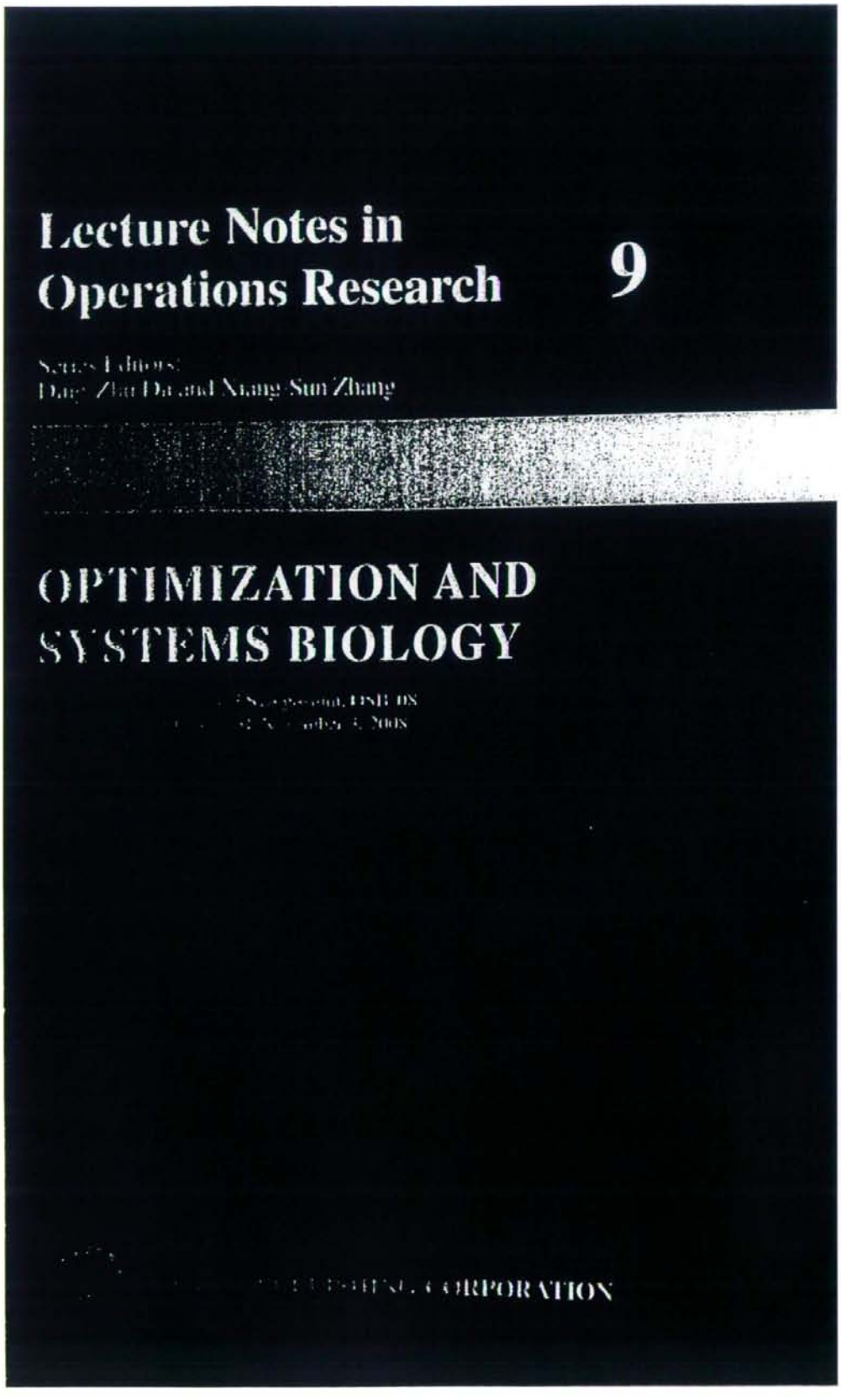
[17] EF. Glynn, J. Chen, AR. Mushegian, Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms, *Bioinformatics*, 22, 2006, 310-316.

[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 Biolоby*, 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


Lecture Notes in
Operations Research
9
OPTIMIZATION AND
SYSTEMS BIOLOGY

The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
 pp. 198–203

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 $[g_1^i, g_2^i, \cdots, g_l^i]^\top$, 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_j^i. For a time window $W_s^e = \{t_s, t_{s+1}, \cdots, t_e\}(t_s < t_e)$, which is a sequence of consecutive time points, the "windowed" time series data of gene i's expression profile is denoted by ${}_s^e g^i = [g_s^i, g_{s+1}^i, \cdots, g_e^i]^\top$, and the "windowed" time series data of the total n genes' expression profiles are denoted by ${}_s^e X = [{}_s^e g^1, {}_s^e g^2, \cdots, {}_s^e g^n]$.

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

$$ {}_s^e X \approx WH $$

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

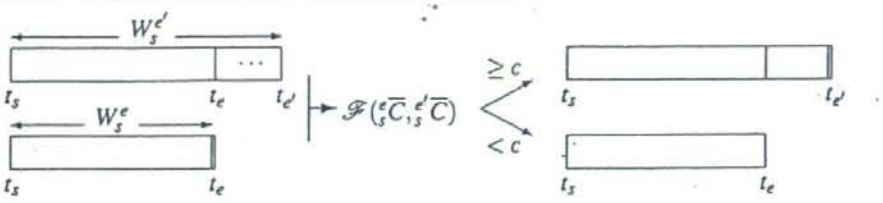


Figure 1: The extension procedure of the time window(thickline: checkpoint; single-line: extended boundary; double-line: putative boundary)

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.

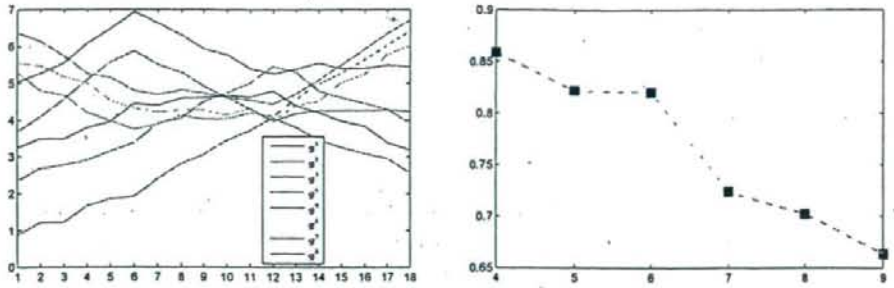
2.3 Inferring directed Cluster-Cluster Regulations using Graphical Gaussian Model

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

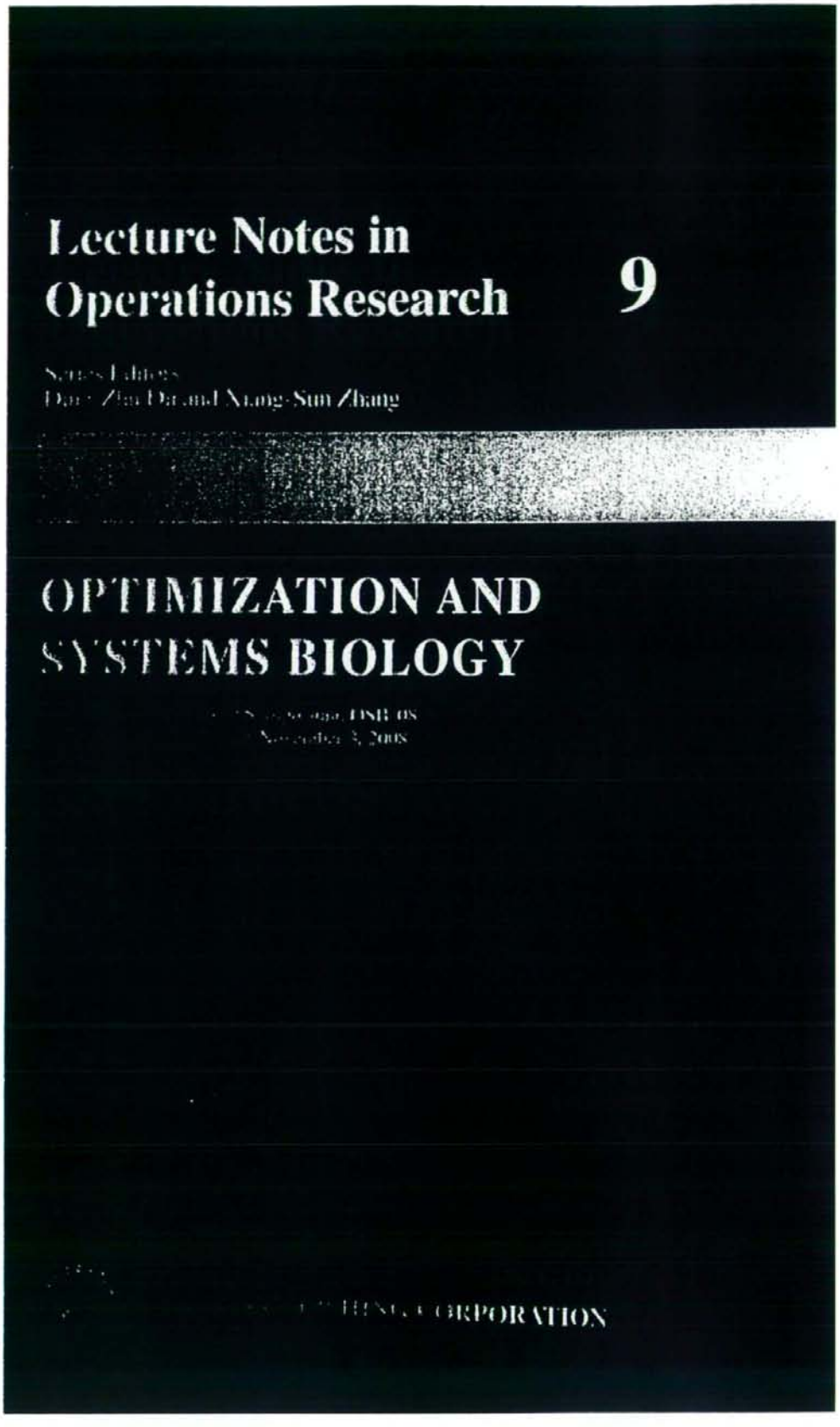
[8] Ma, P. C. H. Ma et al., Inference of Gene Regulatory Networks from Time Series Expression Data: A Data Mining Approach, 2006, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)

[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


Lecture Notes in
Operations Research
9
Series Editors
Ding-Zhu Du and Xiang-Sun Zhang
OPTIMIZATION AND
SYSTEMS BIOLOGY
HUSCO CORPORATION

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

identify cancer related genes. At earlier stage, most of these methods were based on differential expression analysis. In other words, the aberrantly expressed genes are identified as cancer related lesions. Although partical success has made in identifying cancer related genes, these methods are unable either to infer any details on how a protein's behavior has changed or to reveal what specific mechanisms lead to the pathologic transition [2].

To overcome this drawback, some gene-centric identification methods have been developed to embed themself in the context of cellular network. These methods utilize the known disease relatedness of other nodes in the cellular network to infer some node's disease relatedness. The rationale is that if some neighbors (direct or indirect neighbors) of a gene are disease related, then the gene can also be inferred to be disease related with certain confidence [3]. With such a scheme, Kartik M Mani et al. proposed a novel identification method [2]. Their analysis method works in two steps. That is, this method first identifies dysregulated interactions (interactions showing either a gain of correlation or a loss of correlation pattern) in the phenotype of interest, and then ranks genes according to the statistical significance of dysregulated interaction enrichment among the interactions in which they directly participate [2]. This method's rationale is that if a node or gene's relation with most of their neighbors are changed under the disease state, then it can be inferred with high confidence that the gene itself is arch-criminal and disease related.

Some other gene-centric identification methods aim at the entire pathway or a prior defined gene set[4,5,6,7]. Pathway-based methods use a metric to measure the cohesiveness level of the members of the pathway and represent the tightness of relation between its members. Their rationale is that if the cohesiveness level is descended or elevated under disease state, the pathway can be viewed as a disrupted or newly constructed subsystem under disease state [5]. Gene set-based methods first use some metric to measure the differential expression level of each gene and then a ranked list of differentially expressed gene is obtained. Enrichment analysis of differentially expressed gene in the prior defined gene set is conducted to find which gene set's overall differential expression level is statistically significant [6, 7].

At present, there are also some edge-centric computational identification methods, for which the key is how to define the edges between nodes in the network and how to capture the differential behavior of the edge. Essentially, the definition of edge depends on the data at hand. High-throughput technologies are now producing vast amounts of biological data representing the availability of specific molecular species in a cellular population[2]. These include, among many others, gene expression and genotypic profiles [8], DNA-binding profiles[9], genomic sequences, and protein abundance from mass spectrometry [10]. At the same time, another high-throughput experiments have populated the public databases with thousands of protein-protein interaction (PPI) data and genetic interaction data[11].

Some researchers use the gene co-expression to define the edge between genes [12,13 ,14,15,16]: if two genes's mRNA expression levels are highly correlated under certain condition, then it can say that there is a functional association between two genes, in other words, there exists an edge between two genes. Jung-Kyoon Choi et al.[12] constructed a normal and disease coexpression network respectively based on 10 cancer microarray datasets and 10 their normal counterparts, and then identified the differential coexpression in the network. There are also some other methods based on differential co-expression analysis that was proposed to identify disease related lesions[13,14,15,16].

Other research works use the physical PPI and genetic interaction to identify disease related edges. One weakness of the high-throughput PPI data and genetic interaction data is that it contains no information about the conditions under which the interactions may take place[17]. Under the hypothesis that higher expression correlation of the genes implies genuine interactions of the proteins under the investigated conditions, it is a popular way to use the gene expression information to measure the 'activity 'of an interaction in response to the investigated condition. Zheng Guo et al.[17] scored the edge in PPI network based on the correlation coefficient of two genes's expression levels and the deferential expression of two genes, and then used simulated annealing algorithm to find a statistically significant responsive subnetwork.

On the other hand, protein-protein interaction have recently been recognized as challenging but attractive targets for small chemical drugs[18]. Furthermore, recent research works suggest that PPI inhibition could lead treatments for some human disease[18-23]. Motivated by both the potential pharmaceutic and therapeutic applications of disease related interactions and sparseness of computational methods for identifying disease related PPI or genetic interactions, we propose a new method to identify dysrgulated interactions by exploiting the mechanism of diseases in this paper. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated during disease process.

The remainder paper is organized as follows. Firstly, we describe the details of our method as well as the data set we used. Secondly, the results are presented through numerical tests on prostate cancer case. Finally, the features for the new method of identifying disease related interaction are discussed, and a brief conclusion and directions of further research works are presented in the last section.

2 Methods and materials

2.1 Dataset and data processing

The protein-protein interaction and genetic interaction data was first derived from the BIOGRID database(2008, 2.0.36 version). Then the self-interactions and reduplicate interactions were removed from the dataset. Finally, we have 23791 interactions in the interaction data set, which constitute a protein interaction network.

The prostate microarray data set [24] consists of about 7641 genes measured in 71 prostate tumors as well as 41 normal prostate specimens. In the microarray dataset, if there are multiple probes that correspond to the same gene, we choose the one that contains the least amount of missing values. Then, we only retain genes with missing data smaller than one third of the total sample size. Finally, we convert all values $<= 10$ to 10, and then perform a base 2 log transform. The prostate cancer related genes were obtained from Prostate Gene Database (PGDB)[25].

2.2 Estimation of pairwise gene co-expression

In this paper, the Percentage Bend Correlation [26] with $\beta = 0.1$ is applied to obtain a robust correlation estimate. Percentage Bend Correlation is first adopted to detect outliers in expression values of each gene so as to reduce the effects of those outliers in the correlation calculation[15]. Since the Percentage Bend Correlation may have some bias due

to sample size, Fisher's z-transform [27] is also performed to reduce sample size effect, which can be formulated as

$$Z = \frac{\sqrt{n-3}}{2} \times \log\sqrt{\frac{1+r}{1-r}} \quad (1)$$

where r and n denote correlation estimate and sample size respectively, while Z corresponds to the Fisher's Z scores. Z score divided by its theoretical standard deviation theoretically has an asymptotically standard normal distribution. However, Min Xu et al. observed that the distributions of the z-score are still different from dataset to dataset [15]. Hence, we further normalize z-scores to enforce the standard normal distribution. After that, standardized correlations r' are obtained by inverting the z-score with a fixed n of 30 as Min Xu did.

2.3 Active interactions under certain condition

We give different definition of active interaction with respect to physical protein-protein interaction and genetic interaction. Suppose a physical protein-protein interaction connects gene A and gene B in cellular interaction network. We define the interaction to be active under normal state if the expression correlation of gene A and gene B in normal data set is higher than some threshold (in this paper, the threshold is set to be 0.20). Otherwise, the physical interaction between A and B are defined as inactive. For genetic interaction, we define it to be active under normal state if the absolute value of its two genes's expression correlation is higher than some threshold. Otherwise, the genetic interaction between A and B are defined as inactive. Similarly, we can define how an interaction is active or inactive under disease state.

2.4 Downregulated and upregulated interactions under disease state

We define an interaction to be upregulated if it is inactive in normal state but active under disease state. We define an interaction to be downregulated if it is active in normal state but inactive under disease state.

2.5 Enrichment analysis

The GO term enrichment analysis is done by the hypergeometric test on genes involved in downregulated interactions and upregulated interactions respectively through submitting them to DAVID online webserver(http://david.abcc.ncifcrf.gov /home.jsp). The prostate cancer and cancer related gene enrichment analysis are also done by the hypergeometric test.

Finally, the whole procedure of the method is summarized as Figure 1.

3 Results and discussion

Under the different thresholds, there are different numbers of interactions being active under normal state or disease state. In this paper, we present the result obtained when setting threshold being 0.20.

Under the threshold of 0.20, there are 1289 interactions that are active under normal state, while there are 1310 interactions that are active under disease state. Accordingly,

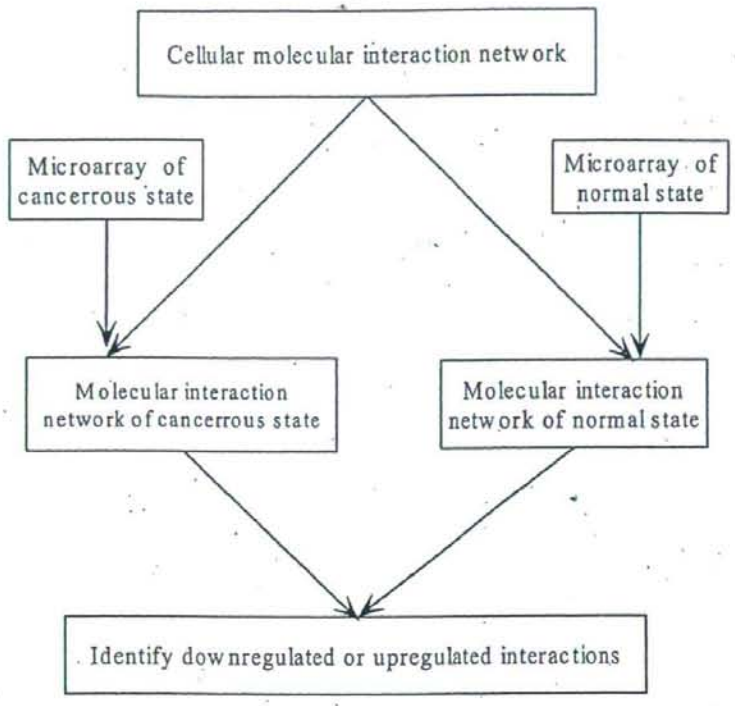


Figure 1: Flowchat of the proposed method

there are 213 interactions that are upregulated and 228 interactions that are downregulated. To evaluate the biological relevance of this identified dysregulated interactions, we perform some enrichment analysis. Firstly, the identified dysregulated interactions involve many genes. If these genes are cancer related, then we can infer that these interactions are also cancer related to some extent. There are 327 genes involved in upregulated interactions, of which 17 genes are known cancer related. There are 337 genes involved in downregulated interactions, of which 18 genes are known cancer related. Furthermore, there are 8042 genes involved in the interaction network. The known 118 cancer related genes that are included in these 8042 genes are used as background. We performed enrichment analysis on genes involved in downregulated and upregulated interactions respectively. The p-value of enrichment analysis is $4.9685e-006$ and $1.7217e-006$ respectively. The small p value shows that the enrichment of cancer related genes on the identified dysregulated interactions is statistically significant, and the identified dysregulated interactions are biological relevant and cancer related.

To further verify its biological relevance and cancer relatedness, we also performed the enrichment analysis of GO terms on the identified dysregulated interactions. There are many GO terms that are enriched. In this paper, we only present GO terms belonging to biological process category for the sake of simplicity. Some representative GO terms are listed in Tables 1 and 2 respectively. Enriched GO terms on downregulated inter-