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 ${}_{s_1}^{e_1}X$ and ${}_{s_2}^{e_2}X$, 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:

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

where $\rho(\cdot,\cdot)$ is the correlation coefficient between random variables of $\overline{C}_i^k = [\overline{C}_{i,1}^k, \cdots, \overline{C}_{i,i-1}^k, \overline{C}_{i,i+1}^k, \cdots, \overline{C}_{i,n}^k]^\top \in \mathbb{R}^{n-1}, k = 1,2$. Note that the diagonal elements $\overline{C}_{i,i}^k (i = 1, \cdots, n;$ $k = 1,2)$ are omitted in the above definition due to $\overline{C}_{i,i}^k \equiv 1$. 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_s^e and its extension $W_s^{e'}, e' > e$. If they are

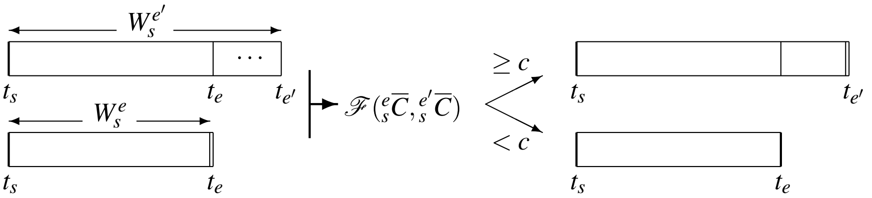


Figure 1: The extension procedure of the time window(thickline: checkpoint; single-line: extended boundary; double-line: putative boundary)

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.

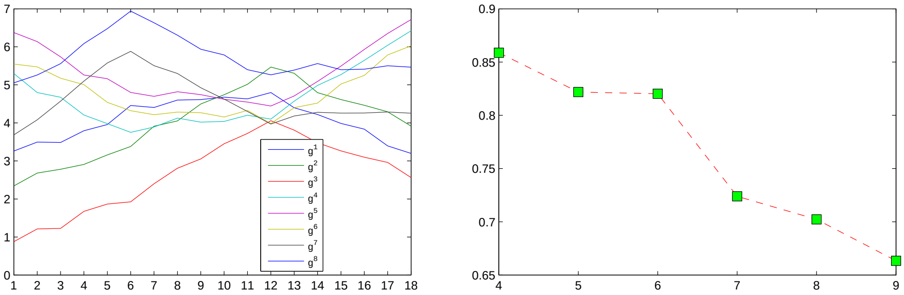
2.3 Inferring directed Cluster-Cluster Regulations using Graphical Gaussian Model

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 $\mathscr{F}({}^3_1\overline{C}, {}^n_1\overline{C}), n = 4,5,\cdots,9$, 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