The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Uncover the Transient Transcriptional Regulations by a Novel Sliding Window Correlation Strategy*

Jiguang Wang[1] Yong Wang[1] Xianwen Ren[1] Wei Guo[2]
Luonan Chen[2,†]

[1]Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing, China.
[2]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.

Abstract

Uncovering the regulation of gene transcription is critical to understand the mechanisms of biological systems. Thanks to the rapid data accumulation by new efficient experimental technologies, many computational methods have been proposed to decipher the transcriptional regulation. However, one of the common shortcomings for most of the previous methods is the ignorance of edge dynamics, i.e., the regulatory relationships between regulators and their targets are intrinsic dynamic both in temporal and spacial. With this in mind, we propose a new sliding window based strategy here to uncover the edge dynamics in transcriptional regulatory network from time-series microarray data. To show the efficiency, we apply our new approach in the yeast cell cycle data. Numerical experiments show that the new strategy achieves a better accuracy in identifying active edges. Particularly, the new approach can identify new edges which are active in particular phase instead of the whole time course. More significantly, some identified edges can even switch from positive regulation to negative regulation in different time phases, which confirms the importance of considering edge dynamics. In addition, we introduce the concept of edge profile describing the dynamic nature of regulation based on the sliding window strategy, which is useful for understanding cell differentiation, cell pathological changes and other cytologic atypia.

Keywords Edge dynamic; Time-series data; Sliding-window correlation; transcriptional regulatory network; Microarray

1 Introduction

Living cells are in some sense the products of gene expression programs. Hence uncovering the regulation of gene transcription is critical to understand the mechanisms of biological systems. Gene expression physically depends on recognition of specific promoter sequences by transcriptional factors (TFs). The technique of ChIP-chip that combines chromatin immunoprecipitation (ChIP) with microarray (chip) allows the identification of genome-wide location analysis for given TFs [1]. *Saccharomyces cerevisiae* is one

*The paper is supported by the NSFC grants 10631070, 10801131 and 2006CB503905 from MST of China.
†Corresponding author: lnchen@sibs.ac.cn

of the well studied model organisms. Several large scale experiments have been carried out to genome-widely identify binding motif of TFs [2, 3] in *Saccharomyces cerevisiae*, which is one of the well-studied model organisms. As a result, the yeast transcriptional regulatory network (TRN) has been constructed by adding directed edges from genes encoding TFs to the target genes being regulated. TRN provides a whole picture and potential pathways that yeast regulates gene expression, which is important in understanding how yeast grows, divides, and response to stimulation [2].

ChIP-chip experiments can easily identify binding motifs of TFs, but it is difficult for this technique to tell when, where, and under which condition TFs regulate genes. More comprehensive understanding of TRN requires knowledge of expression information at different time points or different conditions. Based on the fact that the expression of TFs (at least at some conditions) is correlated with that of target genes [4], a number of reverse engineering methods have been developed to infer TFN combining with microarray data, including dynamic Bayesian network methods [5], optimization based methods [6], information theory based methods [7], and so on. All of these methods make use of the observed gene profile to explore the intrinsic relationships among genes.

A common assumption of the previous approaches is the regulatory relationship does not change in given conditions, but this is usually not valid in practice. For example, in the cell division cycle, different series of events take place in different periods. The cell accumulates nutrients for mitosis and duplicating its DNA in interphase, while it splits itself into two cells in mitosis (M) phase. Regulators in charge of DNA duplicating do not necessarily regulate the same set of target genes at M phase. As a result, some TFs correlate with their targets in a local time interval instead in the whole time interval. More importantly, some TFs even change their roles from activators to repressors when regulating targets. An example is E2f1-3, which switches from activators in progenitor cells to repressors in differentiating cells [8].

In this paper, we firstly check the consistency between two types of data: ChIP-chip data and microarray data. Edges observed by ChIP-chip are considered as active when the interacting partners are significantly associated in microarray data. We then show the observations that just a few ChIP-chip edges are active assessed by traditional Pearson correlated coefficient (PCC) method. Furthermore, we analyze possible reasons. Secondly, we develop a new sliding-window based correlation (SWPCC) approach to assess the ChIP-chip edges. The new approach identifies more active edges by considering edge dynamics in time-series data. More importantly, the concept of edge profile is introduced and defined to explain why our new approach works.

2 Result

2.1 The Sliding Window Based Strategy

Traditional methods take available time points as a whole to infer regulatory relationship between regulators and targets, however the regulatory interactions are transient as we mentioned. To capture the edge dynamics, we design a sliding window based approach to assess the regulatory relationships at each time point. One of the features of time-series data is the dependence of adjacent time points, so we make use of time points within a small time interval to infer the relationship between one given TF-target pair at a particular time point. As shown in Figure 1, an L (L is odd) length sliding window slides

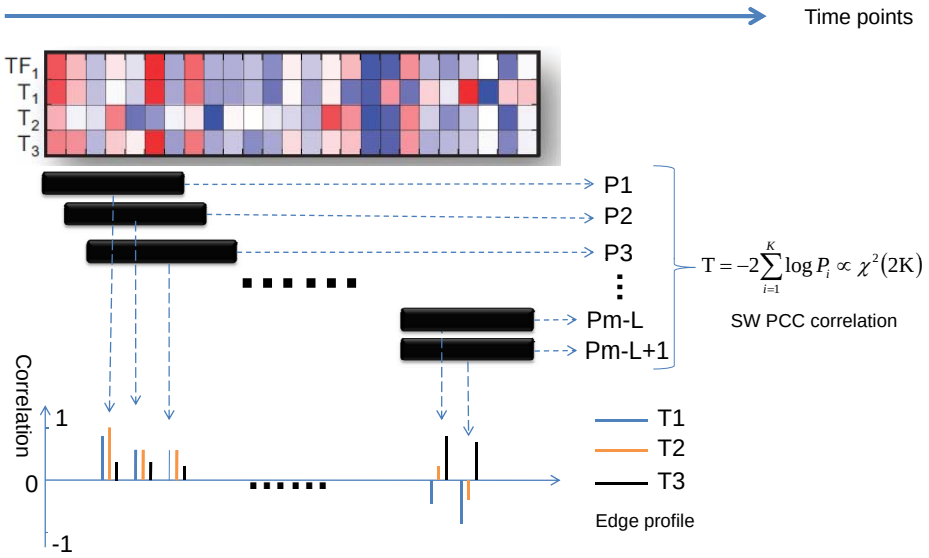


Figure 1: Illustration of the sliding window correlation strategy. Time series data is partitioned into $m-L+1$ time intervals by a sliding window. Inside each window, PCC and its p-value are calculated. The edge profile is a vector consisted of $m-L+1$ PCCs, and the overall p-value is calculated by Fisher combination test integrating the $m-L+1$ p-values.

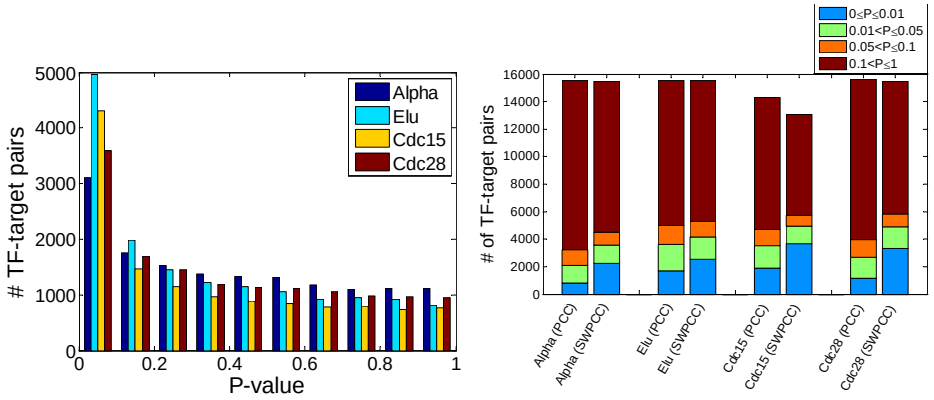
along the time-series microarray data. If there are totally m time points $1,2,\cdots,m$, then we get $m-L+1$ time intervals with $(L-1)/2,(L-1)/2+1,\cdots,m-(L+1)/2$ as centers respectively. We use the correlations between a TF and its targets at a time interval to stand for that at a time point (its center). The window length L, as a single parameter, depends on the practical requirement and data at hand. When L is small, time points within the interval are more reliable and believable because they are closer to the tested points, but the correlation will become un-reliable. Particularly, when $L=m$, the new approach degenerates to traditional methods, in this sense traditional methods are special case of our strategy. In this paper, we artificially choose $L=7$.

In each interval, we can use all traditional methods like dynamic Bayesian network methods [5], optimization based methods [6], information theory based methods [7], and so on. Here our purpose is to perform a proof-of-concept study by use of the simplest correlation methods. Firstly, we prepare all potential physical binding relationships between TF and target genes. Then Pearson correlation coefficient (PCC) is chosen to measure the correlation between time-series TF and its potential target during each time interval. Let X and Y are time-series curves of TFs and their potential targets respectively, then

$$PCC(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{2}(X_i-\bar{X})^n}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.$$

So for each TF-target pair, we get an edge profile with $m-L+1$ length .

Furthermore, we test whether the observed correlation is significantly different with



(a) PCC (b) SWPCC vs PCC

Figure 2: (**a**) Consistency between known TF-target physical binding and microarray data. Known TF-target physical bindings tend to be significantly correlated in microarray data, compared with random pairs. At the same time, most of the tested pairs (about 70%) are not correlated. (**b**) Comparison of sliding-window based Pearson correlation coefficient method (SWPCC) and direct Pearson correlation coefficient method PCC. SWPCC can detect more significantly correlated known TF-target pairs at three different cutoffs (0.01, 0.05 and 0.1).

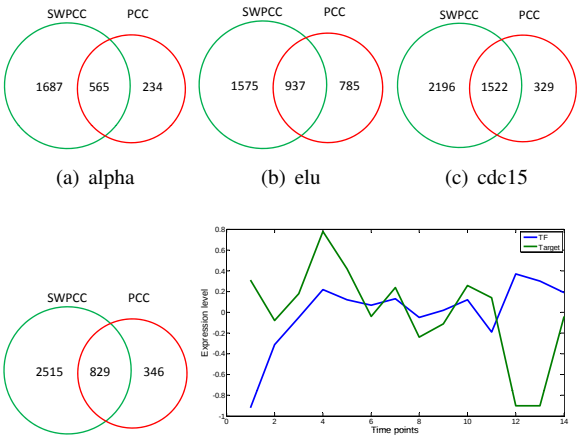
zero. The p-values for Pearson's correlation is calculated by a Student's t distribution for a transformation of the correlation. Particularly, if the underlying variables have a bivariate normal distribution and X is not correlated with Y, then

$$T = \frac{PCC}{\sqrt{\frac{1-PCC^2}{L-2}}}$$

has a student t-distribution [9]. For all these $m-L+1$ intervals, in total we get the $m-L+1$ local p-values. As shown in Figure 1, to integrate information of all time points, we use Fisher's combination test to get the overall p-value.

2.2 Identifying transient regulatory interactions on yeast data

Time-series microarray data has been widely used to infer transcriptional regulatory network, and a few studies evaluated the fact that how consistent the time-series microarray data and real regulatory pairs are. For example, Filkov et al. did a simple evaluation on the cell cycle microarray data of yeast (*Saccharomyces cerevisiae*) [10]. They collected 1,007 regulations from Yeast Protein Database (YPD) in February 2000 [11], and with a simple correlation coefficient analysis find that less than 20% of these known regulation pairs have correlations larger than 0.5. That is to say most of the known transcriptional regulation pairs do not have strong correlation signals. However, this result is not so reliable due to the incomplete of the data and lack of the statistical significance test. In this study, we use a comprehensive dataset of regulation relationship with 15,757 TF-target



(a) alpha (b) elu (c) cdc15

(d) cdc28 (e) An example

Figure 3: Venn graph for the number of significant edges identified by the two methods. Here we choose edges with p-value less than 0.01 as significant ones. (**a**)-(**d**) represent results on alpha, elu, cdc15 and cdc28 dataset respectively. (**e**) A simple example detected by SWPCC. It can not be detected by PCC because the TF switches from activator to repressor.

pairs revealed by ChIP-chip experiments[12]. Four types of cell cycle microarray data (alpha, elu, cdc15 and cdc28) are used respectively to measure how consistent the TF-target pairs are in transcription level. We then calculate the Pearson correlation coefficients additional with their p-value of all known TF-target pairs. As shown in Figure 2 (a), known TF-target pairs tend to be significantly correlated (p-value<0.1), compared with random. At the same time, most of the tested pairs (almost 70%) are not correlated, which is consistent with previous result by Filkov et al. There are several possible reasons. Firstly, mRNA level of TF is approximately used to stand for protein level; Secondly, intensive post-translational modifications may exist for TFs; Thirdly, combinational regulation may weaken the correlation between targets and their TFs.

Next, we will focus on the dynamics of edges. The starting point is that the ability and direction of TFs regulating targets may change according to different time or in different tissues. To describe and identify dynamic TF-target regulatory relationship form the yeast cell cycle process, we apply our sliding window method as described in last section. Particularly, we firstly choose $L = 7$ and calculate local p-values in each interval for all 15,757 edges. Then for each edge we get an overall p-value as an index to measure whether the edge is active. The final results compared with traditional Pearson correlation coefficient (PCC) are shown in Figure 2(b). Three different cutoffs are shown in this figure. Sliding window Pearson correlation coefficient method (SWPCC) found more number of significantly correlated known TF-target pairs at all these three cutoffs (0.01, 0.05 and 0.1). Venn graph shown in Figure 3 compares the two methods in detail at cutoff 0.01. In all four types of data, most of edges identified by PCC can be found by SWPCC, but most of edges found by SWPCC can not be identified by PCC. Figure 3 (e)

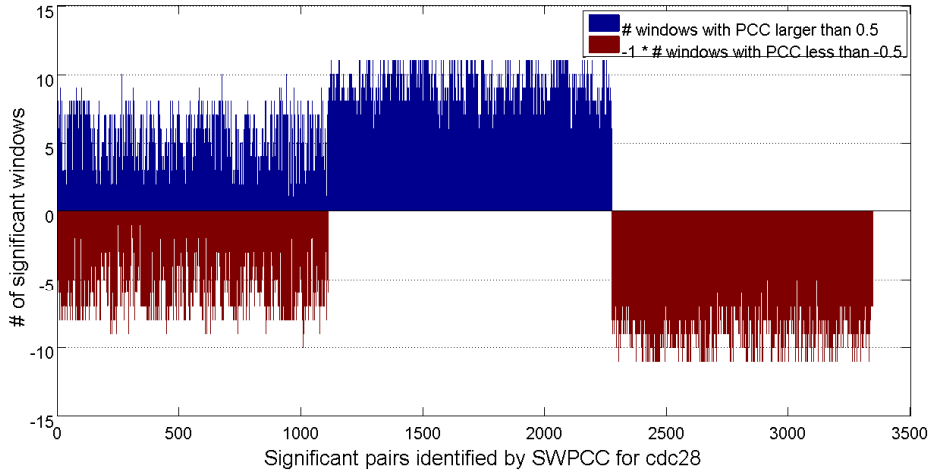


Figure 4: Three types of significant edges identified by SWPCC. Take cdc28 as an example, we partition all significant edges into three types, positive regulation, negative regulation, and a switching regulation. The horizontal axis is significant edges sorted according to three types. The blue bar stands for the number of positive regulation time intervals, while the red bar stands for the number of negative ones.

gives a simple example detected by SWPCC. It can not be detected by PCC because the relationship between TF and targets switches from positively related to negatively related.

Taking the cdc28 dataset as an example, we partition all significant edges into three types. Edges with only positive regulation are labeled as type I, edges with only negative regulation are labeled as type II, while edges containing both positive and negative ones, named as switching regulation, are labeled as type III. The horizontal axis is significant edges sorted according to their types. The blue bar stands for the number of positive regulation time intervals, while the red bar stands for the number of negative ones. Type I edges that can not be detected by PCC are interesting and may play important roles.

2.3 Revealing the dynamic of regulatory relationship between TF and target by edge profile

Based on the above calculation, we can define edge profiles. As shown in Figure 5, different profiles show different aspects of edges. In Figure 5 (a), they are switching edges. PCCs of these edges range from 0.9 to -0.9. We name them as switching edges because the relationship of TF and target may switch from positive regulation to negative regulation. In Figure 5 (b), edge profiles show the relationship between TF and target is always positively correlated or always negatively correlated. In Figure 5 (c), edges profiles show the relationship between TF and target is sometimes positively correlated sometimes non-significantly correlated. In Figure 5 (d), edges profiles show relationship between TF and target is sometimes negatively correlated sometimes non-significantly correlated.

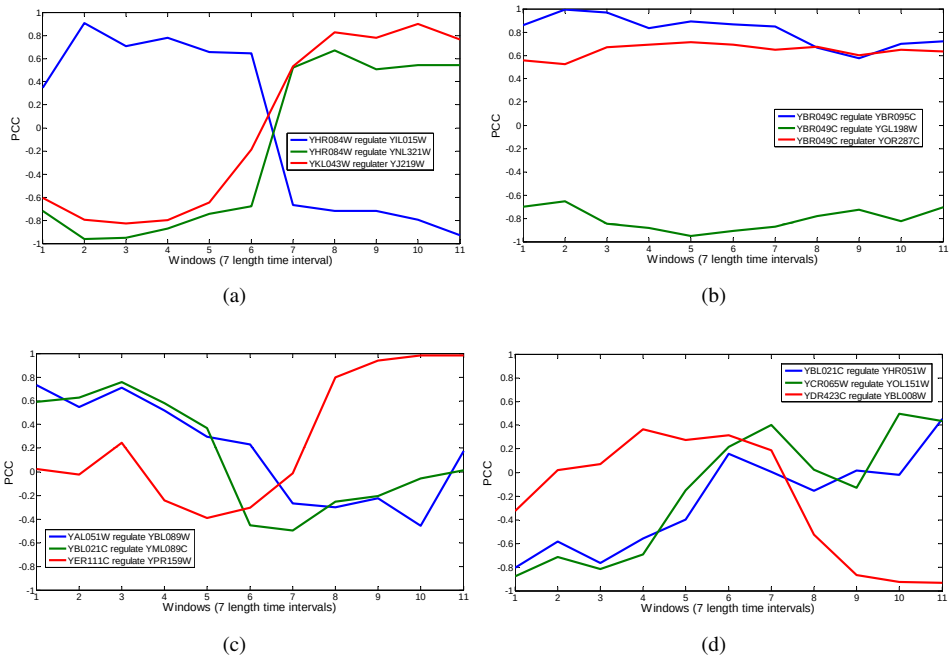


Figure 5: Different types of edge profiles. The horizontal axis stands for different time intervals. (**a**) switching edges. PCCs of these edges range from 0.9 to -0.9. We name them as switching edges because the relationship of TF and target may switch from positive regulation to negative regulation. (**b**) Edges are always positively correlated or always negatively correlated. (**c**) Edges are sometimes positively correlated sometimes non-significantly correlated. (**d**) Edges are sometimes negatively correlated sometimes non-significantly correlated.

3 Materials

3.1 Collection of TF-target Pairs

We use 15,757 high-quality TF-target regulatory relationships revealed by ChIP-chip experiments[12]. Then all self-regulations are filtered and 15,729 edges are analyzed. All gene IDs to mapped to Entrez ID using biomart [13].

3.2 Collection of Microrray Data

The cell cycle microarray data of yeast (*Saccharomyces cerevisiae*) [10] is downloaded from `http://genome-www.stanford.edu/cellcycle/`. If there are missing values, we use the mean of neighboring time points to estimate. If two neighboring time points are both missed, we mark it as NAN. When calculating PCC, genes with less than five non-NAN value are removed, and only non-NAN values are used.

4 Conclusion

We proposed a new sliding window strategy to address the dynamic nature of edges appearing in biological systems. This strategy was proved to be useful in identifying more active TF-target regulatory edges in yeast cell cycle data. Importantly, the defined edge profile describes the dynamic process of edge in a biological process. It can be useful in revealing the mechanisms of cell differentiation, cell pathological changes and other cytologic atypia.

Acknowledges

The authors thank the anonymous reviewers for their valuable suggestions to improve the article.

References

[1] Buck, M. J. and Lieb, J. D. *Genomics* **83**(3), 349–360 March (2004).

[2] Lee, T. I. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. *Science (New York, N.Y.)* **298**(5594), 799–804 October (2002).

[3] Harbison, C. T., Gordon, D. B., Lee, T. I. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. *Nature* **431**(7004), 99–104 September (2004).

[4] He, F., Buer, J., Zeng, A.-P., and Balling, R. *Genome Biology* **8**, R181+ September (2007).

[5] Kim, S. Y., Imoto, S., and Miyano, S. *Brief Bioinform* **4**(3), 228–235 January (2003).

[6] Wang, Y., Joshi, T., Zhang, X.-S., Xu, D., and Chen, L. *Bioinformatics* **22**(19), 2413–2420 October (2006).

[7] Hartemink, A. J. *Nature Biotechnology* **23**(5), 554–555 (2005).

[8] Chong, J.-L., Wenzel, P. L., Sáenz-Robles, M. T., Nair, V., Ferrey, A., Hagan, J. P., Gomez, Y. M., Sharma, N., Chen, H.-Z., Ouseph, M., Wang, S.-H., Trikha, P., Culp, B., Mezache, L., Winton, D. J., Sansom, O. J., Chen, D., Bremner, R., Cantalupo, P. G., Robinson, M. L., Pipas, J. M., and Leone, G. *Nature* **462**(7275), 930–934 (2009).

[9] Kendall, M., Stuart, A., Ord, K. J., and Arnold, S. *Kendall's Advanced Theory of Statistics:Volume 2A -Classical Inference and and the Linear Model (Kendall's Library of Statistics)*. A Hodder Arnold Publication, 6 edition, April (1999).

[10] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. *Mol Biol Cell* **9**(12), 3273–97 Dec (1998).

[11] Filkov, V., Skiena, S., and Zhi, J. *J. Comput. Biol* **9**, 317–330 (2000).

[12] Monteiro, P. T., Mendes, N. D., Teixeira, M. C., d'Orey, S., Tenreiro, S., Mira, N. P., Pais, H., Francisco, A. P., Carvalho, A. M., Lourenc co, A. B., Sá-Correia, I., Oliveira, A. L., and Freitas, A. T. *Nucleic Acids Res* **36**(Database issue), D132–6 Jan (2008).

[13] Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. *Nucleic acids research* **37**(Web Server issue), W23–27 July (2009).