

AN ITERATIVE TIME WINDOWED SIGNATURE ALGORITHM FOR TIME-DEPENDENT TRANSCRIPTION MODULE DISCOVERY

Jia Meng¹ Shou-Jiang Gao^{2,3} and Yufei Huang^{1,3}

¹Dept. of ECE, University of Texas at San Antonio San Antonio, TX 78249

²Dept. of Pediatrics, ³Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, TX 78229

jmeng@lonestar.utsa.edu, gaos@uthscsa.edu, yufei.huang@utsa.edu,

ABSTRACT

An algorithm for the discovery of time varying modules using genome-wide expression data is presented here. When applied to large-scale time serious data, our method is designed to discover not only the transcription modules but also their timing information, which is rarely annotated by the existing approaches. Rather than assuming commonly defined time constant transcription modules, a module is depicted as a set of genes that are co-regulated during a specific period of time, i.e., a time-dependent transcription module (TDTM). A rigorous mathematical definition of TDTM is provided, which serve as an objective function for the retrieving modules. Based on the definition, an effective signature algorithm is proposed that iteratively searches the transcription modules from the time series data. The proposed method was tested on the simulated systems and applied to the human time series microarray data derived from Kaposi's sarcoma-associated herpesvirus (KSHV) infection of human endothelial cells. The result has been verified by Expression Analysis Systematic Explorer.

1. INTRODUCTION

DNA microarray experiments simultaneously monitor the expression profiles of thousands of genes. Using this technology, a large amount of genome-wide expression data has been accumulated and made available.

To reveal insights into the transcriptional network from large-scale expression data, a crucial step is to classify genes and conditions into function modules, i.e., a set of genes sharing similar functions. Although it is well recognized that different transcription modules (TM) exist under different condition, most existing algorithms consider only time static transcription modules under a specific experimental condition, thus failing to capture changes of cell state. We seek in this paper to overcome this limitation.

Rather than assuming time static TM, a more realistic scenario is considered where a module is defined on for a specific period of time, i.e., a time-dependent transcription modules (TDTM). To develop an algorithm for TDTM discovery, a rigorous mathematical definition is provided for TDTM, which defines the information to be extracted from time series expression data. This definition also serves as an objective function, on which an ef-

fective iterative sliding window signature algorithm (ITWSA) is developed that iteratively refines the modules contents and time periods. In order to retrieve the time information, a sliding time window is introduced. Given a sufficiently large number of initial sets, ISWSA is possible to determine all the modules.

2. PROBLEM FORMULATION

Consider a gene expression data matrix Y obtained from time series microarray experiment. Define

Y_{gt} : Expression level of gene g at time t

$$Y_{g(t_i:t_j)} = [Y_{gt_i}, Y_{g(t_i+1)}, \dots, Y_{gt_j}]$$

$$Y_{gT} = [Y_{gt_0}, \dots, Y_{gt_r}]$$

where, $g \in [1, 2, \dots, G], t \in [0, 1, \dots, T]$

Given a pair thresholds τ_C, τ_G , a window width

$W = 2l + 1$, a TDTM is defined by a set of genes G_m and a set of time period $[T_m, L_m]$:

$$M(\tau_r, \tau_G) := \left\{ \begin{array}{l} (G_m, [T_m, L_m]) \forall [t_m^{(i)}, l_m^{(i)}] \in [T_m, L_m]: \\ \frac{1}{|G_m|} \sum_{g \in G_m} \rho \left(Y_{g(t_m^{(i)}, t_m^{(i)}+l_m^{(i)}-1)}, \langle Y_{G_m(t_m^{(i)}, t_m^{(i)}+l_m^{(i)}-1)} \rangle \right) > \tau_C \\ \forall g \in G_m: \\ \frac{1}{\sum_{i=1}^{l_m^{(i)}}} \sum_{[t_m^{(i)}, l_m^{(i)}] \in [T_m, L_m]} \left[t_m^{(i)} \rho \left(Y_{g(t_m^{(i)}, t_m^{(i)}+l_m^{(i)}-1)}, \langle Y_{G_m(t_m^{(i)}, t_m^{(i)}+l_m^{(i)}-1)} \rangle \right) \right] > \tau_G \end{array} \right\}$$

Where, ρ is the Pearson correlation, $|X|$ is the number of component of X , and $\langle Y_{G_m t_i} \rangle = \frac{1}{|G_m|} \sum_{g \in G_m} Y_{gt_i}$. With

this definition, the objective is to design an algorithm that can determine the gene set G_m and time period $[T_m, L_m]$ that satisfy the definition.

3. THE ITERATIVE TIME WINDOWED SIGNATURE ALGORITHM

We describe in this section the detailed ITWSA. From an initial input set, ITWSA first selects time periods during which the selected genes are co-regulated; then it identifies genes that are co-regulated during the selected time periods. Iterating between these two steps will revise the output and finally reach a stable state (genes and time information no longer change). Presumably, when using a sufficient number of initial gene

set, it is possible to retrieve all the TMs in the data. Since the number of possible initial sets increases exponentially with the number of genes, efficient initial sets should be selected.

The algorithm for determining a TDTM can be summarized as follows:

Step 1: A first gene is randomly selected and 29 genes that have the largest Pearson correlation with the first gene were added to form the initial gene set;
Step 2: A set of consecutive time points of width $(2l+1)$ is chosen to measure the convergence of a suggested TM at a particular time; Keep all the t_i that satisfies:

$$S_i^{G_m} = \frac{1}{|G_m|} \cdot \sum_{g \in G_m} \rho \left(Y_{g(t_i-l:t_i+l)}, \left\langle Y_{G_m(t_i-l:t_i+l)} \right\rangle \right) > \tau_T$$

Step 3: All the genes that are co-regulated by this suggested TDTM are identified by calculating its Pearson distance with the center of the TDTM at the selected time period:

$$S_{[T_m, L_m]}^g = \frac{1}{|L_m|} \cdot \sum_{t \in [T_m, L_m]} \left[\rho \left(Y_{g(t:t_m)}, \left\langle Y_{G_m(t:t_m)} \right\rangle \right) \right] > \tau_G$$

where, $[T_m, L_m]$ is calculated from the previous step.

Iterate between steps 2 and 3 until convergence, i.e., the gene set G_m and time period sets $[T_m, L_m]$ no longer change during iterations.

To find another module, the algorithm restarts but in step 1 a gene is randomly selected from the remaining genes excluding genes in the previous initial set.

4. TEST ON SIMULATED SYSTEMS

ITWSA is first validated on simulated systems. Simulated data, transcription modules as shown in Figure 1 were first generated, which is represented as matrix X_g .

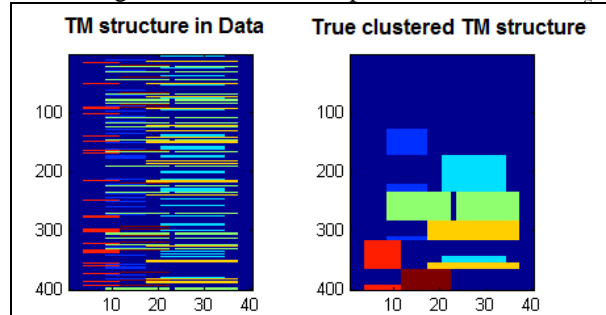


Figure 1. Simulated TMs and true clustered TMs.

The left figure shows the transcription module structure in simulated data (Horizontal label is time; vertical label represents genes), each color represents a separate TM; the right figure shows the same TMs but with genes that belong to the same TM put together. There are 6 TMs, and each TM containing 40 to 60 randomly selected genes and occupying a single (e.g., the red module in Fig. 1.) or two separate time periods (e.g., the green module in Fig. 1). It is possible that a gene is not involved in any TM or is involved in more than one TM during different time points. Data are generated by scaling and adding noise to the left figure.

Then a dynamic state space model is used to generate simulated data:

$$\begin{cases} X_{g^{t+1}} = X_{g^t} + n \\ Y_{g^t} = aX_{g^t} + u_{g^t} \end{cases}$$

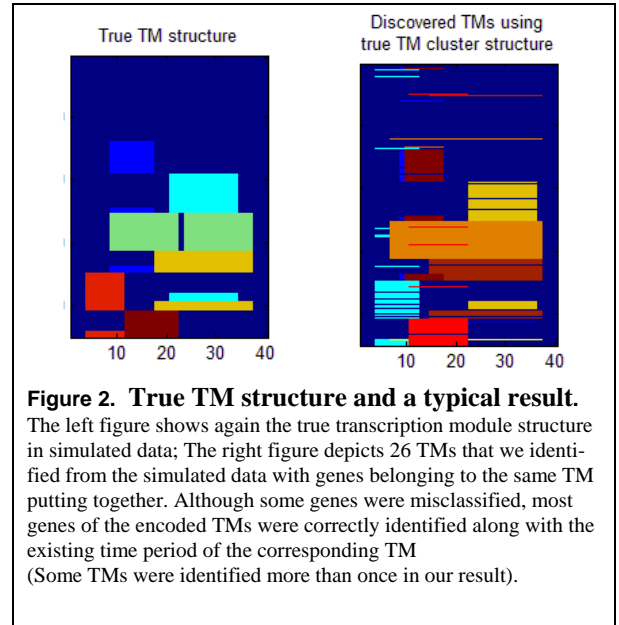
where,

$$n \sim N(0,1), a_g \sim Unif(0,1), u_{g^t} \sim N(u_g, 0.01), u_g \sim Unif(0,1)$$

For genes belonging to the same TM, we use the same $X_{G_m t}$ during the time periods that the correspondent TM exists, i.e.

$$X_{g^t} = X_{G_m t}, \forall g \in G_m, t \in [T_m, L_m]$$

An example of the discovered module is shown in Figure 2.



Our simulation results show that when choosing the thresholds and window size properly, the ITWSA is able to correctly identify not only most genes in the encoded transcription modules but also the time period that TMs exists.

5. TEST ON REAL DATA

We applied the ITWSA algorithm to analyze the human time series microarray data derived from KSHV infection of human endothelial cells. The data were produced with Affymetrix Human Genome U133A Chips, consisting of the expression profiles of 14,500 genes at time $t=[0,1,3,6,10,16,24,36,54,78]$ (hour) after infection. Since priority was given to earlier states, sample times were unevenly chosen.

5.1. Data Preprocessing

An intensity filter (the intensity of a gene should be above 100 in at least 25 percent of the samples), and a variance filter (the inter-quartile range of \log_2 -intensities should be at least 0.5) were first applied to se-

lect 2210 differentially expressed genes. They were further normalized so that all genes contribute equally in the algorithm.

5.2. Results

When $\tau_T = 0.6$, $\tau_G = 0.7$ and window size equals to 5, 145 TMs were identified, and 1230 of 2210 genes were associated with at least one TM. See Figure 3 for detailed discussion of the results.

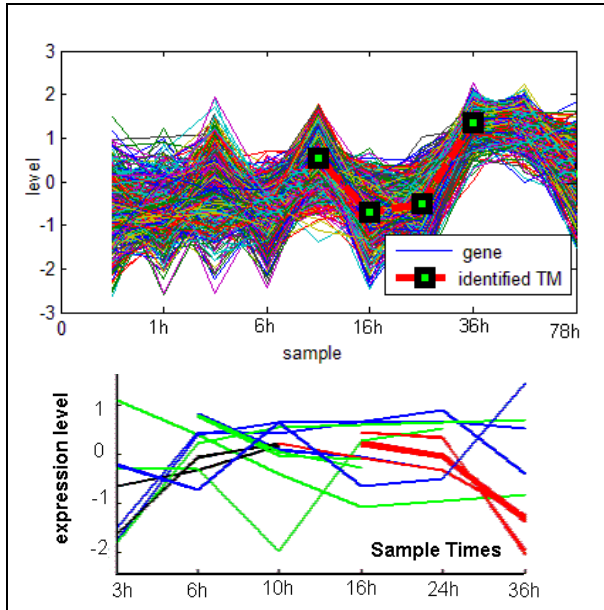


Figure 3. A typical TDTM and centroid plot of all 145 TDTMs identified from data

The upper figure provides an example of a TDTM identified from data. It depicts 1) the expression trajectory of all the genes belong to this TDTM but over the entire sample period, 2) the centroid of TDTM during the identified time period. The TDTM is identified to exist between 10h-36h as described by the centroid plot. It can be seen that, these genes behave only alike during the identified time period but they do not show similarity in other period.

The lower figure shows the centroid plot of all the 145 TDTMs identified. Due to the size of sliding window, we cannot elucidate whether a TDTM exists at time $t=[0h, 1h, 54h, 78h]$. As can be seen:

1. Many TMs identified are similar. (In fact many of them are even identical).
2. A single TM may exist in earlier time period, later period or exist all through the time.
3. Generally, TMs that exist only in earlier time period tend to go up (TDTMs illustrated by black lines in the figure); while TMs exist only in later period tend to go down (TDTMs illustrated by red lines in the figure).

5.3. GO analysis

To further validate the results, EASE (Expression Analysis Systematic Explorer) is used to verify whether an identified TDTM enriches meaningful gene categories. The result is supportive. The EASE results of the TDTM example shown in Figure 3 is tabulated in table 1:

Gene Category	List Hits	List Total	Population Hits	Population Total	EASE score
defense response	26	162	765	11051	1.11E-04
response to external stimulus	36	162	1278	11051	1.51E-04
immune response	24	162	690	11051	1.61E-04
integral to plasma membrane	35	160	1254	10895	2.37E-04
response to biotic stimulus	26	162	830	11051	3.90E-04
integral to membrane	61	160	2795	10895	4.83E-04
membrane	83	160	4191	10895	4.90E-04
receptor activity	34	165	1268	11182	7.02E-04
plasma membrane	43	160	1786	10895	8.31E-04

Table 1. Enriched GO Terms of the TDTM shown in Figure 3.

The table shows the GO analysis result of the specific TDTM shown in Figure 3 by using EASE.

The numbers of genes in the table are all genes assayed or listed (i.e. all genes in the microarray (HGU133A) or TDTM) and annotated within a given system of classifying genes (e.g. the 'Molecular Function' branch of the Gene Ontology). Therefore the population can change from one system to the next. "Hits" refers to genes falling within the gene category in question. Population Hits refer to number of genes in the total group of genes assayed that belong to the specific Gene Category. The EASE score (The upper bound of the distribution of Jackknife Fisher exact probabilities given the List Hits, List Total, Population Hits and Population Total) is essentially a sliding-scale, conservative adjustment of the Fisher exact that strongly penalizes the significance of categories supported by few genes and negligibly penalizes categories supported by many genes. The smaller EASE score is, the gene category is enriched more significantly.

The listed GO categories are enriched significantly by the genes belonging to this TDTM, and these GO categories (defense response, response to external stimulus, etc) are relevant to KSHV infection.

6. DISCUSSION AND FUTURE WORK

In the algorithm, 3 parameters need to be chosen: two thresholds τ_C, τ_G and the width of window $(2l+1)$, which is used when we calculate the Pearson correlation in the step of deciding TM's existing time.

Two thresholds serve as the detection threshold of the search. Presumably, the larger they are, the more TMs can be discovered and more genes they will contain. The width of the window determines the sensitivity of the search towards time. When a smaller window is used, the location time boundary can be more precisely identified; however FDR will also increase.

Future work will seek to increase the stability and precision simultaneously. Post processing of result will also be emphasized.

7. ACKNOWLEDGMENTS

Yufei Huang is supported by an NSF Grant CCF-0546345. Shou-Jiang Gao is supported by NIH grants CA096512 and CA124332.

8. REFERENCES

- [1] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv and N. Barkai, "Revealing modular organization in the yeast transcriptional network", *Nature genetics*, vol. 31, Aug, 2002.

- [2] S. Bergmann, J. Ihmels and N. BarkaiPettinen, "Iterative signature algorithm for the analysis of large-scale gene expression data" *Physical review*, E 67, 031902 (2003).
- [3] J. Supper, M. Strauch, D. Wanke, K. Harter and A. Zell, "EDISA: extracting biclusters from multiple time-series of gene expression profiles", *BMC Bioinformatics* 2007, 8:334.
- [4] DAVID: download EASE
[<http://david.niaid.nih.gov/david/ease.htm>]
- [5] D Hosack, G Dennis, B Sherman, H Lane and R LempickI "Identifying biological themes within lists of genes with EASE" *Genome Biology* 2003, 4:R70