| | |
|---|---|
| Title | Synthetic time series resembling human (HeLa) cell-cycle gene expression data and application to gene regulatory network discovery |
| Author(s) | Tam, GHF; Hung, YS; Chang, C |
| Citation | The 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC 2013), Hangzhou, Zhejiang, China, 26-27 August 2013. In Conference Proceedings, 2013, v. 2, p. 538-541 |
| Issued Date | 2013 |
| URL | http://hdl.handle.net/10722/186742 |
| Rights | International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) Proceedings. Copyright © IEEE Computer Society. |

2013 Fifth International Conference on Intelligent Human-Machine Systems and Cybernetics

# Synthetic Time Series Resembling Human (HeLa) Cell-Cycle Gene Expression Data and Application to Gene Regulatory Network Discovery

Gary Hak Fui Tam, and Yeung Sam Hung
Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong, China
hftam@eee.hku.hk, and yshung@eee.hku.hk

Chunqi Chang
School of Electronic and Information Engineering
Soochow University
Suzhou, Jiangsu Province, China
cqchang@suda.edu.cn

***Abstract*—Evaluation of gene regulatory network (GRN) discovery methods relies heavily on synthetic time series. However, synthetic data generated by traditional method deviate a lot from real data, making such evaluation questionable. Guiding by decaying sinusoids, we propose a new method that generates synthetic data resembling human (HeLa) cell-cycle gene expression data. Using the new synthetic data, a simple comparison between four GRN discovery methods reveals that Granger causality (GC) methods substantially outperform Pearson correlation coefficient (PCC), while time-shifted PCC can give comparable performance as GC methods. The new synthetic data generation would also be useful for generating other kinds of cell-cycle time series. Using data generated by our proposed method, evaluation of GRN discovery methods should be more trustworthy for real-data applications.**

***Keywords-gene regulatory network; synthetic data; cell-cycle; time series; vector autoregressive model; Pearson correlation coefficient; Granger causality.***

## I. INTRODUCTION

In gene regulatory network (GRN) discovery, gene-gene interactions are inferred from gene expression data [1]–[7]. Identified genes involved in disease development can become future drug targets, hence promoting medical advances. Since gene regulation is a temporal process, data from time-series experiment are more informative [1]. A number of recent studies have focused on inferring GRN from time-series data, e.g. [3]–[7]. To evaluate the performance of their proposed methods, usually both synthetic data and real biological data are used. However, the ground truth (i.e. the true underlying biological network of interactions) of real data is usually unknown or unreliable [3], [6], making it difficult (or not reliable) to draw conclusion from results on real data. Thus, evaluation of performance relies heavily on synthetic data. Yet, most synthetic data people used do not resemble real data, making this part of evaluation also questionable.

In this paper, we propose a method that generates synthetic data resembling the real data more closely, so that evaluation using these data should be more reliable. First, we shall briefly describe a dataset widely used in GRN discovery. Secondly, we shall give a typical (traditional) method for generating synthetic data before we describe our new method. The new method will then be compared with the traditional method. Finally, we shall apply our new synthetic data to carry out a simple comparison on four GRN discovery methods.

## II. THE HUMAN (HELA) CELL-CYCLE GENE EXPRESSION DATASET

HeLa dataset [8] is one of the most commonly adopted dataset for GRN discovery. The dataset contains 5 time-series experiments carried out with cDNA microarrays, and about one thousand genes are identified to be periodically expressed with period ≈ 16 hours. Since we intend to compare our new method with the traditional (vector-autoregressive based) method given in 2 papers [3] and [4], and these 2 papers analyzed 9 selected periodic genes using experiment 3, we shall focus on these 9 genes, too. In experiment 3, gene expression levels were measured at time 0, 1, 2, …, 46 hours. So, the number of time points is 47. The expression profiles of these 9 genes are plotted in Fig. 1.

## III. SYNTHETIC DATA GENERATION

### A. Vector Autoregressive Model

Time-series data can be easily generated by vector autoregressive (VAR) model. Suppose we would like to generate data for $n$ genes and number of time points (data length) is $T$. Let an $n\times1$ vector $x_t$ denote the gene expressions at time $t$, the VAR model of order $p$ can be expressed as

$$x_t = \sum_{l=1}^{p} A_l x_{t-l} + e_t , \tag{1}$$

where $A_l$ is an $n\times n$ coefficient matrix containing parameters of the VAR model, and $e_t$ is an $n\times1$ vector representing independent Gaussian white noise of zero mean.

Suppose the GRN has $L$ edges, they can be represented by off-diagonal elements of the coefficient matrices $A_l$ ($l$=1,…,$p$). As long as $A_l$ are specified and initial values of $x_t$ are given, time series of $n$ genes can be obtained by repeatedly applying (1). In this paper, initial values of $x_t$ are taken to have the same nature as $e_t$. After iteration with (1), the first 100 time points (regarded as transient) are dropped and the subsequent $T$ time points are taken as generated data. In the following, model order $p$=2 is adopted.

978-0-7695-5011-4/13 $26.00 © 2013 IEEE
DOI 10.1109/IHMSC.2013.276

### B. Traditional Method

Previous works [3] and [4] specified coefficient matrices $A_l$ by random sampling from Gaussian distribution only. Here we follow the parameters used in [3]. All diagonal elements of $A_l$ are randomly sampled from Gaussian distribution of mean 0 and standard deviation (SD) 0.25. For the GRN edges, $L$ off-diagonal positions are randomly selected from $A_l$ and these positions are filled by values taken from Gaussian distribution of mean 0 and SD 0.25. Other elements of $A_l$ are zeros. (In our investigation, iteration with (1) is highly unstable and time series blow up exponentially if SD>0.35 is used.)

SD for $e_t$ and $x_t$ is 0.1. In fact, this SD is not important because it merely scales the generated data.

### C. Our New Method

To generate synthetic data with a form similar to HeLa data, we consider a decaying sinusoidal function $f(t) = be^{-at}\sin\omega t$, which has a $z$-transform satisfying the following discrete-time relation:

$$\begin{aligned} f_k &= \left(2e^{-a\Delta t}\cos\omega\Delta t\right)f_{k-1} - e^{-2a\Delta t}f_{k-2} + \left(be^{-a\Delta t}\sin\omega\Delta t\right)e_{k-1} \\ &= a_1 f_{k-1} + a_2 f_{k-2} + b_1 e_{k-1} \end{aligned} \quad (2)$$

where

$$\begin{aligned} a_1 &= 2e^{-a\Delta t}\cos\omega\Delta t, \\ a_2 &= -e^{-2a\Delta t}, \\ b_1 &= be^{-a\Delta t}\sin\omega\Delta t \end{aligned}$$

and $e_{k-1}$ is discrete input to the system $f(t)$. To make $f_k$ look similar to HeLa data which have period ≈ 16 hours, we take $\Delta t$=1 and $\omega = 2\pi/16 \approx 0.3927$. The coefficients $a_1$ and $a_2$ are the diagonal elements of coefficient matrices $A_1$ and $A_2$. The $n$ genes may have different parameters $a$ which are randomly sampled from a uniform distribution over the interval [0.6, 1]. These diagonal elements give some degrees of periodic pattern to the synthetic series. To construct $L$ edges for the $n$-gene network, off-diagonal elements are generated as follows. Off-diagonal elements of $A_1$ are obtained by random sampling from a uniform distribution over [0.2, 0.8] with a positive/negative sign also added probabilistically, such that their magnitudes are comparable to the diagonal elements. Off-diagonal elements of $A_2$ are similarly generated but have a smaller magnitude within [0.15, 0.6]. If two off-diagonal elements of $A_1$ and $A_2$ occur at the same position, they only correspond to one edge. Thus, $A_1$ and $A_2$ should be superimposed when counting the number of edges formed. We stop generating off-diagonal elements when exactly $L$ edges are constructed. All remaining off-diagonal elements are zero. The above parameter settings ensure that elements in $A_2$ generally have smaller magnitudes than $A_1$, enabling the iteration with (1) to be more stable and time series are less likely to blow up exponentially.

Initial values of $x_t$ are randomly taken from a Gaussian distribution of mean 0 and SD 0.5 (again this SD is just a scaling), and so is the white noise input $e_t$.

## IV. Comparison of Synthetic Data

To concord with the real data shown in Fig. 1, $n$=9, $T$=47 and $L$=11 are adopted to generate synthetic data with the traditional method and our new method. The plots of these synthetic time series are shown in Fig. 2. It is clear that synthetic data generated by traditional method look like noises and deviate a lot from the real data, whereas synthetic data generated by our proposed method exhibit some form of periodic pattern resembling the real data.

Besides the plots, the following 2 measures are also used for comparison:

### A. Model Consistency

Given time-series data of $n$ genes and length $T$, the underlying GRN can be discovered by conditional Granger causality (CGC) [7], where the input data are regressed by a VAR model. Model consistency measures if the correlation structure of the data is captured by the VAR model properly [9]. Consistency is computed as:

$$C = \left(1 - \frac{\left|R_p - R_i\right|}{\left|R_i\right|}\right) \times 100\% \quad (3)$$

where $R_i$ and $R_p$ are reshaped row vectors (of length $n^2$) from covariance matrices of input time series and predicted time series by the VAR model, respectively. If $C$ is low, the discovered network is unreliable. Here, with the same regression method in CGC [7], a low $C$ means that the standard CGC is not able to reconstruct a reliable network from the time-series data, revealing that the data quality is low.

### B. Statistical Power

An $n$-gene network has a total of $M = n(n-1)$ possible directed edges. Application of CGC returns a $p$-value for each of these $M$ possible edges. The $L$ edges with lowest $p$-values can be taken to constitute the discovered network. If these $p$-values are small, then the time-series data offer higher statistical power, meaning that the data quality is high. Besides looking at the minimum and maximum of these $L$ $p$-values, we also compute the geometric mean of all the $L$ $p$-values – denoted by $p_{\text{gm}}$.

Table I shows the results of the above 2 measures for the real data and the synthetic data generated using both the traditional and our proposed methods. Obviously, the traditional method gives a $C$ much lower than that of real data; whereas our method gives a comparable $C$ as real data. Regarding the statistical power, the traditional method yields $p$-values one order of magnitude higher than that of real data, revealing a much lower statistical power. The maximum $p$-value of 0.16 is not statistically significant. On the other hand, synthetic data generated by our proposed method yield

even higher statistical power than that of real data. If this statistical power is higher than desired for specific applications, it can be easily decreased by adding noise.

TABLE I. RESULTS OF 2 MEASURES REVEALING DATA QUALITY

| | | Real Data | Synthetic Data | |
|---|---|---|---|---|
| | | | ***Traditional*** | ***Ours*** |
| ***C*** | | **93 %** | **43 %** | **90 %** |
| Lowest $L$ $p$-values | Min | $4.5\times10^{-4}$ | $3.5\times10^{-3}$ | $1.0\times10^{-6}$ |
| | Max | $5.3\times10^{-2}$ | $1.6\times10^{-1}$ | $5.9\times10^{-2}$ |
| | $p_{gm}$ | $\mathbf{6.3\times10^{-3}}$ | $\mathbf{5.3\times10^{-2}}$ | $\mathbf{1.2\times10^{-3}}$ |

## V. APPLICATION OF OUR SYNTHETIC DATA

This section gives a simple example for applying our new synthetic data generation to compare GRN discovery methods. The synthetic data series in Fig. 2(b) come from an underlying network as shown in Fig. 3(a), which corresponds to the non-zero off-diagonal elements of $A_1$. In the following, we reconstruct GRNs by four methods using the data series in Fig. 2(b). Fig. 3(a) is used as the ground truth network for computing precisions [7] of reconstructed (or discovered) networks.

Two methods belong to the class of correlation score, where Pearson correlation coefficient (PCC) [10] is used. Since the simple PCC can only detect an interaction between two genes but direction of influence cannot be determined, we also implement a time-shifted PCC (TSPCC) as follows. For each pair of genes $i$ and $j$, gene $j$ series is shifted by –2, –1, 0, 1 or 2 time points, and its PCC with gene $i$ series is computed for each time shift. The maximum PCC and the corresponding time shift are taken. A positive time shift infers that $j$ influences $i$, and vice versa. In this way, a directed network can also be obtained by correlation method. The maximum time shift is 2 because the synthetic data are generated by a second order ($p$=2) VAR model. Fig. 3 (b) and (c) show the reconstructed networks by simple PCC and TSPCC, where each reconstruction takes the $L$ edges with largest PCC magnitude.

The other two methods are pairwise Granger causality (PGC) and CGC [7]. Model validation as described in [7] is used with PGC. For both PGC and CGC, the $L$ edges with lowest $p$-values are taken to be the reconstructed network. Fig. 3 (d) and (e) show the results.

Fig. 3 clearly shows that simple PCC is not effective in reconstructing GRN, whereas Granger causality (GC) methods work much better. (The outperformance of CGC over PGC was explained in [7].) Yet, it is a bit surprising that a straight-forward time-shifted version of PCC also gives comparable performance as GC methods. TSPCC is also good in the sense that if a correct interaction is identified between two genes, the direction of influence is obtained correctly, while PGC is not able to do so. (cf. CGC can also achieve this.) Comprehensive comparisons will be left to future works.

## VI. CONCLUSION

Synthetic data in GRN discovery are commonly generated by VAR model, of which the coefficient matrices are crucial. Traditional method specifies the coefficient matrices by random sampling and synthetic data resulted deviate a lot from real data, so evaluation of GRN discovery methods using these synthetic data is questionable. In this paper, we have proposed a new method where the choice of coefficient matrices is guided by decaying sinusoids. Synthetic data so generated resemble the human (HeLa) cell-cycle gene expression data fairly well. Evaluation of GRN discovery methods using these data should be more reliable and valuable. Other cell-cycle time series may be synthesized by adjusting the parameters in our proposed method.

A simple comparison of four GRN discovery methods is carried out using our synthetic data. Results show that GC methods outperform simple PCC. However, a time-shifted version of PCC gives comparable performance to GC methods.

## ACKNOWLEDGMENT

We would like to acknowledge funding support by Hong Kong SAR Research Grants Council (Project No HKU762111M) and CRCG of the University of Hong Kong.

## REFERENCES

[1] Z. Bar-Joseph, "Analyzing time series gene expression data," Bioinformatics, vol. 20, no. 16, pp. 2493–2503, November 2004.

[2] L. Chen, R. S. Wang, and X. S. Zhang, Biomolecular Networks: Methods and Applications in Systems Biology. NJ: Wiley, 2009, p. 47.

[3] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical Granger modeling for gene expression regulatory networks discovery," Bioinformatics, vol. 25, no. 12, pp. i110–118, June 2009.

[4] A. Shojaie, and G. Michailidis, "Discovering graphical Granger causality using the truncating lasso penalty," Bioinformatics, vol. 26, no. 18, pp. i517–i523, September 2010.

[5] X. Zhang, et al., "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," Bioinformatics, vol. 28, no. 1, pp. 98–104, January 2012.

[6] G. H. F. Tam, A Granger Causality Approach to Gene Regulatory Network Reconstruction based on Data from Multiple Experiments. PhD thesis, The University of Hong Kong, 2012.

[7] G. H. F. Tam, C. Chang, and Y. S. Hung, "Application of Granger causality to gene regulatory network discovery," Proc. IEEE 6th International Conference on Systems Biology, Xi'an, China, pp. 232–239, August 2012.

[8] M. L. Whitfield, et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," Mol. Biol. Cell, vol. 13, pp. 1977–2000, June 2002.

[9] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," J. Neurosci. Methods, vol. 186, pp. 262–273, February 2010.

[10] D. Steinhauser, L. Krall, C. Mussig, D. Bussis, and B. Usadel, "Correlation Networks," in Analysis of Biological Networks, B. H. Junker, and F. Schreiber, Eds. NJ: Wiley, 2008, pp. 305–333.

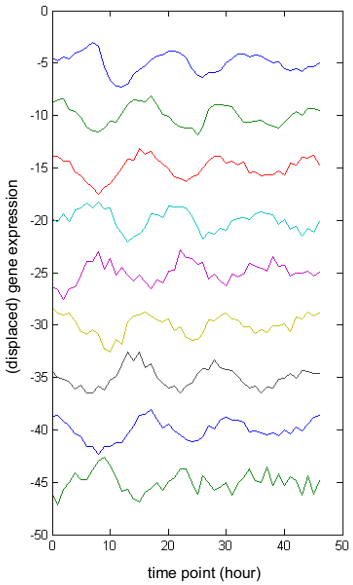


Figure 1. Time series of the 9 genes in HeLa dataset. For visualization, each series is normalized to zero mean and unit variance, and then each displaced by 5 units. Gene names from the top to bottom are CDC2, CDC6, E2F1, CCNA2, CDKN3, RFC4, CCNE1, PCNA, CCNB1.

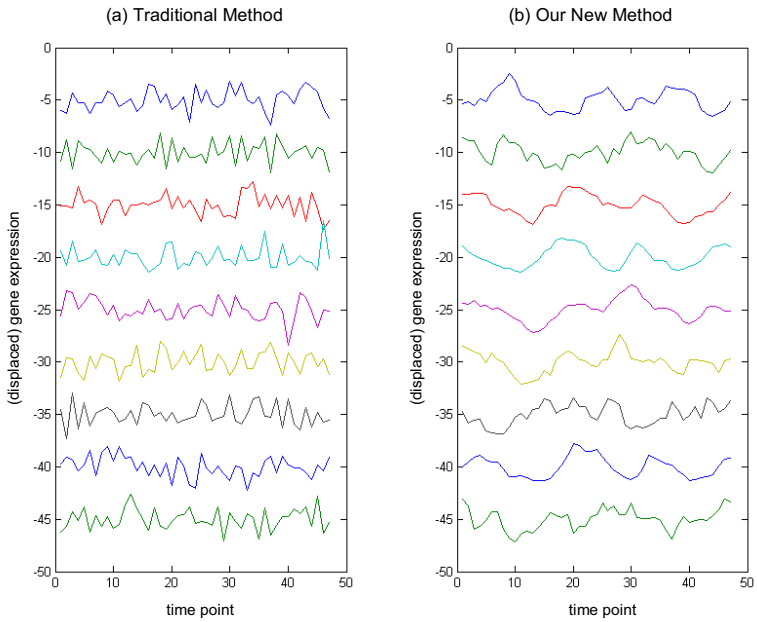


Figure 2. Time series of synthetic data generated by (a) the traditional method, and (b) our new method. For visualization, each series is normalized to zero mean and unit variance, and then each displaced by 5 units.

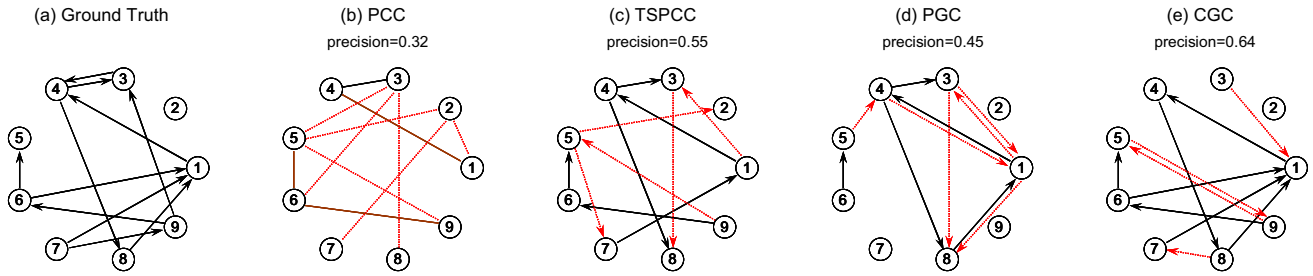


Figure 3. GRNs. (a) is the ground truth network corresponding to our synthetic data shown in Fig. 2(b). Other GRNs are reconstructed from these data by four methods. Solid black arrows represent true positives, dotted red arrows represent false positives. For (b) PCC, edges are undirected and solid lines represent correct edges.