Application of Granger Causality to Gene Regulatory Network Discovery

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

***Abstract*—Granger causality (GC) has been applied to gene regulatory network discovery using DNA microarray time-series data. Since the number of genes is much larger than the data length, a full model cannot be applied in a straightforward manner, hence GC is often applied to genes pairwisely. In this paper, firstly we investigate with synthetic data and point out how spurious causalities (false discoveries) may emerge in pairwise GC detection. In addition, spurious causalities may also arise if the order of the vector autoregressive model is not high enough. Therefore, besides using a suitable model order, we recommend a full model over pairwise GC. This is possible if pairwise GC is first used to identify a network of interactions among only a few genes, and then all these interactions are validated with a full model again. If a full model is not possible, we recommend using model validation techniques to remove spurious discoveries. Secondly, we apply pairwise GC with model validation to a real dataset (HeLa). To estimate the model order, the Akaike information criterion is found to be more suitable than the Bayesian information criterion. Degree distribution and network hubs are obtained and compared with previous publications. The hubs tend to act as sources of interactions rather than receivers of interactions.**

Keywords-Granger causality, gene regulatory network, DNA microarray, pairwise, spurious discovery, model validation.

I. INTRODUCTION

Gene regulatory network (GRN) discovery is a hot research topic. It identifies gene-gene interactions from mRNA experiment data to help elucidate biological process in disease development, hence promoting medical advances [1]–[4]. Recently, Granger causality (GC) has been applied to GRN discovery using DNA microarray time-series data, e.g. [5], [6]. However, due to the curse of dimensionality, i.e. the number of genes is much larger than the data length, a full model cannot be applied to all genes simultaneously, and thus GC is often applied to genes pairwisely. For example, [5] and [6] used a low model order of one only, which we doubt is not sufficient. In this paper, first we investigate the effects of applying GC pairwisely and the choice of model order. We would also like to see if model validation techniques provided by Granger causal connectivity analysis (GCCA) toolbox [7] can help in the above situations. Since ground truth network is usually unknown for real data, we adopt synthetic data for the purpose of conducting a more reliable evaluation. Secondly, we apply pairwise GC with model validation to a real dataset and compare results with previous publications.

II. GRANGER CAUSALITY

Granger causality (GC) [8] is well suited for identifying causal relations among multiple time series, hence we adopt it for our GRN discovery.

A. Definition of Granger Causality

Suppose there are two time series X and Y, if Y can help predicting the future of X, then Y "Granger-causes" X. This means that, including past observations of Y can reduce the prediction error of X, compared to the prediction made using past observations of X only.

B. Bivariate and Multivariate Autoregressive Models

Vector autoregressive (VAR) model is often used to detect GC.

1. Bivariate autoregressive model

First, consider a system with two variables x_1 and x_2, with measurements constituting two time series both of data length T. Denoting their measurements at time t by $x_{1,t}$ and $x_{2,t}$ (t=1,2,…,T), respectively, the system can be modeled by a bivariate autoregressive model:

$$\begin{aligned} x_{1,t} &= \sum_{l=1}^{p}\left(A_{11,l}x_{1,t-l} + A_{12,l}x_{2,t-l}\right) + e_{1,t} \\ x_{2,t} &= \sum_{l=1}^{p}\left(A_{21,l}x_{1,t-l} + A_{22,l}x_{2,t-l}\right) + e_{2,t} \end{aligned} \tag{1}$$

where p is the model order, which is the number of time lag to be included in the model. Coefficients of the model can be collectively denoted as

$$A_l = \begin{bmatrix} A_{11,l} & A_{12,l} \\ A_{21,l} & A_{22,l} \end{bmatrix}.$$

$e_{1,t}$ and $e_{2,t}$ represent residuals (prediction errors) of the two time series.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Since $t = p+1,\cdots,T$ for (1), the bivariate model (1) represents $m = T-p$ pairs of equations. Using matrix notation and taking transpose, these equations can be stacked and written in the form of standard linear regression:

$$Y = XB + E \tag{2}$$

where

$$\underset{(m\times 2)}{Y} = \begin{bmatrix} x_{1,p+1} & x_{2,p+1} \\ x_{1,p+2} & x_{2,p+2} \\ \vdots & \vdots \\ x_{1,T} & x_{2,T} \end{bmatrix},$$

$$\underset{(m\times 2p)}{X} = \begin{bmatrix} x_{1,p} & x_{2,p} & x_{1,p-1} & x_{2,p-1} & \cdots & x_{1,1} & x_{2,1} \\ x_{1,p+1} & x_{2,p+1} & x_{1,p} & x_{2,p} & \cdots & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1,T-1} & x_{2,T-1} & x_{1,T-2} & x_{2,T-2} & \cdots & x_{1,T-p} & x_{2,T-p} \end{bmatrix},$$

$$\underset{(2p\times 2)}{B} = \begin{bmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{bmatrix}, \quad \underset{(m\times 2)}{E} = \begin{bmatrix} e_{1,p+1} & e_{2,p+1} \\ e_{1,p+2} & e_{2,p+2} \\ \vdots & \vdots \\ e_{1,T} & e_{2,T} \end{bmatrix}.$$

The coefficient matrix B can be estimated by ordinary least squares (OLS):

$$\hat{B} = (X'X)^{-1}X'Y, \tag{3}$$

provided that that $X'X$ is invertible. Then, the prediction error of x_1 can be measured by the residual sum of squares (RSS):

$$\mathrm{RSS}_1 = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right), \tag{4}$$

where y and $\hat{\beta}$ are the 1st column of Y and $\hat{B}$, respectively.

To detect if x_2 Granger-causes x_1, the above regression procedure is repeated by removing x_2. i.e. prediction of x_1 is made using past observations of x_1 only:

$$x_{1,t} = \sum_{l=1}^{p} A_{11,l}x_{1,t-l} + e_{1,t}.$$

The RSS of x_1 obtained by regression without x_2 (restricted model) is denoted by RSS_{21}. Since RSS_1 denotes RSS of unrestricted model, where both x_1 and x_2 are included, the F statistic can be constructed as [9], [10]:

$$F = \frac{\left(\mathrm{RSS}_{21} - \mathrm{RSS}_1\right)/p}{\mathrm{RSS}_1/(m-2p)}. \tag{5}$$

Under the null hypothesis that x_2 does not Granger-cause x_1, which is equivalent to $A_{12,l}$ being all zero, the F statistic has a $F(p,m-2p)$ distribution, so the corresponding p-value can be calculated. If RSS_1 is much smaller than RSS_{21}, (5) yields a large F value, so resulting in a small p-value, in which case a significant GC is detected and we can conclude that x_2 Granger-causes x_1.

Similarly, to see if x_1 Granger-causes x_2, the aforementioned routine can be repeated by exchanging the roles of x_1 and x_2.

2. Multivariate autoregressive model

GC can be extended to multivariate case [4], [7], where the number of variables of the system $n \geq 3$. Suppose there are n time series. If including the history of variable j reduces the prediction error of variable i, compared to exclusion of variable j, with series of all other variables always included in the prediction model, then variable j Granger-causes variable i. Notice that the detection of causality is conditioned on all other series, hence it is also called conditional GC (CGC). Analogous to bivariate case, we can write out the procedure of CGC explicitly as follows.

Suppose n variables are measured at T time instants. Let an $n\times 1$ vector x_t denote the measurements at time t, the VAR model of order p can be expressed as

$$x_t = \sum_{l=1}^{p} A_l x_{t-l} + e_t \quad , \quad t = p+1,...,T \tag{6}$$

where A_l is an $n\times n$ coefficient matrix containing parameters of the VAR model, and e_t is a $n\times 1$ error vector containing residuals (prediction errors).

Taking transpose of the $m = T-p$ equations in (6) and then stacking them, (6) can be written in the form of standard linear regression:

$$Y = XB + E \tag{7}$$

where

$$\underset{(m\times n)}{Y} = \begin{bmatrix} x_{p+1}' \\ x_{p+2}' \\ \vdots \\ x_T' \end{bmatrix}, \quad \underset{(m\times np)}{X} = \begin{bmatrix} x_p' & x_{p-1}' & \cdots & x_1' \\ x_{p+1}' & x_p' & \cdots & x_2' \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-1}' & x_{T-2}' & \cdots & x_{T-p}' \end{bmatrix},$$

$$\underset{(np\times n)}{B} = \begin{bmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{bmatrix}, \quad \underset{(m\times n)}{E} = \begin{bmatrix} e_{p+1}' \\ e_{p+2}' \\ \vdots \\ e_T' \end{bmatrix}.$$

The coefficient matrix B can be estimated by OLS, provided that $m \geq np$ so $X'X$ is invertible:

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

$$\hat{B} = (X'X)^{-1}X'Y. \quad (8)$$

Then, the prediction error of variable i can be measured by RSS:

$$\mathrm{RSS}_i = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right), \quad (9)$$

where y and $\hat{\beta}$ are the i-th column of Y and $\hat{B}$, respectively.

To detect if there is GC from variable j to variable i, we remove variable j and repeat the above linear regression with n–1 variables to get the RSS of variable i of this restricted model – denoted RSS_{ji}. The F statistic can be constructed as:

$$F = \frac{\left(\mathrm{RSS}_{ji} - \mathrm{RSS}_i\right)/p}{\mathrm{RSS}_i/(m - np)}. \quad (10)$$

Under the null hypothesis that variable j does not Granger-cause variable i, the F statistic has a $F(p, m-np)$ distribution, then p-value can be calculated.

For the purpose of GRN discovery, we calculate the F statistics of all $M = n(n-1)$ combinations of directed variable pair j to i, so the p-values of all possible edges of the n-variable network are obtained.

C. Conditional and Pairwise GC

For a system of n variables, to detect GC from variable j to variable i, if conditional GC (CGC) is applied to all variables simultaneously, this is a full model application, where only one coefficient matrix B for unrestricted model is estimated.

However, in GRN discovery using DNA microarray time-series data, the number of variables (genes) is much larger than the data length. A full model cannot be applied to all genes simultaneously because the condition $m \geq np$ for OLS is usually violated. Therefore, bivariate autoregressive model is often applied to the n genes pairwisely. i.e. to detect if gene j Granger-causes gene i, the procedure described in "bivariate autoregressive model" above is executed using time series of genes i and j only. To find p-values of all $M = n(n-1)$ possible edges of the n-gene network, totally $n(n-1)/2$ coefficient matrices B for unrestricted models are estimated because there are totally $n(n-1)/2$ pairs of genes $\{i,j\}$. We call this implementation pairwise GC (PGC).

D. Correction to Multiple Testing

Both CGC and PGC return $M = n(n-1)$ p-values for an n-variable network. The M hypothesis tests for GC detection need correction and there are mainly two approaches [7]: Bonferroni correction and Benjamini-Hochberg false discovery rate (FDR) controlling procedure [11].

The Bonferroni approach is also known as controlling the family-wise error rate. If there are M tests, to discover a network at a significance level α, each individual test should be executed at level α/M. i.e. only those causalities having p-values $\leq \alpha/M$ are considered as significant and included as edges in the discovered network. Using Bonferroni correction, the probability of having one or more false positives is controlled at α [11]. In this paper, for the purpose of minimizing spurious causalities (false positives), we adopt the stricter Bonferroni correction with α=0.05 for simulations using synthetic data (Section III).

The Benjamini-Hochberg FDR controlling procedure [11] is described as follows. Suppose we need to test M null hypotheses H_1, H_2, …, H_M with corresponding p-values p_1, p_2, …, p_M. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$ be the ordered p-values, and denote $H_{(i)}$ to be the null hypothesis corresponding to $p_{(i)}$. To control FDR at q: if K is the largest i that

$$p_{(i)} \leq \frac{i}{M}q,$$

then reject all $H_{(i)}$ where i = 1, 2, …, K. In other words, the FDR of these K discoveries is q. In this paper, to compare results with previous publications, we adopt this FDR controlling procedure with q=0.05 in real data application (Section IV).

E. Model Validation

If a VAR model does not adapt to the data well, correlations between the variables cannot be captured properly, implying that the GC detected is not reliable. The GCCA toolbox offers three implementations to check if a model is valid [7]:

1. Model consistency: to measure if a VAR model can capture the correlation structure of the data sufficiently. Consistency is computed as:

$$C = \left(1 - \frac{\left|R_p - R_i\right|}{\left|R_i\right|}\right) \times 100\% \quad (11)$$

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. We take the 80% threshold suggested by Seth [7] and treat a model with lower consistency as invalid.

2. Adjusted RSS: in GCCA, the adjusted RSS of variable i is calculated as:

$$\nu_i = 1 - \frac{\mathrm{RSS}_i/(m - np)}{y'y/m}. \quad (12)$$

In (12), the numerator is the estimated variance of prediction errors and the denominator is the variance of input data. Thus ν_i measures the amount of input data captured by the model. We require ν_i of all variables ≥ 0.3 for a model to pass this validation.

3. The Durbin-Watson test (whiteness test): to test if the residuals are serially uncorrelated. GCCA has implemented the

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

procedure in [12]. We follow the default setting of GCCA to use Bonferroni correction at significant level 0.05 for this whiteness test, too.

III. Simulation Results

Our simulations are carried out using the GCCA toolbox [7]. Consider a 3-variable model used in [13] (i.e. n=3):

$$\begin{aligned} X_t &= 0.8X_{t-1} - 0.5X_{t-2} + 0.4Z_{t-1} + \varepsilon_t \\ Y_t &= 0.9Y_{t-1} - 0.8Y_{t-2} + \xi_t \\ Z_t &= 0.5Z_{t-1} - 0.2Z_{t-2} + 0.5Y_{t-1} + \eta_t \end{aligned} \tag{13}$$

where ε_t, ξ_t and η_t are independent Gaussian white noise processes of zero mean and unit variance. Initial values of X, Y and Z have the same nature as ε, ξ and η. Model (13) has 2 causal inferences "$Z\to X$" and "$Y\to Z$". After simulations using (13), we drop the first 100 time points which are transient and take the subsequent 200 time points as our synthetic data, i.e. T=200. Applying CGC to these 3 time series with model order p=2, a discovered network of 2 edges is obtained. Fig. 1(a) shows this discovered network, we can see this CGC recovers the 2 causal inferences in model (13) exactly.

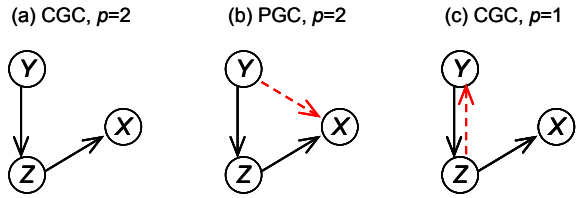

Figure 1. Discovered networks on the 3-variable model. Solid black arrows represent true positives, dashed red arrows represent false positives.

A. False Discoveries in Pairwise GC

Using the same synthetic data generated above, if PGC with p=2 is applied, a discovered network of 3 edges is obtained, which is shown in Fig. 1(b). The spurious edge "$Y\to X$" is a false discovery. It comes from the bivariate modeling on time series X and Y only. Since including Y helps to predict X, a causal inference from Y to X is detected. However, we know that this is actually an indirect inference through Z. Compare Fig. 1(a) obtained by CGC, where the inference "$Y\to X$" is tested by conditioning on Z (i.e. Z is included in the regression model). Since Z already helps predicting X, including Y does not help any more, CGC does not identify "$Y\to X$" as a causal inference. In short, CGC can distinguish direct and indirect inferences, but PGC cannot, resulting in false discoveries.

B. False Discoveries in Conditional GC if Model Order is Not High Enough

Using the same synthetic data as before, if CGC with p=1 is applied, the discovered network also has false discovery, as shown in Fig. 1(c). This time, a spurious edge "$Z\to Y$" exists, which can be explained by Fig. 2. Solid arrows mark the true inferences. However, if the VAR model does not have time lag = 2 part, as crossed out by dashed lines in Fig. 2, since Z_{t-1} contains information of Y_{t-2}, and Y_{t-2} helps predicting Y_t, these imply Z_{t-1} helps predicting Y_t. Hence, spurious edge "$Z\to Y$" appears.

A closer examination of model (13) and applying the same argument as above, the edge "$X\to Z$" should also appear in Fig. 1(c). Indeed, we observe that if data length T=1000, this edge also appears because a longer data length has higher statistical power. Since the spurious feedback causality "$X\to Z$" is relatively weaker, it is not identified as a significant causality when T=200. In short, using an inappropriately low model order can easily make acyclic system appear as cyclic.

Nevertheless, even we use a long T=1000, figures 1(a) and 1(b) still remain the same as T=200. On the other hand, if CGC of p=3 is applied, the discovered network is the same as p=2 in Fig. 1(a).

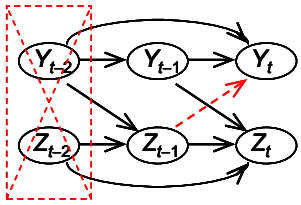

Figure 2. Arising of spurious feedback causality. Solid black arrows represent true inferences, dashed red arrow indicates spurious inference.

C. Reducing False Discoveries in PGC by Model Validation

For the purpose of investigating PGC more clearly, a 5-variable model is adopted in this sub-section such that more pairs of variables are available. We follow the 5-variable model used in [7] and [14]:

$$\begin{aligned} x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + w_1(t) \\ x_2(t) &= 0.5x_1(t-2) + w_2(t) \\ x_3(t) &= -0.4x_1(t-3) + w_3(t) \\ x_4(t) &= -0.5x_1(t-2) + 0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + w_4(t) \\ x_5(t) &= -0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + w_5(t) \end{aligned} \tag{14}$$

where w_i and initial values of x_i (i=1,2,3,4,5) are independent Gaussian white noise processes of zero mean and unit variance. Similar as before, the first 100 time points are dropped and subsequent 200 time points are taken as our synthetic data. i.e. now T=200, n=5. The total number of all possible edges is $M = n(n-1) = 20$. The model (14) has order p=3. Applying CGC to synthetic data with model order p=3 completely recovers the ground truth network, which is shown in Fig. 3(a).

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

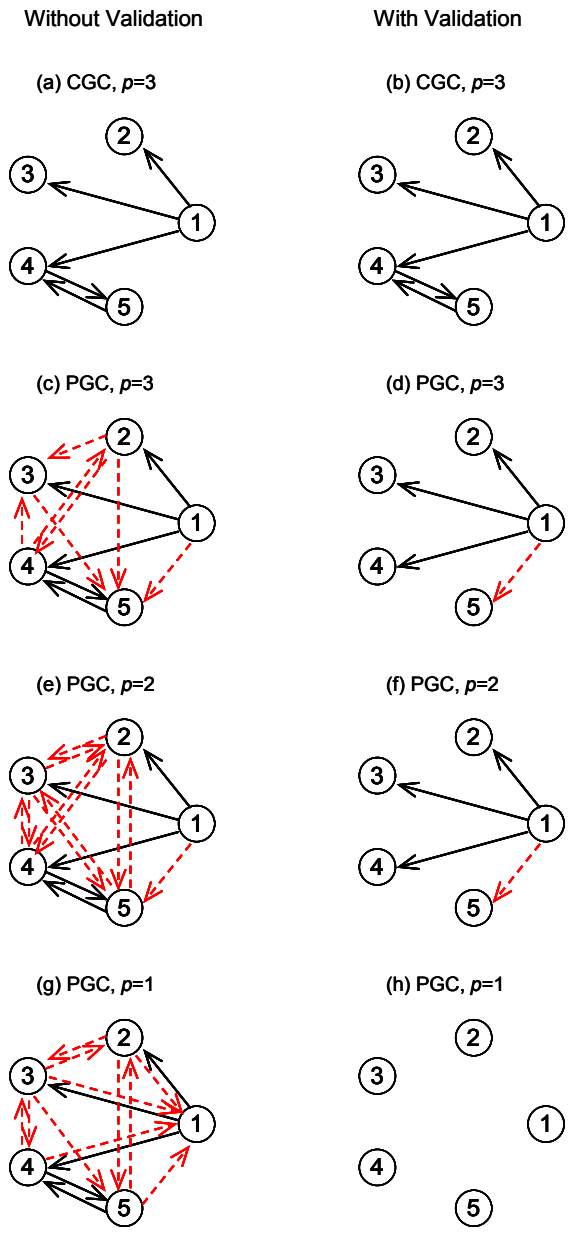

Figure 3. Discovered networks on the 5-variable model. Solid black arrows represent true positives, dashed red arrows represent false positives.

Before moving to PGC, let us introduce some measures on discovered network. Suppose a discovered network has H edges, e of them are true positives (TP) and the rest are false positives (FP). Then, precision $P = e/H$ and false discovery rate $Q = (H-e)/H = 1-P$. Denote the number of edges of the ground truth network by L, then recall $R = e/L$, the harmonic mean between P and R is $F_1 = 2PR/(P+R)$. A good discovered network should have high P and R, and thus also F_1, implying that Q should be low. As an example, the 5-variable model (14) has L=5. The discovered network in Fig. 3(a) has H=5 and e=5, so P=R=F_1=1 and Q=0, hence it is a perfect discovered network.

Now, we apply PGC to the synthetic data. The total number of variable pairs $\{i,j\}$ is $n(n-1)/2 = 10$, so totally 10 bivariate models should be applied. We try p=3,2,1 and results are shown in Table I. We can see PGC without model validation performs badly. H is much larger than L, meaning that many spurious edges exist. The middle part of Table I shows how many unrestricted models pass validation. CGC has one unrestricted model only and it passes all validation tests, so using validation or not makes no difference on its results. PGC has 10 unrestricted models, some models fit well thus pass validation, but some are not, hence using validation makes a difference.

Take p=2 row as an example, without validation, 16 edges out of 20 are detected as significant, but we know most of them are spurious. Applying validation, only 0.4×10=4 unrestricted models can pass all the three tests. That means only 4×2=8 candidate edges are considered in significance test subject to Bonferroni correction, where p-values ≤ α/8 are considered as significant. It turns out only 4 edges out of 8 are significant. They are shown in Fig. 3(f). Now, only 1 edge is false, giving Q=0.25 which is much lower than Q=0.69 without validation.

Similar observation can be seen at PGC p=3 row. Since results without validation are not so bad as p=2, improvement made by validation is relatively smaller. For p=1, no model can pass all the three validation tests. That means model order 1 is too low that all the 10 bivariate models cannot fit well with the data, thus there is no discovery (H=0). Higher p should be used for valid discovery.

TABLE I. PERFORMANCE OF CGC AND PGC ON THE 5-VARIABLE SYNTHETIC DATA

		Cons[a] (%)		Without Validation			Proportion Passing Validation[c]				With Validation			Improvement[f]	
	p	*mean*	*SD*[b]	*H*	F_1	*Q*	*Cons*	*Adj. RSS*[d]	*Whiteness*	*All 3*[e]	*H*	F_1	*Q*	F_1	*Q*
CGC	3	91	0	5	1	0	1	1	1	1	5	1	0	---	---
PGC	3	74	17	12	0.59	0.58	0.4	1	1	0.4	4	0.67	0.25	0.08	0.33
	2	70	20	16	0.48	0.69	0.4	1	0.7	0.4	4	0.67	0.25	0.19	0.44
	1	52	16	16	0.48	0.69	0	0.5	0.2	0	0	---	---	---	---

a. model consistency (total number of unrestricted models are 1 and 10 for CGC and PGC, respectively); b. standard deviation; c. proportion passing validation = (number of unrestricted models passing validation) / (total number of unrestricted models); d. Adjusted RSS; e. all the 3 validation tests; f. improvement (in F_1 or Q) by validation in PGC case, measured in magnitude.

Fig. 3 shows all 8 discovered networks corresponding to Table I. In short, these results show that PGC may have many false discoveries, and these false discoveries can be substantially reduced by model validation.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

In real data application, if PGC can identify a network of inferences among a small number (say n) of gene for which the condition $m \geq np$ for OLS can be met, all these inferences can be validated again by applying CGC to the n genes, hence false discoveries can be further minimized.

IV. Application of Pairwise GC to Real Data

When applying GC to real data from DNA microarray experiment, the model order p can be estimated by the Akaike information criterion (AIC) [15] or the Bayesian information criterion (BIC) [16]. Publications [5] and [6] have applied PGC with p=1 to the HeLa cell-cycle dataset [17]. From previous investigation using synthetic data, we doubt if p=1 is sufficient. Thus, we estimate the model order p first.

A. Estimating the Model Order by AIC and BIC

Experiment 3 of the HeLa dataset have the longest time series and all time points are equally spaced, so many people have applied GC on data of this experiment. Reference [17] identified 1134 periodic genes, 1099 of them do not have missing values and we focus our study on these 1099 genes (same as [6]). Experiment 3 has time points t=0,0,1,2,3,…,46. For each gene, we average the two measurements of t=0. Similar to [5] and [6], we have not executed any other trend removal or pre-processing that may distort the data. Now, n=1099 and T=47, obviously CGC cannot be applied on all genes simultaneously, hence PGC is used.

For the 1099 genes, the total number of gene pairs is $n(n-1)/2 = 603351$, so totally 603351 bivariate models are applied. For each pair, equivalently each bivariate model, we estimate the model order p by AIC and BIC giving candidate p=1,2,3,4,5. For comparison, bivariate models with fixed p=1,2,3 are also implemented. For each case above, one row in Table II shows the results of the 603351 bivariate models. Distributions of the 603351 model orders estimated by AIC and BIC are shown in Fig. 4. We can see AIC generally returns higher model orders than BIC. From Table II, the proportion of bivariate models passing all the 3 validation tests is 9.0%, which is the highest among the five cases. So, model orders estimated by AIC are the most suitable. Note that though BIC gives a mean model order 1.6 which is lower than p=2, BIC still has much more bivariate models (5.3%) passing all the 3 validation tests than p=2 case (2.6%), meaning than using BIC is still much better than fixing p=2 for all gene pairs.

TABLE II. Different Cases of Fixing the Model Order

	Model Order		**Cons[a] (%)**		**Proportion Passing Validation (%)[c]**			
	Mean	***SD[b]***	***Mean***	***SD***	***Cons***	***Adj. RSS[d]***	***Whiteness***	***All 3[e]***
AIC	3.3	1.5	51	23	12.9	99.7	37.4	9.0
BIC	1.6	1.0	43	23	7.4	98.7	30.3	5.3
p=3	3	0	48	21	8.1	99.2	32.3	5.5
p=2	2	0	44	19	4.0	97.6	29.6	2.6
p=1	1	0	39	20	1.1	93.9	22.7	0.6

a. model consistency; b. standard deviation; c. proportion passing validation in % = (number of bivariate models passing validation) / (total number of bivariate models) × 100%; d. Adjusted RSS; e. all the 3 validation tests.

B. Results from AIC and Model Validation

Since AIC is the most suitable, we further analyze its results. Its 9.0% bivariate models passing all the 3 validation tests correspond to 54376 gene pairs. That means we need to test 54376×2=108752 null hypotheses. Similar to [5] and [6], we also apply Benjamini-Hochberg FDR controlling procedure with q=0.05 for the multiple testing. The corrected threshold for p-value is 0.015, and 33601 causalities are found to be significant.

Before proceeding, let us introduce degree distribution which is a useful illustration in gene network study. In a network, the in-degree of a gene is the number of edges pointing into it and the out-degree is the number of edges going out of it. The degree of a gene is the sum of its in-degree and out-degree [18]. For a large discovered network that cannot be drawn as Fig. 1 and Fig. 3, the distribution of the (in-/out-) degrees of all the genes in the network can be plotted.

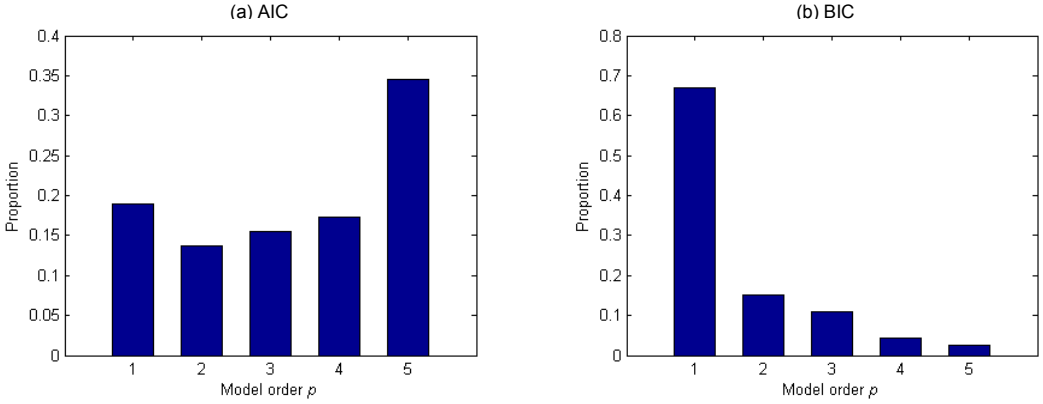

Figure 4. Distributions of the model orders estimated by AIC and BIC.

Regarding the 54376 gene pairs, 29882 pairs are detected to have significant GC, where 3719 pairs (12% of 29882) have cyclic GC. These are consistent with the 33601 edges (causalities) of the discovered network. Among the 1099 genes, 968 genes have in-degree > 0 and 826 genes have out-degree > 0. The union set involves 986 genes, which is the set of genes

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

having degree > 0. The degree distributions of the 986 genes are plotted in Fig. 5, which shows similar decaying trend as in [5] and [6]. Compare the skewness and kurtosis of Fig. 5 (a) and (b) with [6], they are comparable to that of [6] for in-degree but much larger for out-degree.

Table III shows the top 10 genes with maximum (in-/out-) degrees corresponding to the three cases in Fig. 5. The hubs with highest degrees shown in Table III(c) generally have higher out-degrees than in-degrees, meaning that the hubs tend to act as sources of interactions rather than receivers of interactions. Compare Table III (a) and (b) with [6], our top 10 genes do not overlap with [6]. On the other hand, 5 out of the 10 genes in Table III(c) overlap with [19], which implemented their grouped graphical Granger modeling on the HeLa dataset. These 5 genes are: 5.CDC2, 23.KNSL5, 26.CDC2, 42.DJ616B8.3 and 87.USF1.

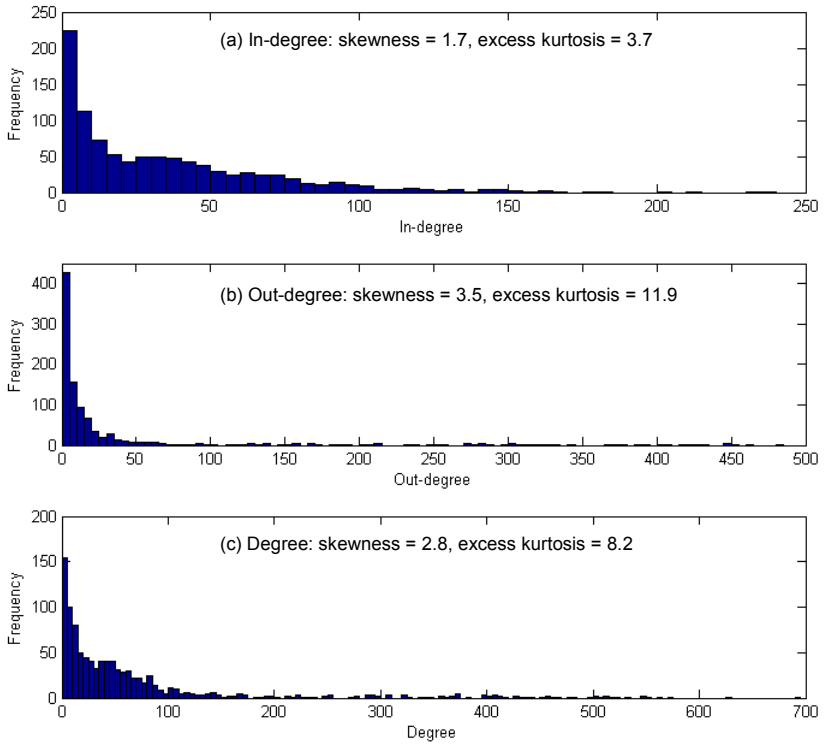

Figure 5. Degree distributions of the 986 genes involved in significant causalities. Excess kurtosis = kurtosis – 3. So, skewness and excess kurtosis are both zero for Gaussian distribution.

TABLE III. TOP 10 GENES WITH MAXIMUM (IN-/OUT-) DEGREES.

(a) In-degree	
In-degree	***Gene name***
238	333.MCM6
232	257.ESTs
211	5.CDC2
202	73.TTK
180	26.CDC2
178	190.MDS025
165	100.ESTs
161	68.SMAP
160	66.CKAP2
155	203.TUBB2

(b) Out-degree	
Out-degree	***Gene name***
480	5.CDC2
480	47.KPNA2
461	42.DJ616B8.3
452	87.USF1
447	3.UBE2C
447	26.CDC2
445	20.STK15
445	22.UBE2C
430	10.FLJ10468
426	16.TOP2A

(c) Degree	
Degree	***Gene name***
691	5.CDC2
627	26.CDC2
572	3.UBE2C
562	6.TOP2A
552	47.KPNA2
546	23.KNSL5
545	42.DJ616B8.3
532	22.UBE2C
521	87.USF1
515	11.CCNF

The number in front of each gene name is the row number in the data file dataPlusScores_all5.txt downloaded at the web link shown in the abstract of [17].

V. DISCUSSION

Some readers may think that the condition $m \geq np$ for solving (7) by OLS is too strict. Actually, if this condition is not satisfied, (7) can be solved by regularization techniques [4] [20]. However, rigorous and practical statistical tests on that aspect are still needed to be developed. Our preliminary study (not shown here) reveals that if the condition $m \geq np$ can be met for a small network, OLS will give better performance than regularization because regularization usually imposes bias [4], [20]. Thus, here we choose to report our results obtained by OLS first. Different GC implementations (including

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

regularization) will be investigated later. Their comparisons with other GRN discovery methods (e.g. Bayesian networks, mutual information approaches) [2]–[4] are also our future work.

Our synthetic data simulations use data length T=200, which is much longer than T=47 in real data. It is because the plots of the time series generated by models (13) and (14) look like noises. If T is 50 or 100, the statistical power yielded by these synthetic data is too low that sometimes true causalities cannot be detected, which makes our investigation difficult. However, for the real dataset HeLa, the genes exhibit periodic patterns, causalities can be detected more easily though they are shorter.

On the other hand, real data of $T \geq 100$ is not impossible, since many studies carry out multiple experiments, e.g. [17] have done 5 experiments. These experiments may be used together such that the total number of time points can be ≥ 100. However, integrating multiple time series is not a trivial task, and we have started research on this topic.

VI. Conclusions

Using synthetic data, we have shown that false discoveries easily arise in pairwise GC implementation, where indirect inferences are often mistakenly taken as direct inferences. Even in full model (CGC) application, if the model order is not sufficiently high, false discoveries may cause acyclic system to appear as cyclic. To remedy these problems, we have demonstrated that model validation can effectively reduce the number of false discoveries. We also recommend using full model instead of pairwise GC if possible.

The application of pairwise GC to the HeLa dataset shows that AIC is better than BIC in estimating the model order, in the sense that AIC leads more bivariate models passing all the 3 validation tests. With model validation, degree distributions of the discovered network show similar decaying trends as previous publications. Network hubs tend to act as sources of interactions rather than receivers of interactions.

We have also discussed a few related issues and mentioned future work.

Acknowledgment

We would like to thank Z. G. Zhang for discussion on implementation of GC, and the University Research Committee of the University of Hong Kong for funding support. We also benefit from comments of anonymous reviewers and Z. G. Zhang on revising the manuscript.

Y. S. Hung and C. Q. Chang would like to acknowledge support by Hong Kong SAR Research Grants Council (Project No. HKU 762111M) and CRCG of the University of Hong Kong.

References

[1] S. Zhang, G. Jin, X. S. Zhang, and L. Chen, "Discovering functions and revealing mechanisms at molecular level from biological networks," Proteomics, vol. 7, no. 16, pp. 2856–2869, August 2007.

[2] G. Karlebach, and R. Shamir, "Modelling and analysis of gene regulatory networks," Nat. Rev. Mol. Cell Biol., vol. 9, no. 10, pp. 770–780, October 2008.

[3] L. Chen, R. S. Wang, and X. S. Zhang, Biomolecular Networks: Methods and Applications in Systems Biology. NJ: Wiley, 2009, pp. 72–73.

[4] Z. G. Zhang, Y. S. Hung, S. C. Chan, W. C. Xu, and Y. Hu, "Modeling and identification of gene regulatory networks: a Granger causality approach," in Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, China, vol. 6, pp. 3073–3078, July 2010.

[5] N. Mukhopadhyay, and S. Chatterjee, "Causality and pathway search in microarray time series experiment," Bioinformatics, vol. 23, no. 4, pp. 442–449, February 2007.

[6] R. Nagarajan, and M. Upretiy, "Granger causality analysis of human cell-cycle gene expression profiles," Stat. Appl. Gen. & Mol. Bio., vol. 9, iss. 1, art. 31, 2010.

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

[8] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica, vol. 37, no. 3, pp. 424–438, August 1969.

[9] J. D. Hamilton, Time Series Analysis. NJ: Princeton University Press, 1994, p. 305.

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer-Verlag, 2009, p. 48.

[11] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. Roy. Statist. Soc. Ser. B, vol. 57, no. 1, pp. 289–300, 1995.

[12] J. Durbin, and G. S. Watson, "Testing for serial correlation in least squares regression: I," Biometrika, vol. 37, pp. 409–428, December 1950.

[13] M. Ding, Y. Chen, and S. L. Bressler, "Granger causality: basic theory and application to neuroscience," in Handbook of Time Series Analysis, S. Schelter, M. Winterhalder, and J. Timmer, Eds. Wienheim: Wiley, 2006, pp. 438–460.

[14] L. A. Baccalá, and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," Biol. Cybern., vol. 84, no. 6, pp. 463–474, June 2001.

[15] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Automat. Contr., vol. 19, iss. 6, pp. 716–723, December 1974.

[16] G. Schwartz, "Estimating the dimension of a model," Ann. Stat., vol. 6, no. 2, pp. 461–464, 1978.

[17] 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.

[18] F. Schreiber, "Graph theory," in Analysis of Biological Networks, B. H. Junker, and F. Schreiber, Eds. NJ: Wiley, 2008, pp. 15–28.

[19] 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.

[20] J. Fan, and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," J. Am. Stat. Assoc., vol. 96, no. 456, pp. 1348–1360, December 2001.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE