IDENTIFICATION OF GRANGER CAUSALITY BETWEEN GENE SETS

ANDRÉ FUJITA

*Computational Science Research Program, RIKEN, 2-1, Hirosawa, Wako, Saitama, 351-0198,
Japan.*
andrefujita@riken.jp

JOÃO RICARDO SATO

*Center of Mathematics, Computation and Cognition, Universidade Federal do ABC, Rua Santa
Adélia, 166 - Santo André.*
joao.sato@ufabc.edu.br

KANAME KOJIMA

*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai,
Minato-ku, Tokyo, 108-8639, Japan.*
kaname@ims.u-tokyo.ac.jp

LUCIANA RODRIGUES GOMES

*Chemistry Institute, University of São Paulo, Av. Lineu Prestes, 748 - São Paulo, 05508-900,
Brazil.*
lugomes@iq.usp.br

MASAO NAGASAKI

*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai,
Minato-ku, Tokyo, 108-8639, Japan.*
masao@ims.u-tokyo.ac.jp

MARI CLEIDE SOGAYAR

*Chemistry Institute, University of São Paulo, Av. Lineu Prestes, 748 - São Paulo, 05508-900,
Brazil.*
mcsoga@iq.usp.br

SATORU MIYANO

*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai,
Minato-ku, Tokyo, 108-8639, Japan.*
miyano@ims.u-tokyo.ac.jp

Wiener and Granger have introduced an intuitive concept of causality (Granger causal-

ity) between two variables which is based on the idea that an effect never occurs before its cause. Later, Geweke has generalized this concept to a multivariate Granger causality, i.e., n variables Granger-cause another variable. Although Granger causality is not "effective causality" in the Aristothelic sense, this concept is useful to infer directionality and information flow in observational data. Granger causality is usually identified by using VAR (Vector Autoregressive) models due to their simplicity. In the last few years, several VAR-based models were presented in order to model gene regulatory networks. Here, we generalize the multivariate Granger causality concept in order to identify Granger causalities between *sets* of gene expressions, i.e., whether a set of n genes Granger-causes another set of m genes, aiming at identifying the flow of information between gene networks (or pathways). The concept of Granger causality for *sets* of variables is presented. Moreover, a method for its identification with a bootstrap test is proposed. This method is applied in simulated and also in actual biological gene expression data in order to model regulatory networks. This concept may be useful to understand the complete information flow from one network or pathway to the other, mainly in regulatory networks. Linking this concept to graph theory, sink and source can be generalized to node sets. Moreover, hub and centrality for sets of genes can be defined based on total information flow. Another application is in annotation, when the functionality of a set of genes is unknown, but this set is Granger caused by another set of genes which is well studied. Therefore, this information may be useful to infer or construct some hypothesis about the unknown set of genes.

Keywords: Granger causality; Canonical correlation analysis; Vector autoregressive model.

1. Introduction

Norbert Wiener [52] conceived the notion that, if the prediction of one time series could be statistically improved by incorporating the knowledge of past values of a second one, then the latter is said to have a "causal" influence on the former. Clive Granger [18] has formalized the prediction idea in the context of VAR (Vector Autoregressive) models. More specifically, if the variance of the autoregressive prediction error of one time series at the present time is statistically reduced by inclusion of past measurements from another time series, then the latter is said to have a Granger causal influence on the former. From this definition it is clear that the flow of time plays a crucial role in allowing inferences to be made about directional causal influences from time series data. Due to its simplicity and intuitive idea that an effect never occurs before its cause, it has been widely used in several areas such as Econometrics [35] [53] [34], Neuroscience [4] [42] [43] and more recently, in Bioinformatics [36] [11] [12] [13] [14] [27] [37] [38]. The idea is that, if a variable x affects a variable y, past values of the former should be useful in generating predictions for the latter variable.

Later, Geweke [17] has generalized the bivariate Granger causality to a multivariate fashion in order to identify conditional Granger causality. To illustrate the bivariate and multivariate Granger causalities, suppose three processes, where one drives the other two with differential time delays. A pairwise analysis would indicate a causal influence from the process that receives an early input to the process that receives a late input. Conditional Granger causality may be useful to disambiguate these situations, since it has the ability to resolve whether the interaction

between two time series is direct or is mediated by another recorded time series and whether the causal influence is simply due to differential time delays in their respective driving inputs.

Nagarajan and Upreti (2008) [37] and Nagarajan (2009) [38] investigated the use of bivariate VAR for acyclic approximations in the case of networks composed of two genes by exploring the parameters defined as transcriptional noise variance, autoregulatory feedback, and transcriptional coupling strength, which may influence some measures of Granger Causality. These authors have shown that under some specific conditions, VAR parameters may have a pronounced effect on the statistical tests used for identifying significant Granger causality.

Frequently, conditional Granger causality is identified by using VAR models due to its simplicity [31]. In the last few years, several extensions of the standard VAR model namely, Dynamic VAR (DVAR) [11], Sparse VAR (SVAR) [12] [39] [44] [30] and Nonlinear VAR (NVAR) [13] [8] [16] [32] [33] have been proposed to model gene regulatory networks. DVAR were developed in order to identify time-varying Granger causalities, i.e., to identify structural changes along time (cell cycle) in regulatory networks. However, DVAR identifies only linear Granger causalities. In order to overcome this limitation, a model capable to identify nonlinear Granger causalities, namely, NVAR, was introduced. Another problem in bioinformatics is the high dimensional characteristic of gene expression data, where the number of parameters (genes) is higher than the number of observations (microarrays). Therefore, constructing regulatory networks with statistical tests are difficult. In order to identify linear Granger causalities in this context, SVAR was proposed [12] [39] [44].

These models are useful to identify Granger causalities between genes (gene expression signals), however, sometimes, it is necessary to identify Granger causality between *sets* of genes. Thus, for example, one may need to identify Granger causality from one gene cluster to the other gene cluster in order to study the information flow between different regulatory pathways and to identify which pathway is crucial. Another application may be the identification of Granger causality between a set of genes whose biological function is yet unknown and another set of genes which are better studied and for which more characteristics are known. This latter case may help to determine and infer the possible biological process affected by the set of genes whose functionalities are not yet known.

The theoretical generalization of Granger causality between *sets* of variables has not been sufficiently explored. Here, we propose a definition of Granger causality between *sets* of time series, where m time series Granger-cause n other time series, thus generalizing the two previous definitions (bivariate and multivariate). Furthermore, a method for identification of Granger causality and a statistical test based on nonparametric bootstrap are also proposed. The results are illustrated by simulations and, also, by application of this new concept to actual biological gene expression data.

2. Methods

2.1. *Granger causality*

In order to formalize the concept of Granger causality between *sets* of time series, suppose that $\Im_t$ is a set containing all relevant information available up to and including time-point t. Let $\mathbf{Y}_t$, $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$ be sets of time series containing k, m and n time series, respectively, where $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$ are disjoint subsets of $\mathbf{Y}_t$, i.e., each time series only belongs to one set, and thus, $k \geq m + n$. Let $\mathbf{Y}_t(h|\Im_t)$ be the optimal (i.e., the one which produces the minimum mean squared error (MSE) prediction) h-step predictor of the set of m time series $\mathbf{Y}_t^i$ from the time point t, based on the information in $\Im_t$. The forecast MSE of the linear combination of $\mathbf{Y}_t^i$ will be denoted by $\Omega_Y(h|\Im_t)$. The set of n time series $\mathbf{Y}_t^j$ is said to Granger-cause the set of m time series $\mathbf{Y}_t^i$ if

$$\Omega_Y(h|\Im_t) < \Omega_Y(h|\Im_t \backslash \{\mathbf{Y}_s^j | s \leq t\}) \quad \text{for at least one } h = 1, 2, \ldots, \quad (1)$$

where $\Im_t \backslash \{\mathbf{Y}_s^j | s \leq t\}$ is the set containing all relevant information except for the information in the past and present of $\mathbf{Y}_t^j$. In other words, if $\mathbf{Y}_t^i$ can be predicted more accurately when the information in $\mathbf{Y}_t^j$ is taken into account, then $\mathbf{Y}_t^j$ is said to be Granger-causal for $\mathbf{Y}_t^i$.

Applying the idea of Granger causality to regulatory networks, a set of gene expression time series $\mathbf{Y}_t^j$ Granger-causes another set of gene expression time series $\mathbf{Y}_t^i$, if linear combinations of $\mathbf{Y}_t^j$ provide statistically more significant information about future values of linear combinations of $\mathbf{Y}_t^i$ than considering only the past values of $\mathbf{Y}_t^i$. Thus, past gene expression values of $\mathbf{Y}_t^j$ allow the prediction of more accurate gene expression values of $\mathbf{Y}_t^i$. Notice that since this relationship is not reciprocal, Granger causality may be interpreted as information flow [2]. Moreover, it is important to highlight that Granger causality is not actually inferring "effective" causality in the Aristothelic sense, i.e., interaction of gene products (or protein-protein interactions), since the former is based solely on prediction and quantitative criteria, as described before, however, this concept may be useful to suggest some insights on molecular interactions which may then be experimentally tested.

Notice that this definition generalizes not only the original bivariate Granger causality (where $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$ are 1-dimensional) but, also, the multivariate Granger causality (where $\mathbf{Y}_t^j$ is m-dimensional and $\mathbf{Y}_t^i$ is 1-dimensional), i.e., the bivariate and multivariate cases are special cases of the Granger causality for *sets* of time series.

2.2. *Identification*

Due to the simplicity of notation and concepts, only the identification of Granger causality in an Autoregressive (AR) process of order one will be presented, generalization for other orders being straightforward.

Hotelling [21] [22] was the first to tackle the problem of identifying and measuring relationships between two *sets* of variables, introducing the Canonical Correlation

Analysis (CCA) [54] [41] [48] [15]. Canonical correlation analysis can be understood as the bivariate correlation of two synthetic variables which are the linear combinations of the two sets of observed variables. The original variables of each set are linearly combined to produce pairs of synthetic variables which have maximal correlation.

The present problem consists in verifying whether $\mathbf{Y}_t^j$ Granger-causes $\mathbf{Y}_t^i$. In the linear case, Granger causality for *sets* of variables may be identified based on the idea initially proposed by Hotelling. Notice that linear combinations provide simple summarized measures of a set of variables.

Therefore, for the linear case, let $\mathbf{Y}_t^i$ (m-dimensional) and $\mathbf{Y}_t^j$ (n-dimensional) be two disjoint subsets of $\mathbf{Y}_t$, where $\mathbf{Y}_t$ is a k-dimensional set of stationary time series ($k \geq m + n$), in other words, each time series only belongs to one cluster. Then, $\mathbf{Y}_t^j$ is Granger non-causal for $\mathbf{Y}_t^i$ if the following condition holds:

$$CCA(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^j | \mathbf{Y}_t \backslash \{\mathbf{Y}_{t-1}^j\}) = \rho = 0 \qquad (2)$$

where ρ is the largest correlation calculated by Canonical Correlation Analysis. Notice that CCA is applied to the time lags in a partialized manner and not to the instantaneous time step, i.e., the CCA is performed between time series sets $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^j$ partialized by the set $\mathbf{Y}_t \backslash \{\mathbf{Y}_{t-1}^j\}$. Furthermore, correlation may be interpreted as the square root of R^2 of a linear regression model, where R^2 represents the variance of the response variable which may be explained by the regressors, i.e., when partial CCA is applied between the past values of one subset and present values of another subset, it is identifying Granger causality between groups of time series.

In order to calculate ρ, set:

$$\mathbf{u} = \mathbf{a}' \mathbf{Y}_t^i, \qquad (3)$$

and

$$\mathbf{v} = \mathbf{b}' \mathbf{Y}_{t-1}^j, \qquad (4)$$

for some pair of coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. Then, we obtain

$$Var(\mathbf{u}) = \mathbf{a}' Cov(\mathbf{Y}_t^i) \mathbf{a} = \mathbf{a}' \mathbf{\Sigma}_{\mathbf{Y}_t^i \mathbf{Y}_t^i} \mathbf{a}, \qquad (5)$$

$$Var(\mathbf{v}) = \mathbf{b}' Cov(\mathbf{Y}_{t-1}^j) \mathbf{b} = \mathbf{b}' \mathbf{\Sigma}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j} \mathbf{b}, \qquad (6)$$

$$Cov(\mathbf{u}, \mathbf{v}) = \mathbf{a}' Cov(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^j) \mathbf{b} = \mathbf{a}' \mathbf{\Sigma}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j} \mathbf{b}. \qquad (7)$$

We shall seek coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ such that

$$Corr(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{a}' \mathbf{\Sigma}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j} \mathbf{b}}{\sqrt{\mathbf{a}' \mathbf{\Sigma}_{\mathbf{Y}_t^i \mathbf{Y}_t^i} \mathbf{a}} \sqrt{\mathbf{b}' \mathbf{\Sigma}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j} \mathbf{b}}} \qquad (8)$$

is maximized.

However, notice that the calculations performed above do not identify conditional Granger causalities (Granger causalities partialized by a third set of

6   *André Fujita*

time series), but, only, pairwise correlations between lagged sets of time series $(CCA(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^j))$, therefore, it is necessary to develop a partial canonical correlation analysis.

The crucial point in performing partial canonical correlation analysis [40] is to derive from variance-covariance matrices, linear coefficients for combining original variables into canonical variables. It turns out that the residualized variance-covariance matrices of $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^j$ after partialing out the effect of vector $\mathbf{X} = \mathbf{Y}_{t-1} \backslash \{\mathbf{Y}_{t-1}^j\}$ from both $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^j$, provide the solutions. More specifically, suppose that, when combined, the three vectors of variables, $\mathbf{Y}_t^i$, $\mathbf{Y}_{t-1}^j$ and $\mathbf{X}$ have the following partitioned variance-covariance matrix:

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y_t^i Y_{t-1}^j X}} = \begin{pmatrix} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{X}} \\ \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{X}} \\ \hat{\mathbf{\Sigma}}_{\mathbf{X} \mathbf{Y}_t^i} & \hat{\mathbf{\Sigma}}_{\mathbf{X} \mathbf{Y}_{t-1}^j} & \hat{\mathbf{\Sigma}}_{\mathbf{X X}} \end{pmatrix} \tag{9}$$

where the (r, s)th entry of $\hat{\mathbf{\Sigma}}_{\mathbf{UV}}$ is given by $(T-1)^{-1} \sum_{k=1}^T (\mathbf{U}_{kr} - \bar{\mathbf{U}}_r)(\mathbf{V}_{ks} - \bar{\mathbf{V}}_s)$ (T: time series' length).

The conditional variance-covariance matrix of $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^j$, partialing out the effect of $\mathbf{X}$ is given as [1] [25] [47]:

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y_t^i Y_{t-1}^j | X}} = \begin{pmatrix} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i | \mathbf{X}} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j | \mathbf{X}} \\ \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i | \mathbf{X}} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j | \mathbf{X}} \end{pmatrix}, \tag{10}$$

where

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i | \mathbf{X}} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i} - \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{XY}_t^i} \tag{11}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j | \mathbf{X}} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j} - \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{XY}_{t-1}^j} \tag{12}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i | \mathbf{X}} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i} - \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{XY}_t^i} \tag{13}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j | \mathbf{X}} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j} - \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{XY}_{t-1}^j} \tag{14}$$

Similar to regular canonical correlation analysis [25], the eigenvalues λ_d ($d = 1, \ldots, \min(m, n)$) from the following two matrices $\mathbf{A}$ and $\mathbf{B}$, will be the squared partial canonical correlation coefficients for the dth canonical functions and the eigenvectors $\mathbf{a}_d$ and $\mathbf{b}_d$ associated with the eigenvalue λ_d will be the linear coefficient vectors which combine the original variables into synthetic canonical variables:

$$\mathbf{A} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i | \mathbf{X}}^{-1/2} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j | \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j | \mathbf{X}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i | \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i | \mathbf{X}}^{-1/2} \tag{15}$$

$$\mathbf{B} = \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j | \mathbf{X}}^{-1/2} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_t^i | \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_t^i | \mathbf{X}}^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i \mathbf{Y}_{t-1}^j | \mathbf{X}} \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j \mathbf{Y}_{t-1}^j | \mathbf{X}}^{-1/2} \tag{16}$$

.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{\min(m,n)}$ be the ordered eigenvalues of matrices $\mathbf{A}$ and $\mathbf{B}$, therefore $CCA(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^j | \mathbf{Y}_t \backslash \{\mathbf{Y}_{t-1}^j\}) = \hat{\rho} = \sqrt{\lambda_1}$.

Again, as in regular canonical correlation analysis, the two matrices $\mathbf{A}$ and $\mathbf{B}$, have the same eigenvalues but with different eigenvectors, i.e, each eigenvector $\mathbf{b}_d$ is proportional to $\hat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{Y}^j_{t-1}\mathbf{Y}^j_{t-1}|\mathbf{X}}\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}^j_{t-1}\mathbf{Y}^i_t|\mathbf{X}}\hat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{Y}^i_t\mathbf{Y}^i_t|\mathbf{X}}\mathbf{a}_d$.

2.3. *Statistical test*

The hypothesis to be tested in order to verify the existence of Granger causality between *sets* of variables is as follows:

$H_0 : CCA(\mathbf{Y}^i_t, \mathbf{Y}^j_{t-1}|\mathbf{X}) = \rho = 0$ (Granger noncausality)

$H_1 : CCA(\mathbf{Y}^i_t, \mathbf{Y}^j_{t-1}|\mathbf{X}) = \rho \neq 0$ (Granger causality)

The bootstrap for our method is based on the block bootstrap [29]. It consists of splitting the data into blocks of observations (with overlapping) and sampling the blocks randomly with replacement. To describe this bootstrap more precisely, let the data consist of observations $\mathbf{Y}^i_t$ and $\mathbf{Y}^j_t$ $(t = 1, \ldots, T)$.

With overlapping blocks of length l, block 1 is observations $\mathbf{Y}_h : h = 1, \ldots, l$, block 2 is observations $\mathbf{Y}_{h+1} : h = 1, \ldots, l$, block 3 is observations $\mathbf{Y}_{h+2} : h = 1, \ldots, l$ and so forth. The bootstrap sample $\mathbf{Y}^{i*}_t$ is obtained by sampling blocks randomly with replacement from $\mathbf{Y}^i_t$ and laying them end-to-end in the order sampled. The bootstrap sample $\mathbf{Y}^{j*}_t$ is obtained in an analogous way. This block re-sampling is carried out in order to capture the dependence structure of neighbourhood obsevations, i.e., autocorrelation. Moreover, the resampling of all time series of $\mathbf{Y}^i_t$ is carried out together, in order to capture contemporaneous correlations between time series. The same is performed for the time series of $\mathbf{Y}^j_t$. However, $\mathbf{Y}^i_t$ and $\mathbf{Y}^j_t$ are re-sampled independently, in order to break the relationship between the response and predictor variables.

After constructing the bootstrap samples $\mathbf{Y}^{i*}_t$ and $\mathbf{Y}^{j*}_t$, calculate $CCA(\mathbf{Y}^{i*}_t, \mathbf{Y}^{j*}_{t-1}|\mathbf{X}^*) = \hat{\rho}^*$, where $\mathbf{X}^* = \mathbf{Y}^*_{t-1}\backslash\{\mathbf{Y}^{j*}_{t-1}\}$. Repeat these steps until obtaining the desired number of bootstraps. Then, use the empirical distribution of ρ^* to test whether $\hat{\rho} = 0$ (the bootstrap scheme is illustrated in Figure 1).

Regardless of the block bootstrap that is used, the block length l must increase with increasing time series length T to render bootstrap estimators of moments and distribution functions consistent [7] [28] [19]. Similarly, the block length must increase with increasing sample size to enable the block bootstrap to achieve asymptotically correct coverage probabilities for confidence intervals and rejection probabilities for hypothesis tests. For the special case of an autoregressive process of order one, [7] showed that the block length l that minimizes the asymptotic mean-square error of the variance estimator increases at the rate of $l \propto T^{\frac{1}{3}}$. Since time series gene expression data are generally short, it is unfeasible to fit an AR model of higher orders. However, if a longer time series data becomes available, one may use the algorithm proposed by [6] in order to select the block length for the bootstrap procedure.
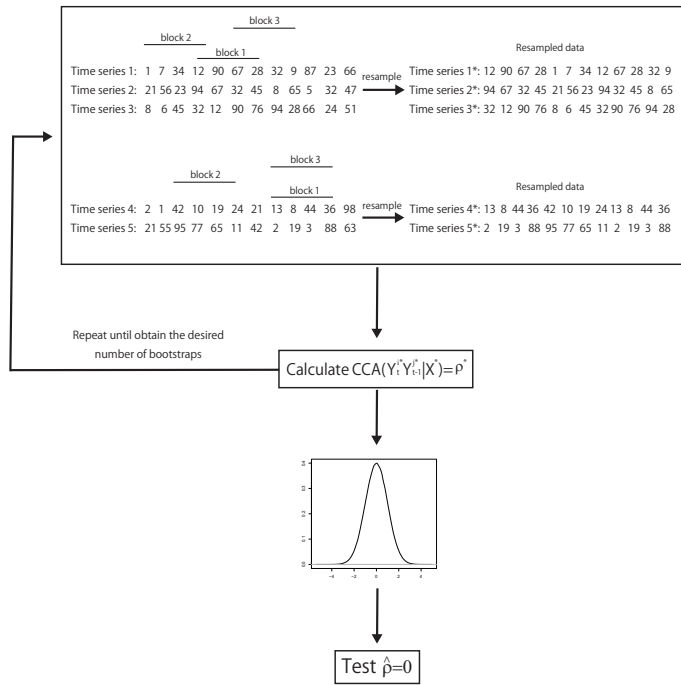
8   *André Fujita*

Fig. 1. Illustrative diagram of the bootstrap. Time series 1, 2 and 3 belongs to the set $\mathbf{Y}_t^i$ while time series 4 and 5 belong to the $\mathbf{Y}_t^j$ set. Notice that the blocks may be overlapped.

2.4. *Computational simulation*

Before analyzing actual gene expression data, we conducted Monte Carlo simulations to examine whether our approach is able to identify Granger causality between sets of time series and effectively control the rate of false positives.

The artificial network is set as:

$$\text{simulation 1} \begin{cases} Z^{\mathrm{I}}_{1,t} = \beta \times Z^{\mathrm{I}}_{1,t-1} - \beta \times Z^{\mathrm{I}}_{4,t-1} + \varepsilon_{1,t} \\ Z^{\mathrm{I}}_{2,t} = \beta \times Z^{\mathrm{I}}_{1,t-1} + \varepsilon_{2,t} \\ Z^{\mathrm{I}}_{3,t} = \beta \times Z^{\mathrm{I}}_{1,t-1} - \beta \times Z^{\mathrm{I}}_{5,t-1} + \varepsilon_{3,t} \\ Z^{\mathrm{I}}_{4,t} = \beta \times Z^{\mathrm{I}}_{2,t-1} + \varepsilon_{4,t} \\ Z^{\mathrm{I}}_{5,t} = \varepsilon_{5,t} \\ Z^{\mathrm{II}}_{6,t} = \beta \times Z^{\mathrm{II}}_{8,t-1} + \varepsilon^{\mathrm{II}}_{6,t} \\ Z^{\mathrm{II}}_{7,t} = \beta \times Z^{\mathrm{I}}_{3,t-1} - \beta \times Z^{\mathrm{II}}_{6,t-1} + \varepsilon_{7,t} \\ Z^{\mathrm{II}}_{8,t} = \beta \times Z^{\mathrm{II}}_{10,t-1} + \varepsilon_{8,t} \\ Z^{\mathrm{II}}_{9,t} = \beta \times Z^{\mathrm{I}}_{5,t-1} - \beta \times Z^{\mathrm{II}}_{7,t-1} + \beta \times Z^{\mathrm{III}}_{14,t-1} + \varepsilon_{9,t} \\ Z^{\mathrm{II}}_{10,t} = \beta \times Z^{\mathrm{II}}_{9,t-1} - \beta \times Z^{\mathrm{III}}_{13,t-1} + \varepsilon_{10,t} \\ Z^{\mathrm{III}}_{11,t} = \beta \times Z^{\mathrm{II}}_{8,t-1} + \varepsilon_{11,t} \\ Z^{\mathrm{III}}_{12,t} = \varepsilon_{12,t} \\ Z^{\mathrm{III}}_{13,t} = \varepsilon_{13,t} \\ Z^{\mathrm{III}}_{14,t} = \varepsilon_{14,t} \end{cases}$$

for $t = 1, \ldots, T$, where the noises $\varepsilon_{i,t}$ $(i = 1, \ldots, 14)$ are normally distributed with mean of zero and variance of one and I, II and III represent the sets to which each time series $Z_{i,t}$ belong. The time series length is set to 100, i.e., $T = 100$, and β is set to 0.4.

In order to evaluate the power in identifying Granger causalities between sets of time series, a comparison between the proposed method based on CCA and the standard VAR model was performed. The following artificial network is used for the comparison:

$$\text{simulation 2} \begin{cases} W^{\mathrm{I}}_{1,t} = \gamma \times W^{\mathrm{I}}_{1,t-1} + \epsilon_{1,t} \\ W^{\mathrm{I}}_{2,t} = \gamma \times W^{\mathrm{I}}_{1,t-1} - \gamma \times W^{\mathrm{I}}_{5,t-1} + \epsilon_{2,t} \\ W^{\mathrm{I}}_{3,t} = \gamma \times W^{\mathrm{I}}_{1,t-1} - \gamma \times W^{\mathrm{I}}_{4,t-1} + \epsilon_{3,t} \\ W^{\mathrm{I}}_{4,t} = \gamma \times W^{\mathrm{I}}_{2,t-1} - \gamma \times W^{\mathrm{I}}_{4,t-1} + \epsilon_{4,t} \\ W^{\mathrm{I}}_{5,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} - \gamma \times W^{\mathrm{III}}_{13,t-1} + \epsilon_{5,t} \\ W^{\mathrm{II}}_{6,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} + \epsilon_{6,t} \\ W^{\mathrm{II}}_{7,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} - \gamma \times W^{\mathrm{II}}_{6,t-1} + \epsilon_{7,t} \\ W^{\mathrm{II}}_{8,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} - \gamma \times W^{\mathrm{II}}_{7,t-1} + \epsilon_{8,t} \\ W^{\mathrm{II}}_{9,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} - \gamma \times W^{\mathrm{II}}_{8,t-1} + \epsilon_{9,t} \\ W^{\mathrm{II}}_{10,t} = \gamma \times W^{\mathrm{I}}_{3,t-1} - \gamma \times W^{\mathrm{II}}_{9,t-1} + \gamma \times W^{\mathrm{III}}_{12,t-1} + \epsilon_{10,t} \\ W^{\mathrm{III}}_{11,t} = \gamma \times W^{\mathrm{III}}_{11,t-1} - \gamma \times W^{\mathrm{III}}_{13,t-1} + \epsilon_{11,t} \\ W^{\mathrm{III}}_{12,t} = \gamma \times W^{\mathrm{III}}_{11,t-1} + \epsilon_{12,t} \\ W^{\mathrm{III}}_{13,t} = \gamma \times W^{\mathrm{III}}_{12,t-1} + \epsilon_{13,t} \end{cases}$$

for $t = 1, \ldots, T$, where the noises $\epsilon_{i,t}$ $(i = 1, \ldots, 13)$ are normally distributed with mean of zero and variance of one and I, II and III represent the sets which each time series $W_{i,t}$ belongs to. The time series length is set to 100, i.e., $T = 100$, and γ is set to 0.2.

For each simulation, 1 and 2, 10,000 networks were generated. For each artificial network, the proposed approach was applied to identify Granger causalities between

Table 1. Simulation 1. Number of Granger causalities obtained using our approach in 10,000 simulations. The rate of false positives was controlled in 5%, i.e., it is expected to obtain ~ 500 false-positives where there is no Granger causality. In bold, are the Granger causalities which actually exist.

from/to	I	II	III
I	**10,000**	**8,543**	507
II	461	**9,979**	**6,646**
III	500	**8,015**	437

the sets of variables. The threshold to discriminate the existence of an edge was $p < 0.05$, i.e., the false positives rate was controlled to 5% and the number of bootstraps was set to 1,000. For the standard VAR model, the Wald's test was used in order to identify and test Granger causality between sets. In other words, suppose three sets I, II and III as in simulation 2. The coefficients of the VAR model related to the time series of set I to II were simultaneously tested in order to verify whether they are equal to zero or not. The same was performed with the coefficients I $\rightarrow$ I, I $\rightarrow$ III, II $\rightarrow$ I, II $\rightarrow$ II, III $\rightarrow$ I, III $\rightarrow$ II and III $\rightarrow$ III.

3. Results and Discussions

3.1. *Simulation*

In order to study the properties of our method to identify Granger causalities between sets of genes, we have designed a simulation study which contains 14 artificial genes in a time series of length $T = 100$ (simulation 1).

Figure 2A illustrates an example of time series obtained by our simulation study. Figure 2B illustrates the structure of the artificial regulatory network and the interaction between sets (I, II and III). Figure 2C represents the true model, i.e., if our approach is working correctly, it must identify the structure illustrated in Figure 2C.

Notice in Table 1 that where there is no edge, the number of false-positives is actually controlled to 5% (~ 500). In bold, the edges which actually are present in the network are illustrated. Notice that, as expected, where there is an edge, the number of identifications is higher than where there is no edge. The number of false-positives is actually controlled since the p-value distribution under the null hypothesis is close to a uniform distribution (Figure 3), demonstrating that the bootstrap procedure is working adequately.

Interpreting the loops (auto-regressions) from I to I and from II to II, they are representing the total information flow contained in networks I and II, respectively. Another interpretation for the loops may be the density of the network, since this density may represent the *sum* of the information flow. On the other hand, no Granger causality is found in set III, therefore, no loop is present. It is necessary
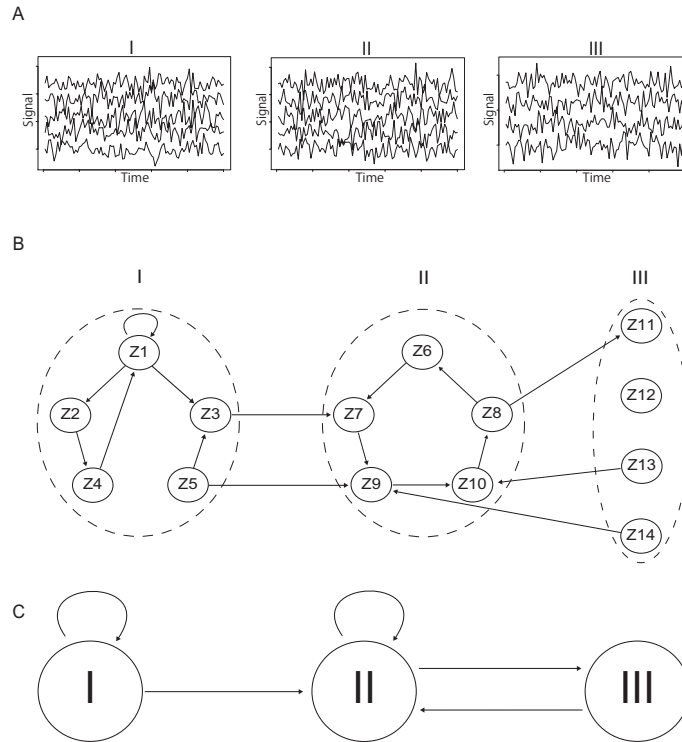
Fig. 2. Simulation 1. (A) Time series data for each simulated variable; (B) Time series Z_i ($i =$ $1, \ldots, 14$) represent variables. I, II and III are the sets of time series composed by Z_i. Arrows represent the Granger causalities; (C) Granger causality between sets of time series.

to point out that these loops are exactly the $CCA(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^i | \mathbf{Y}_t \backslash \{\mathbf{Y}_{t-1}^i\})$, i.e., the total connectivity inside set $\mathbf{Y}^i$, where $i=$I, II or III. Notice that CCA is always applied between the past of one subset and the present of another subset of the time series.

Figure 4A illustrates the structure of an artificial network and the interaction between sets of time series (Figure 4B) described in simulation 2. In this simulation 2, both approaches, based on CCA and blockwise VAR proposed by [50] were applied in order to evaluate the power of the proposed method. The blockwise VAR model was applied using all the variables, i.e., a large network was constructed using VAR. The test between sets for VAR were performed by testing the estimated coefficients between sets using the Wald's test (whether all of them are equal to zero simultaneously). The Wald's test is equivalent to test the ratio of the prediction's error variance proposed by [17]. Notice, in Table 2 that, again, the false-positives are actually controlled to 5%. In bold, are the number of times Granger causality was identified in a total of 10,000 simulations. Analyzing the number of Granger causalitites identified by both methods, the method based on CCA is more powerful
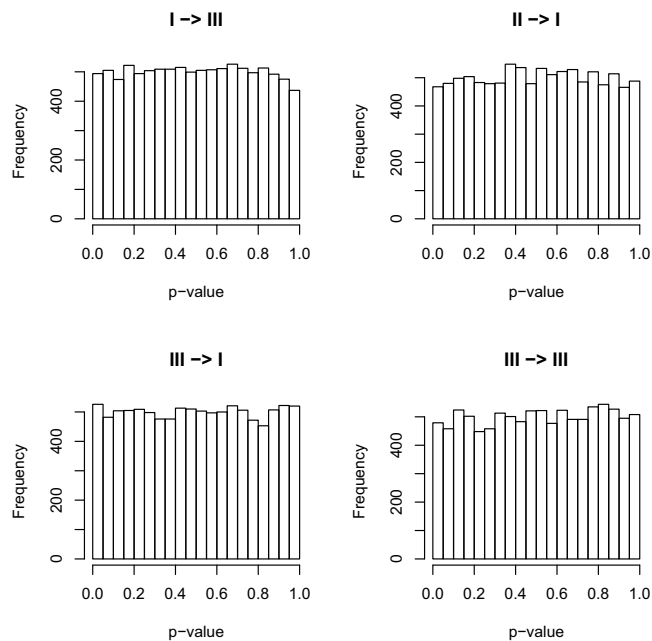
Fig. 3. P-value distribution under the null hypothesis. Notice that the p-value distributions are close to a uniform distribution, confirming the consistency of the proposed bootstrap approach.

Table 2. Simulation 2. Number of Granger causalities obtained using our approach in 10,000 simulations. In parenthesis are the number of Granger causalities obtained by standard VAR. The rate of false positives was controlled in 5%, i.e., it is expected to obtain ~ 500 false-positives where there is no Granger causality. In bold, are the Granger causalities which actually exist.

from/to	I	II	III
I	**7,154 (1,079)**	**10,000 (9,980)**	505 (589)
II	466 (547)	**10,000 (9,997)**	463 (522)
III	**1,490 (1,015)**	**10,000 (9,827)**	**6,263 (2,107)**

than the standard VAR. This is due to the CCA characteristic which maximizes the correlation by calculating the optimum linear combination of both sets while the blockwise VAR sets equal weights to each time series. In a biological sense, time series W_3 (Figure 4) may be interpreted as a transcription factor which Granger causes several other genes. The method based on CCA is also clearly more powerful in the case of auto loops (from I to I, from II to II and from III to III).
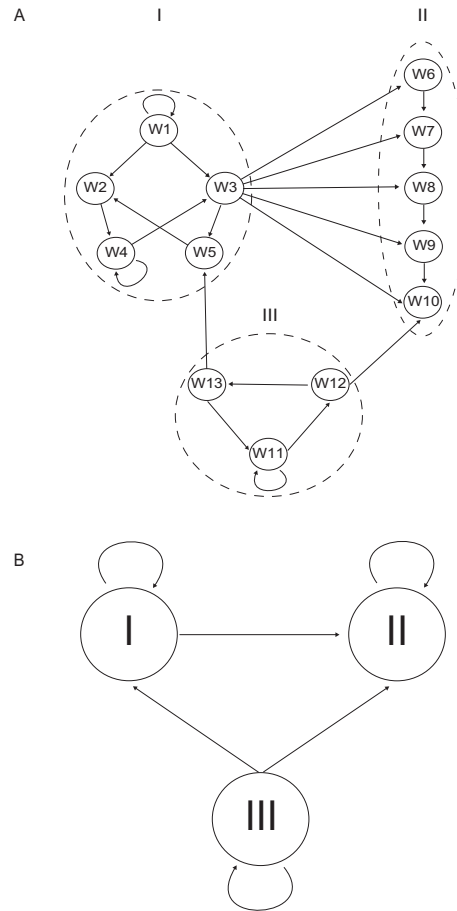
Fig. 4. Simulation 2. (A) The time series W_i $(i = 1, \ldots, 13)$ represent the variables. I, II and III are the sets of time series composed by W_i. Arrows represent the Granger causalities; (B) Granger causality between sets of time series.

3.2. *Biological data*

In order to illustrate a practical application of the proposed model, an actual biological regulatory network was modeled. The following 15 genes, namely, RECK, SRC, C-MYC, THBS1, TP53, CDKN1A, GADD45A, FGFRL1, FGF2, FGFR2, FGFR1, IKBKG, CHUK, NFKBIA and NFKB1, were selected, the same which were used in our previous works [11] [12] [13]. Further details on these genes may be found in Table 3 and their expression profile shown in Figure 5. The proposed approach was applied to the analysis of HeLa cell cycle gene expression data collected by [51]. HeLa is an immortal cell line, derived from a human epithelial cervical carcinoma. Its cell cycle is of approximately 16 hours. The gene expression data used in this

14   *André Fujita*

Table 3. List of the 15 genes and their respective IMAGE IDs used in the construction of the biological gene network.

IMAGE ID	Gene name
IMAGE:324383	FGF2 fibroblast growth factor 2 (basic)
IMAGE:809464	FGFR2 fibroblast growth factor receptor 2
IMAGE:837908	FGFRL1 fibroblast growth factor receptor-like 1
IMAGE:711857	FGFR1 fibroblast growth factor receptor 1
IMAGE:591683	GADD45A growth arrest and DNA-damage-inducible
IMAGE:268652	CDKN1A cyclin-dependent kinase inhibitor 1A (p21, Cip1)
IMAGE:24415	TP53 tumor protein p53
IMAGE:417226	MYC v-myc avian myelocytomatosis viral oncogene homolog
IMAGE:133096	RECK reversion-inducing-cysteine-rich protein with kazal motifs
IMAGE:2107602	SRC v-src avian sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog
IMAGE:429042	THBS1 thrombospondin 1
IMAGE:488499	CHUK conserved helix-loop-helix ubiquitous kinase
IMAGE:789357	NFKB1 nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
IMAGE:41072	IKBKG inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
IMAGE:340734	NFKBIA nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor

study contains three complete cell cycles, i.e., 48 time points distributed at intervals of one hour. The HeLa gene expression data is freely available at: http://genome-www.stanford.edu/Human-CellCycle/HeLa/.
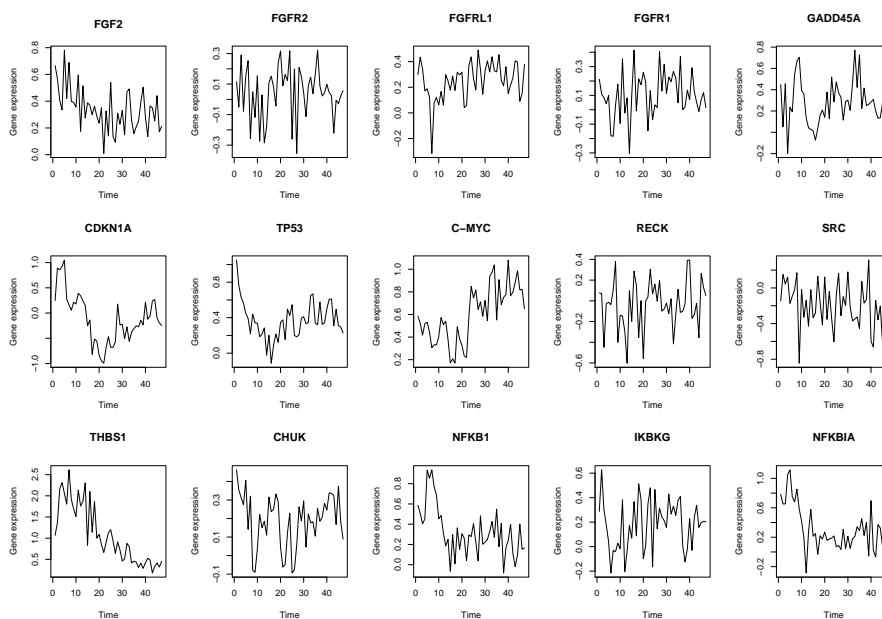

Fig. 5. Gene expression kinetics profile of the 15 genes analysed.

The sets of genes were composed as follows: (I) RECK, SRC, C-MYC and THBS1; (II) TP53, CDKN1A and GADD45A; (III) FGFRL1, FGF2, FGFR2 and FGFR1 and (IV) IKBKG, CHUK, NFKBIA and NFKB1. The networks illustrated

inside each cluster (Figure 6) were obtained by applying the standard VAR model. The edge links between the clusters were obtained by the method proposed here. Unfortunately, the auto-loops in sets III and IV were not identified probably due to the power of the test. Increasing time series length may increase the power of the test and, consequently, identifying the auto-loops. Notice that in the simulations, the proposed approach was able to correctly identify the Granger causalities.
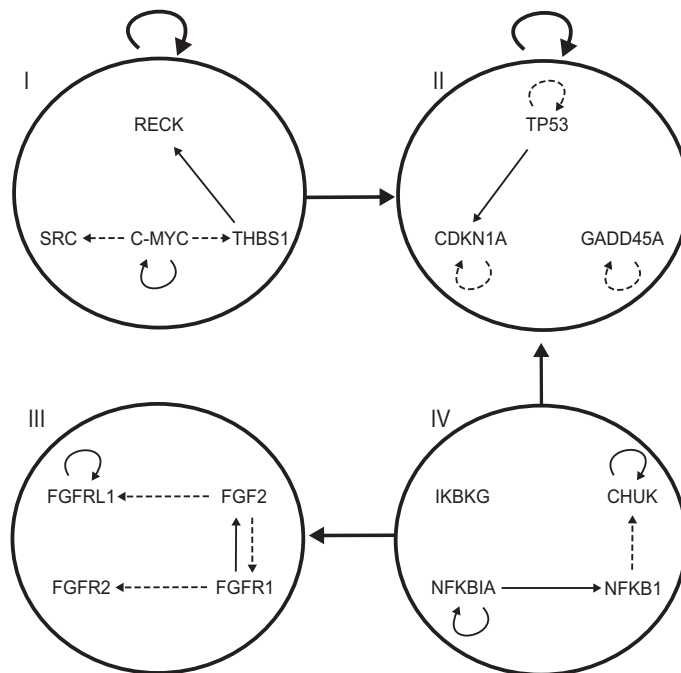

Fig. 6. Actual biological network. Identification of Granger causality to a network composed by I: RECK, SRC, c-Myc, THBS1; II: TP53, CDKN1A, GADD45A; III: FGFRL1, FGF2, FGFR2, FGFR1 and IV: IKBKG, CHUK, NFKBIA and NFKB1 genes. Connectivities inside each set were identified by using the standard VAR model. Dashed arrows are significant connectivities with $p < 0.1$ and full arrows are significant connectivities with $p < 0.05$.

The tumor progression processes, as other pathological and physiological events, are regulated by responses to changes in the external environment, and by following coordinated modulation of gene expression in order to appropriately adapt to these changes. Several molecules are responsible for controlling cell behavior, among which, growth and transcriptional factors display prominent functions in the context of tumorigenesis. The networks of FGF, C-MYC, TP53 and NFKB1, used in this work to demonstrate a biological application of the proposed model, are able to control almost all processes involved in development and tumor progression. Proliferation and differentiation, cell survival and death, cell division and cycle arrest,

cell migration, adhesion and invasion are just some examples of cellular phenotypes modulated by these molecules [23]. Therefore, the choice of these sets of genes is appropriate for validating the biological application of the proposed model. Thus, prediction of the interaction between these important gene clusters could elucidate the complex relationship among them and, consequently, allow construction of hypotheses about how the interaction between these pathways orchestrate the regulation of processes that are often mutually exclusive.

Several members of the fibroblast growth factors (FGFs) family and their respective specific cell surface receptors (FGFRs) play important roles in a variety of normal and pathological processes, including tumor transformation. In the tumor context, FGFs are well known for promoting tumor angiogenesis and invasiveness, and for acting as mitogenic agents and cell death inhibitors in many cell types [26]. However, recent reports have shown that some of these molecules can also promote tumor suppression, depending on the cell type and the number and strength of stimuli. FGF2, a gene belonging to set III, is also implied in restoring tumor defense mechanisms in malignant cells [10].

By controlling a variety of downstream target genes, the p53 transcription factor has many tumor-suppressor activities. The p53 protein is modified in more than 50% of human tumors. Thus, many processes are subverted due to p53-alteration, such as, for instance: cell cycle arrest, inhibition of metastasis and angiogenesis, DNA repair and cell death induction [24]. On the other hand, aberrant c-myc activity is associated with genomic instability and tumor progression. This transcriptional factor is encoded by a proto-oncogene, whose product modulates transcriptional processes during normal cell growth and proliferation [49]. However, in the tumoral context, deregulated c-myc activity is a classic molecular marker of poor prognostics in several types of cancers. Moreover, the c-Myc protein is able to regulate other important events involved in tumor transformation, including hormone dependence, invasiveness and metastatic potential [45]. In all of these processes, c-Myc positively associates with tumor progression. Like c-Myc, the NFKB1 family of transcription factors has been shown to be constitutively activated in various human malignancies, namely: leukemias, lymphomas, and a number of solid tumors. As the other three pathways analyzed in this work, NFKB1 is able to control a variety of events by regulating the expression of genes involved in angiogenesis, metastasis, cell growth, proliferation and apoptosis. Therefore, NFKB1 has been implicated in transcriptional upregulation of several growth factors, cytokines, adhesion molecules, antiapoptotic proteins and oncogene products [9].

Here, we were able to identify the Granger causality between the previously summoned networks. We predicted that the set of genes (I) Granger-causes (II), and that (IV) Granger-causes (II) and (III). Therefore, by applying our model, we predicted that the c-Myc network is able to induct p53 related genes, and that the NFKB1 pathway positively regulates the FGF and p53 sets of genes (Figure 5). By applying the standard VAR model and jointly testing the coefficients [17], it was only possible to identify Granger causality from clusters IV to II, confirming the

simulation's results (simulation II) that the power of the proposed method in actual biological data is superior to the traditional one.

Previous reports have demonstrated that p53-induced G1/S cell-cycle arrest is attributed mainly to its ability to the transcriptionally upregulate CDKN1A and repress c-myc [5]. Induction of CDKN1A by p53 is in accordance with the predicted connectivity previously reported [11] [12] [13]. However, the well established p53, acting as a c-myc inhibitor contradicts the identified Granger causality. On the other hand, there are other described mechanisms demonstrating that c-myc is able to indirectly induce p53 through *Arf* regulation. Thus, p14$^{\text{ARF}}$ (the *Arf* product in humans) inhibits Mdm2 and, consequently, stabilizes p53 [3]. Therefore, this last mechanism demonstrates that the anticipated relationship of p53 induction by c-myc, predicted by our model, is biologically relevant. Several reports linking NFKB1 and p53 networks are available. As mentioned for the previous sets of genes, these transcriptional factors can reciprocally regulate each other. NFKB1 has been reported to induce Mdm2, consequently inhibiting p53 [46]. In contrast, NFKB1 has also been described as an inducer of the p53 promoter [20]. Thus, the positive modulation of the p53 pathway by NFKB1 set of genes, determined by our model, corroborates previously described mechanisms. Moreover, these results suggest the occurrence of a synergistic mechanism to induce p53, in view of the fact that both c-myc and NFKB1 networks are involved in up-regulation of this transcriptional factor.

By examining the sign and the magnitude of the canonical weight assigned to each variable, i.e., the eigenvectors **a** and **b** (see Methods section), it is possible to verify that variables with larger weights contribute more to the variables. Moreover, variables whose weights have opposite signs exhibit an inverse relationship with each other, and variables with weights of the same sign exhibit a direct relationship.

4. Conclusions

Comparison between this method here proposed and other approaches such as Bayesian networks, is difficult, since, by definition, they analyze/identify different kinds of causality. Moreover, although our previous reports [11] [12] [13] [14] identify time-varying, linear or nonlinear Granger causalities, they are limited to the usual "between time series (genes)". The method we propose here generalizes the standard Granger causality to relationships between *sets* of time series, which, to the best of our knowledge, is the first to address this issue.

Ordinarily, when the number of genes is relatively large, interpreting the pairwise Granger causalities is hopeless. Moreover, linear combinations of variables are often interesting and useful for predictive or comparative purposes. In addition, pairwise Granger causality does not distinguish direct from indirect causalities because it is not conditioned by a third set of time series. The purpose of the definition we present for Granger causality is to summarize the associations between the $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$ sets in terms of *few* carefully chosen covariances. This summarized measure illustrates the strength of Granger causality between sets $\mathbf{Y}_t^i$ and $\mathbf{Y}_t^j$. This may be

useful to understand the complete information flow from one network or pathway to the other, mainly in regulatory networks, where the number of genes is high and interpreting gene-gene Granger causalities may be cumbersome. When modeling gene-gene regulatory networks, generally a selection of the probe which will be used to represent a certain gene is carried out, or an average between the probes is calculated. However, this problem can be overcome in a straightforward manner with the method proposed here, since one gene can be represented by the set of all probes. Linking the concept of Granger causality for sets of time series to graph theory, some graph theoretical properties, such as sink and source can be generalized to node sets. Moreover, hub and centrality for sets of genes can be defined based on its total information flow. Another application is in annotation, i.e., when the functionality of a set of genes is unknown, but this set is Granger caused by another set of genes which is well studied. Therefore, this information may be useful to infer or construct some hypothesis about the unknown set of genes.

Here, we have assumed that the sets of time series are given. However, depending on the problem, it is necessary to identify which time series belongs to each set. To this purpose, biological *a priori* information may be used to define the different gene sets. Thus, for example, if it is known that one set of genes respond to a specific drug and there is another set of genes which respond to the same drug, but no literature information is available about the existence of a pathway between these two sets, one may use our approach in order to obtain some clues about the existence (or not) of this pathway. Another application may be the search for crosstalks between under-studied pathways. On the other hand, an objective and systematic method based on clustering may be used. The latter is the object of our future studies.

Acknowledgments

References

1. Anderson TW, *An introduction to multivariate statistical analysis*, Wiley, New York, 1984.
2. Baccala L, Sameshima K, Partial directed coherence: a new concept in neural structure determination, *Biological Cybernetics* **84**:463–474, 2001.
3. Baudino TA, Maclean KH, Brennan J, Parganas E, Yang C, Aslanian A, Lees JA, Sherr CJ, Roussel MF, Cleveland JL, Myc-mediated proliferation and lymphomagenesis, but not apoptosis, are compromised by E2f1 loss, *Molecular Cell* **11**:905–914, 2003.
4. Brovelli A, Ding M, Ledberg A, Chen Y, Nakamura R, Bressler SL, Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality, *PNAS* **101**:9849–9854, 2004.
5. Brown L, Boswell S, Raj L, Lee SW, Transcriptional targets of p53 that regulate cellular proliferation, *Critical reviews in eukaryotic gene expression* **17**:71–85, 2007.
6. Bühlmann P, Künsch R, Block length selection in the bootstrap for time series, *Computational Statistics & Data Analysis* **31**:295–310, 1999.

7. Carlstein E, The use of subseries methods for estimating the variance of a general statistic from a stationary time series, *Annals of Statistics* **14**:1171–1179, 1986.

8. Chen Y, Rangarajan G, Feng J, Ding M, Analyzing multiple nonlinear time series with extended Granger causality, *Physics Letters A*, **324**:26–35, 2004.

9. Chen F, Beezhold K, Castranova V, Tumor promoting or tumor suppressing of NF-kappa B, a matter of cell context dependency, *International reviews of immunology*, **27**:183–204, 2008.

10. Costa ET, Forti FL, Matos TG, Dermargos A, Nakano F, Salotti J, Rocha KM, Asprino PF, Yoshihara CK, Koga MM, Armelin HA, Fibroblast growth factor 2 restrains Ras-driven proliferation of malignant cells by triggering RhoA-mediated senescence, *Cancer Research*, **68**:6215–6223, 2008.

11. Fujita A, Sato JR, Garay-Malpartida HM, Morettin PA, Sogayar MC, Ferreira CE, Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method, *Bioinformatics*, **23**:1623–1630, 2007.

12. Fujita A, Sato JR, Garay-Malpartida HM, Yamaguchi R, Miyano S, Sogayar MC, Ferreira CE. Modeling gene expression regulatory networks with the sparse vector autoregressive model, *BMC Systems Biology*, **1**:39, 2007.

13. Fujita A, Sato JR, Garay-Malpartida HM, Sogayar MC, Ferreira CE, Miyano S, Modeling nonlinear gene regulatory networks from time series gene expression data, *Journal of Bioinformatics and Computational Biology*, **6**:961–979, 2008.

14. Fujita A, Patriota AG, Sato JR, Miyano S. The impact of measurement errors in the identification of regulatory networks, *BMC Bioinformatics*, **10**:412, 2009.

15. Guo X, Tatsuoka K, Liu R, Histone acetylation and transcriptional regulation in the genome of Saccharomyces cerevisiae, *Bioinformatics*, **2**:392-399, 2006.

16. Guo S, Wu J, Ling M, Feng J, Uncovering interactions in the frequency domain, *PLoS Computational Biology*, **4**:e1000087, 2008.

17. Geweke J, Measures of conditional linear dependence and feedback between time series, *J Am Stat Assoc*, **79**:907–915, 1984.

18. Granger CWJ, Investigating causal relationships by econometric models and cross-spectral methods, *Econometrica*, **37**:424–438, 1969.

19. Hall P, Resampling a coverage process, *Stochastic Process Applications*, **19**:259–269, 1985.

20. Hellin AC, Calmant P, Gielen J, Bours V, Merville MP, Nuclear factor - kappaB-dependent regulation of p53 gene expression induced by daunomycin genotoxic drug, *Oncogene*, **16**:1187–1195, 1998.

21. Hotelling H, The most predictable criterion, *J. Educ. Psychol.*, **26**:139–142, 1935.

22. Hotelling H, Relations between two sets of variates, *Biometrika*, **28**:321–377, 1936.

23. Huang YH, Wu JY, Zhang Y, Wu MX, Synergistic and opposing regulation of the stress-responsive gene IEX-1 by p53, c-Myc, and multiple NF-kB/rel complexes, *Oncogene*, **21**:6819-6828, 2002.

24. Jin S, Levine A, The p53 functional circuit, *J. Cell Sci.*, **114**:4139–4140, 2001.

25. Johnson RA, Wichern DW, *Applied multivariate statistical analysis*, Prentice Hall, New Jersey, 2002.

26. Kwabi-Addo B, Ozen M, Ittmann M, The role of fibroblast growth factors and their receptors in prostate cancer, *Endocrine-related cancer*, **11**:709–724, 2008.

27. Kojima K, Fujita A, Shimamura T, Imoto S, Miyano S, Estimation of nonlinear gene regulatory networks via L1 regularized NVAR from time series gene expression data, *Genome Informatics*, **20**:37–51, 2008.

28. Künsch HR, The jacknife and the bootstrap for general stationary observations, *Annals of Statistics*, **17**:1217–1241, 1989.

29. Lahiri SN, *Resampling methods for dependent data*, Springer-Verlag, New York, 2003.
30. Lozano AC, Abe N, Liu Y, Rosset S, Grouped graphical Granger modeling for gene expression regulatory networks discovery, *Bioinformatics*, **25**:i110-i118, 2009.
31. Lütkepohl H, *New introduction to multiple time series analysis*, Springer-Verlag, New York, 2006.
32. Marinazzo D, Pellicoro M, Stramaglia S, Kernel Granger causality and the analysis of dynamical networks, *Physical Review E*, **77**:056215, 2008.
33. Marinazzo D, Pellicoro M, Stramaglia S, Kernel method for nonlinear Granger causality, *Physical Review Letters*, **100**:144103, 2008.
34. McCracken MW, Asymptotics for out of sample tests of Granger causality, *Journal of Econometrics*, **140**:719–752, 2007.
35. McCrorie JR, Chambers MJ, Granger causality and the sampling of economic process, *Journal of Econometrics*, **132**:311–336, 2006.
36. Mukhopadhyay ND, Chatterjee S, Causality and pathway search in microarray time series experiment, *Bioinformatics*, **23**:442–449, 2007.
37. Nagarajan R, Upreti M, Comment on causality and pathway search in microarray time series experiment, *Bioinformatics*, **24**:1029–1032, 2008.
38. Nagarajan R, A note on inferring acyclic network structures using Granger causality tests, *The International Journal of Biostatistics*, **5**:10, 2009.
39. Opgen-Rhein R, Strimmer K, Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process, *BMC Bioinformatics*, **8**:S3, 2008.
40. Rao BR, Partial canonical correlations, *Trabajos de Estadistica y de Investigacion operative*, **20**:211–219, 1969.
41. Revell LJ, Harrison AS, PCCA: a program for phylogenetic canonical correlation analysis, *Bioinformatics*, **24**:1018–20, 2008.
42. Roebroeck A, Formisano E, Goebel R, Mapping directed influence over the brain using Granger causality and fMRI, *Neuroimage*, **25**:230–242, 2006.
43. Sato JR, Amaro Jr E, Takahashi DY, de Maria Felix M, Brammer MJ, Morettin PA, A method to produce evolving functional connectivity maps during the course of an fMRI experiment using wavelet-based time-varying Granger causality, *Neuroimage*, **31**:187–196, 2006.
44. Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M, Miyano S, Recursive regularization for inferring gene networks from time-course gene expression profiles, *BMC Systems Biology*, **3**:41, 2009.
45. Shiu RP, Watson PH, Dubik D, c-myc oncogene expression in estrogen-dependent and -independent breast cancer, *Clinical Chemistry*, **39**:353–355, 1993.
46. Tergaonkar V, Pando M, Vafa O, Wahl G, Verma I, p53 stabilization is decreased upon NFKB1 activation: A role for NFKB1 in acquisition of resistance to chemotherapy, *Cancer Cell*, **1**:493–503, 2002.
47. Timm NH, *Multivariate analysis with application in education and psychology*, Brooks/Cole, Monterey, 1975.
48. Waaijenborg S, Zwinderman AH, Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers, *BMC Proceedings*, **1**:S122, 2007.
49. Wade M, Wahl GM, c-Myc, genome instability, and tumorigenesis: the devil is in the details, *Curr Top Microbiol Immunol*, **302**:169–203, 2006.
50. Wang X, Chen Y, Bressler SL, Ding M, Granger causality between multiple interdependent neurobiological time series: blockwise versus pairwise methods, *International Journal of Neural Systems*, **17**:71–78, 2006.

51. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Molecular Biology of the Cell*, **13**:1977–2000, 2002.

52. Wiener N, *The theory of prediction. In E.F. Beckenbach, editor, Modern Mathematics for Engineers*, McGraw-Hill, New York, 1956.

53. Wong T, Granger causality tests among openness to international trade, human accumulation and economic growth in China: 1952-1999, *International Economic Journal*, **20**:285–302, 2006.

54. Yamanishi Y, Vert J-P, Nakaya A, Kanehisa M, Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, *Bioinformatics*, **19**:i323–i330, 2003.