BMC Systems Biology

METHODOLOGY ARTICLE Open Access

# Functional clustering of time series gene expression data by Granger causality

André Fujita[1*], Patricia Severino[2], Kaname Kojima[3], João Ricardo Sato[4], Alexandre Galvão Patriota[1] and Satoru Miyano[3]

## Abstract

**Background:** A common approach for time series gene expression data analysis includes the clustering of genes with similar expression patterns throughout time. Clustered gene expression profiles point to the joint contribution of groups of genes to a particular cellular process. However, since genes belong to intricate networks, other features, besides comparable expression patterns, should provide additional information for the identification of functionally similar genes.

**Results:** In this study we perform gene clustering through the identification of Granger causality between and within sets of time series gene expression data. Granger causality is based on the idea that the cause of an event cannot come after its consequence.

**Conclusions:** This kind of analysis can be used as a complementary approach for functional clustering, wherein genes would be clustered not solely based on their expression similarity but on their topological proximity built according to the intensity of Granger causality among them.

## Background

Gene network analysis of complex datasets, such as DNA microarray results, aims to identify relevant structures that help the understanding of a certain phenotype or condition. These networks comprise hundreds to thousands of genes that may interact generating intricate structures. Consequently, pinpointing genes or sets of genes that play a crucial role becomes a complicated task.

Common analyses explore gene-gene level relationships and generate broad networks. Although this is a valuable approach, genes might interact more intensely to a few members of the network, and the identification of these so-called sub-networks should lead to a better comprehension of the entire regulatory process.

Several *in silico* methodologies are available for the identification of sub-networks, or clusters, within a given dataset [1-5]. Most of the times, the identified clusters group genes based on similar patterns of expression in time. In a different manner, the identification of Granger causality [6] within a network allows the clustering of genes based on their topological proximity in the network [7,8]. Briefly, Granger causality [6] analysis identifies interaction in terms of temporal precedence (the cause comes before its effect) [6] and may generate a set of sub-networks within which Granger causality is intense among genes. As a result, genes are grouped depending on how close they are in terms of Granger causality. Figure 1a illustrates the clustering based on the network topological proximity while Figure 1b shows the clustering based on similar expression patterns.

The concept of Granger causality [6] has been previously shown to help in the identification and interpretation of regulatory networks in time series gene expression datasets [9-18]. The main advantage of Granger causality analysis in the context of gene expression datasets consists in the fact that each edge of the network represents the information flow from one gene to another [19]. Nevertheless, it is necessary to point out that Granger causality is not effective causality in the Aristothelic sense because it is based on prediction and numerical calculations. Fujita *et al.* [20-22] suggested a concept for the identification of Granger causality between groups of time series. The application was, however, limited to scenarios

*Correspondence: fujita@ime.usp.br
[1] Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo 05508-090, Brazil
Full list of author information is available at the end of the article

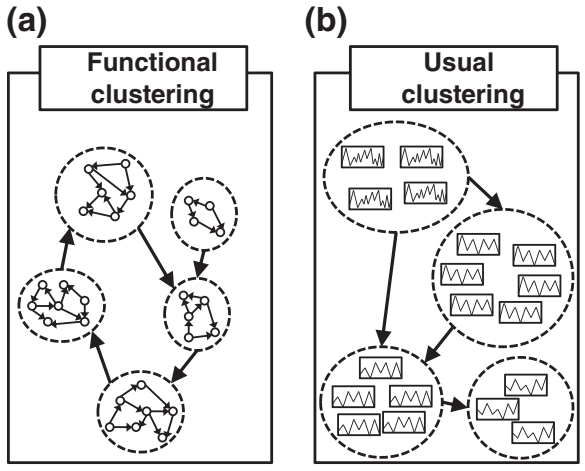


**Figure 1 Regulatory networks. (a)** Functional clustering. Genes are clustered based on their topological proximity given by Granger causality. **(b)** Usual clustering. Genes are clustered based on the similarity between gene expression levels.

when clusters could be previously defined based on particular data characteristics. Here, we propose a method to define clusters by their topological proximity in the network. For this purpose we introduce an extension of the concept of *functional clustering*, initially proposed by [23] in neuroscience. In [23], they applied mutual information in order to group the most active brain regions. We are interested in clustering the genes by using the concept of information flow [19] between *sets* of time series [20]. The gene expression time series are grouped depending on the hidden structure underlying the network topology, in a way that genes which are topologically close in terms of Granger causality are clustered (Figure 1a). We use the generalization of Granger causality for sets of time series datasets proposed by [20,21] in order to define concepts of distance, degree and flow useful to determine gene sets that highly interact in terms of Granger causality. In other words, we will derive the Granger causality-based functional clustering directly from the time series gene expression data. For this purpose, an approach that allows the identification of the optimum number of clusters for a given dataset is also presented.

## Materials and Methods

### Granger causality for sets of time series

Granger causality identification is a potential approach for the detection of possible interactions in a data driven framework couched in terms of temporal precedence. The main idea is that temporal precedence does not imply, but may help to identify causal relationships, since a cause never occurs after its effect.

A formal definition of Granger causality for sets of time series [20] can be given as follows.

**Definition 1.** [20] *Granger causality for sets of time series: Suppose that $\Im_t$ is a set containing all relevant information available up to and including time-point t. Let $\mathbf{X}_t$, $\mathbf{X}_t^i$ and $\mathbf{X}_t^j$ be sets of time series containing p, m and n time series, respectively, where $\mathbf{X}_t^i$ and $\mathbf{X}_t^j$ are disjoint subsets of $\mathbf{X}_t$, i.e., each time series only belongs to one set, and thus, $p \geq m + n$. Let $\mathbf{X}_t(h|\Im_t)$ 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 $\mathbf{X}_t^i$ from the time point t, based on the information in $\Im_t$. The forecast MSE of the linear combination of $\mathbf{X}_t^i$ will be denoted by $\Omega_{\mathbf{X}}(h|\Im_t)$. The set of n time series $\mathbf{X}_t^j$ is said to Granger-cause the set of m time series $\mathbf{X}_t^i$ if*

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

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

For the linear case, $\mathbf{X}_t^j$ is Granger non-causal for $\mathbf{X}_t^i$ if the following condition holds:

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

where $\rho$ is the largest correlation calculated by Canonical Correlation Analysis (CCA).

In order to simplify both notation and concepts, only the identification of Granger causality for sets of time series in an Autoregressive process of order one is presented. Generalizations for higher orders are straightforward.

### Functional clustering in terms of Granger causality

There are numerous definitions for clusters in networks in the literature [24]. A functional cluster in terms of Granger causality can be defined as a subset of genes that strongly interact among themselves but interact weakly with the rest of the network.

A usual approach for network clustering when the structure of the graph is known is the spectral clustering proposed by [25]. However, in biological data, the structure of the regulatory network is usually unknown.

In order to overcome this limitation, we developed a framework to cluster genes by their topological proximity using the time series gene expression information. We developed concepts of distance and degree for sets of time series based on Granger causality, and combined

them to the modified spectral clustering algorithm. The procedures are detailed below.

#### *Functional clustering*

Given a set of time series $x_t^1, x_t^2, \ldots, x_t^p$ (where $p$ is the number of time series) and a definition of similarity $w_{ij} \geq 0$ between all pairs of data points $x_t^i$ and $x_t^j$, the intuitive goal of clustering is to divide the time series into several groups such that time series in the same group are highly connected by Granger causality and time series in different groups are not connected or show few connections to each other. One usual representation of the connectivity between time series is in the form of graph $G = (V, E)$. Each vertex $v_i$ in this graph represents a time series gene expression $x_t^i$. Two vertices are connected if the similarity $w_{ij}$ between the corresponding time series $x_t^i$ and $x_t^j$ is not zero (the edge of the graph is weighted by $w_{ij}$). In other words, a $w_{ij} > 0$ represents existence of Granger causality between time series $x_t^i$ and $x_t^j$ and $w_{ij} = 0$ represents Granger non-causality. The problem of clustering can now be reformulated using the similarity graph: we want to find a partition of the graph such that there is less Granger causality between different groups and more Granger causality within the group.

Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, \ldots, v_p\}$ (where each vertex represents one time series) and weighted edges set $E$. In the following we assume that the graph $G$ is weighted, that is each edge between two vertices $v_i$ and $v_j$ carries a non-negative weight $w_{ij} \geq 0$. The weighted adjacency matrix of the graph is the matrix $\mathbf{W} = w_{ij}; i, j = 1, \ldots, p$. If $w_{ij} = 0$, this means that the vertices $v_i$ and $v_j$ are not connected by an edge. As $G$ is undirected, we require $w_{ij} = w_{ji}$. Therefore, in terms of Granger causality, $w_{ij}$ can be set as the distance between two time series $x_t^i$ and $x_t^j$. This distance can be defined as

**Definition 2.** *Distance between two (sets of) time series $x_t^i$ and $x_t^j$:*

$$dist\left(x_t^i, x_t^j\right) = 1 - \frac{|CCA(x_t^i, x_{t-1}^j)| + |CCA(x_t^j, x_{t-1}^i)|}{2}. \tag{3}$$

Notice that $\mathrm{CCA}(x_t^i, x_{t-1}^j)$ is the Granger causality from time series $x_t^j$ to $x_t^i$. In the case of sets of time series, just replace $x_t^i$ and $x_t^j$ by the set of time series $\mathbf{X}_t^i$ and $\mathbf{X}_t^j$ [20,21]. Since absolute value of CCA ranges from zero to one and the higher the CCA, the higher is the quantity of information flow, it is possible to see that the higher the CCA, the shorter the distance is. Furthermore, it is necessary to point out that the average between $\mathrm{CCA}(x_t^i, x_{t-1}^j)$ and $\mathrm{CCA}(x_t^j, x_{t-1}^i)$ is calculated because the distance must be symmetric. The intuitive idea consists on the fact that the higher is the CCA coefficient, the lower is the distance between the time series (or sets of time series) independent of the direction of Granger causality.

Moreover, notice that the CCA is the Pearson correlation after dimension reduction, therefore, $\mathrm{dist}(x_t^i, x_t^j)$ satisfies three out of four criteria for distances: (i) non-negativity; (ii) identity of indiscernible; and (iii) symmetry; and does not satisfy the (iv) triangular inequality, therefore, Pearson correlation is not a real metric. However, it is commonly used as a distance measure in several gene expression data analysis [26,27]. The main advantage with this definition of distance is the fact that it is possible to interpret the clustering process by a Granger causality concept.

Another necessary concept is the idea of degree of a time series $x_t^i$ (vertex $v_i$) which can be defined as

**Definition 3.** *Degree of $x_t^i$ is defined by:*

$$degree(x_t^i) = \frac{in\text{-}degree(x_t^i) + out\text{-}degree(x_t^i)}{2}, \tag{4}$$

*where in-degree and out-degree are respectively*

$$in\text{-}degree(x_t^i) = |CCA(x_t^i, \mathbf{X}_{t-1} | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}\})| \tag{5}$$

$$out\text{-}degree(x_t^i) = |CCA(\mathbf{X}_t, x_{t-1}^i | \mathbf{X}_t \backslash \{x_{t-1}^i\})|. \tag{6}$$

Notice that in-degree and out-degree represent the total information flow that "enters" and "leaves" the vertex $v_i$, respectively. Therefore, the degree of vertex $v_i$ contains the total information flow passing through vertex $v_i$.

Without loss of generality, it is possible to extend the concept of degree of a vertex $v_i$ (time series $x_t^i$) to a set of time series (sub-network) $\mathbf{X}_t^u$, where $u = 1, \ldots, k$ and $k$ is the number of sub-networks.

**Definition 4.** *Degree of sub-network $\mathbf{X}_t^u$ is defined by:*

$$degree(\mathbf{X}_t^u) = \frac{in\text{-}degree(\mathbf{X}_t^u) + out\text{-}degree(\mathbf{X}_t^u)}{2}, \tag{7}$$

*where in-degree and out-degree are respectively*

$$in\text{-}degree(\mathbf{X}_t^u) = |CCA(\mathbf{X}_t^u, \mathbf{X}_{t-1} | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}\})|, \tag{8}$$

$$out\text{-}degree(\mathbf{X}_t^u) = |CCA(\mathbf{X}_t, \mathbf{X}_{t-1}^u | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}^u\})|. \tag{9}$$

Now, by using the definitions of distance and degrees for time series and sets of time series in terms of Granger

causality, it is possible to develop a spectral clustering-based algorithm to identify sub-networks (set of time series that are highly connected within sets and poorly connected between sets) in the regulatory networks. The algorithm based on spectral clustering [25] is as follows:

**Input:** The $p$ time series $(x_t^i; i = 1, \ldots, p)$ and the number $k$ of sub-networks to construct.
**Step 1:** Let $\mathbf{W}$ be the $(p \times p)$ symmetric weighted adjacency matrix where
$w_{i,j} = w_{j,i} = 1 - \text{dist}(x_t^i; x_t^j), i, j = 1, \ldots, p.$
**Step 2:** Compute the non-normalized $(p \times p)$ Laplacian matrix $\mathbf{L}$ as (Mohar, 1991)

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{10}$$

where $\mathbf{D}$ is the $(p \times p)$ diagonal matrix with the degrees $d_1, \ldots, d_p$ $(\text{degree}(x_t^i) = d_i; i = 1, \ldots, p)$ on the diagonal.
**Step 3:** Compute the first $k$ eigenvectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ (corresponding to the $k$ largest eigenvalues) of $\mathbf{L}$.
**Step 4:** Let $\mathbf{U} \in \Re^{p \times k}$ be the matrix containing the vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ as columns.
**Step 5:** For $i = 1, \ldots, p$, let $\mathbf{y}_i \in \Re^k$ be the vector corresponding to the $i$th row of $\mathbf{U}$.
**Step 6:** Cluster the points $(\mathbf{y}_i)_{i=1,\ldots,p} \in \Re^k$ with the $k$-means algorithm into clusters $\{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$. For $k$-means, one may select a large number of initial values to achieve (or to be closer) the global optimum configuration. In our simulations, we generated 100 different initial values.
**Output:** Sub-networks $\{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$.

Notice that this clustering approach does not infer the entire structure of the network.

#### *Estimation of the number of clusters*

The method presented so far describes a framework for clustering genes (time series) using their topological proximity in terms of Granger causality.

Now, the challenge consists in determining the optimum number of sub-networks $k$. The choice of the number of sub-networks $k$ is often difficult depending on what the researcher is interested in. In our specific problem, one is interested in identifying the clusters presenting dense connectivity within a cluster and sparse connectivity between clusters.

In order to determine the most appropriate number of clusters in this specific context, we used a variant of the silhouette method [28].

Let us first define the cluster index $s(i)$ in the case of dissimilarities. Take any time series $x_t^i$ in the data set, and denote by $\mathbf{A}$ the sub-network to which it has been assigned. When sub-network $\mathbf{A}$ contains other time series apart from $x_t^i$, then we can compute: $a(i) = \text{dist}(x_t^i, \mathbf{A})$, which is the average dissimilarity of $x_t^i$ to $\mathbf{A}$. Let us now consider any sub-network $\mathbf{C}$ which is different from $\mathbf{A}$ and compute: $\text{dist}(\mathbf{x}_t^i, \mathbf{C})$ which is the dissimilarity of $x_t^i$ to $\mathbf{C}$. After computing $\text{dist}(x_t^i, \mathbf{C})$ for all sub-networks $\mathbf{C} \neq \mathbf{A}$, we set the smallest of those numbers and denote it by $b(i) = min_{\mathbf{C} \neq \mathbf{A}} \text{dist}(x_t^i, \mathbf{C})$. The sub-network $\mathbf{B}$ for which this minimum value is attained (that is, $\text{dist}(x_t^i, \mathbf{B}) = b(i)$) we call the neighbor sub-network, or cluster of $x_t^i$. The neighbor cluster would be the second-best cluster for time series $x_t^i$. In other words, if $x_t^i$ could not belong to sub-network $\mathbf{A}$, the best sub-network to belong to would be $\mathbf{B}$. Therefore, $b(i)$ is very useful to know the best alternative cluster for the time series in the network. Note that the construction of $b(i)$ depends on the availability of other sub-networks apart from $\mathbf{A}$, thus it is necessary to assume that there is more than one sub-network $k$ within a given network [28].

After calculating $a(i)$ and $b(i)$, the cluster index $s(i)$ can be obtained by combining them as follows:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}. \tag{11}$$

Indeed, from the above definition we easily see that $-1 \leq s(i) \leq 1$ for each time series $x_t^i$. Therefore, there are at least three cases to be analyzed, namely, when $s(i) \approx 1$ or $s(i) \approx 0$ or $s(i) \approx -1$. For cluster index $s(i)$ to be close to one we require $a(i) \ll b(i)$. As $a(i)$ is a measure of how dissimilar $i$ is to its own sub-network, a small value means it is well matched. Furthermore, a large $b(i)$ implies that $i$ is badly matched to its neighboring sub-network. Thus, a cluster index $s(i)$ close to one means that the gene is appropriately clustered. If $s(i)$ is close to negative one, then by the same logic we see that $x_t^i$ would be more appropriate if it was clustered in its neighboring sub-network. A cluster index $s(i)$ near zero means that the gene is on the border of two sub-networks. In other words, the cluster index $s(i)$ can be interpreted as the fitness of the time series $x_t^i$ to the assigned sub-network.

The average cluster index $s(i)$ of a sub-network is a measure of how tightly grouped all the genes in the sub-network are. Thus, the average cluster index $s(i)$ of the entire dataset is a measure of how appropriately the genes have been clustered in a topological point of view and in terms of Granger causality.

#### *Estimation of the number of clusters in biological data*

In order to estimate the most appropriate number of sub-networks present in the data set, we estimate the average cluster index $s$ of the entire dataset for each number of clusters $k$. When the number of identified sub-networks is equal or lower than the adequate number of sub-networks, the cluster index values are very

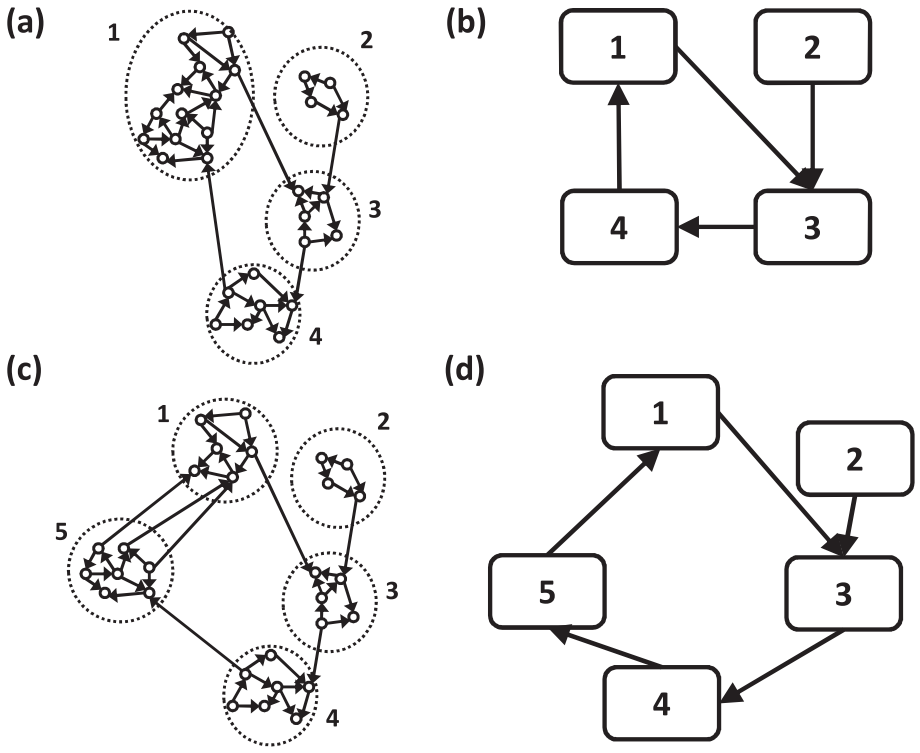


**Figure 2 (a) The representation of network with four clusters; (b) The network obtained by applying the proposed method with four clusters (*k*=4); (c) The representation of the network when the number of clusters is set to five; (d) The network obtained by applying the proposed method with five clusters (*k*=5).** The solid edges represent Granger causality. Notice that the structure of the "true" network (**(a)** and **(c)**) is not observed and it can only be estimated.

similar. However, when the number of identified sub-networks becomes higher than the adequate number of sub-networks, the cluster index value $s$ decreases abruptly. This is due to the fact that one of the highly connected sub-networks is split into two new sub-networks. Notice that these two new sub-networks present high connectivity between them because they are in fact, only one sub-network. In order to illustrate this event, see Figure 2 for an example. In Figure 2a, genes in cluster 1 are highly interconnected. Now, suppose that one wants to increase the number of clusters by splitting cluster 1 into two clusters namely clusters 1 and 5 (Figure 2c). Notice that clusters 1 and 5 are highly connected between them. If the number of clusters is higher than the adequate number of clusters (four, in our case), the value $s$ decreases substantially, since the Granger causality between clusters increases and the within cluster decreases. The breakpoint where the value $s$ decreases abruptly can be used to determine the adequate number of sub-networks. In fact, this can be visually identified by analyzing the breakpoint at the plot similarly to the standard elbow method used in k-means. However, if one wants to determine the breakpoint in an objective manner, this can be done by adjusting two linear regressions, one with the first $q$ dots and another with the remaining dots, thus identifying the breakpoint (the value $q$) that minimizes the sum of squared errors (Figure 3).

### Network construction

The network connecting clusters is constructed following procedures previously described [20,21]. Briefly, after Classification Expectation Maximization (CEM) [29] Principal Component Analysis (PCA) is used to remove redundancy and to extract the eigen-time series from each cluster. PCA allows us to keep only the most significant components leading to variability in the dataset, thus reducing the number of variables for subsequent processing. In this study, we retained only components accounting for more than 5% of the temporal variance in each cluster [22]. The eigen-time series are then clustered as described in the section *Functional clustering* and the network can be inferred by applying the method proposed by [20,21].

The Granger causality between cluster is identified by:

$$\mathrm{CCA}(\mathbf{X}_t^i, \mathbf{X}_{t-1}^j | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}^j\}) = \hat{\rho} \text{ for all } i,j = 1,\ldots,k \quad (12)$$

where $\hat{\rho}$ is the sample canonical correlation between the sets $\mathbf{X}_t^i$ and $\mathbf{X}_{t-1}^j$ partialized by all information contained in $\mathbf{X}_t$ minus the set $\mathbf{X}_{t-1}^j$.

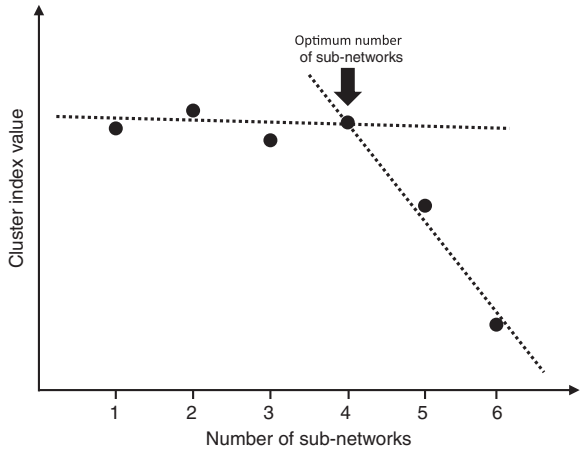


**Figure 3 The optimum number of sub-networks is indicated by the breakpoint in the graph.** The breakpoint appears when the number of sub-networks is greater than the adequate number of sub-networks. The breakpoint selection criterion is based on two linear regressions that best fit the data.

Then, test

$H_0 : \mathrm{CCA}(\mathbf{X}_t^i, \mathbf{X}_{t-1}^j | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}^j\}) = \hat{\rho} = 0$ (Granger non-causality)

$H_1 : \mathrm{CCA}(\mathbf{X}_t^i, \mathbf{X}_{t-1}^j | \mathbf{X}_t \backslash \{\mathbf{X}_{t-1}^j\}) = \hat{\rho} \neq 0$ (Granger causality) where $H_0$ and $H_1$ are the null and alternative hypothesis, respectively.

### Simulations

Four sets of Monte Carlo simulations were carried out in order to evaluate the proposed approach under controlled conditions. The first scenario represents four sub-networks without Granger causality between them (Figure 4a). The second scenario consists of four sub-networks constituting a cyclic graph (Figure 4b). The third scenario presents a feedback loop between sub-networks A and B (Figure 4c). The fourth scenario is composed of a network with one sub-network (sub-network D) that only receives Granger causality and one sub-network (sub-network A) that only sends Granger causality (Figure 4d). Since biological data usually possess several highly correlated genes (genes which hold the same information from a statistical stand point), we constructed 10 highly correlated time series for each $x_t^i, i = 1, \ldots, 20$. In other words, $x_t^1$ is represented by 10 time series with correlation of 0.6 between them, $x_t^2$ is represented by 10 time series with correlation of 0.6 between them and so on. Therefore, instead of 20 time series, each scenario is in fact composed of 200 time series.

For each scenario, time series lengths varied: 50, 75, 1000 and 200 time points. The number of repetitions for each scenario is 1,000. The synthetic gene expression time series data in sub-networks A, B, C and D were generated by the following equations described below.

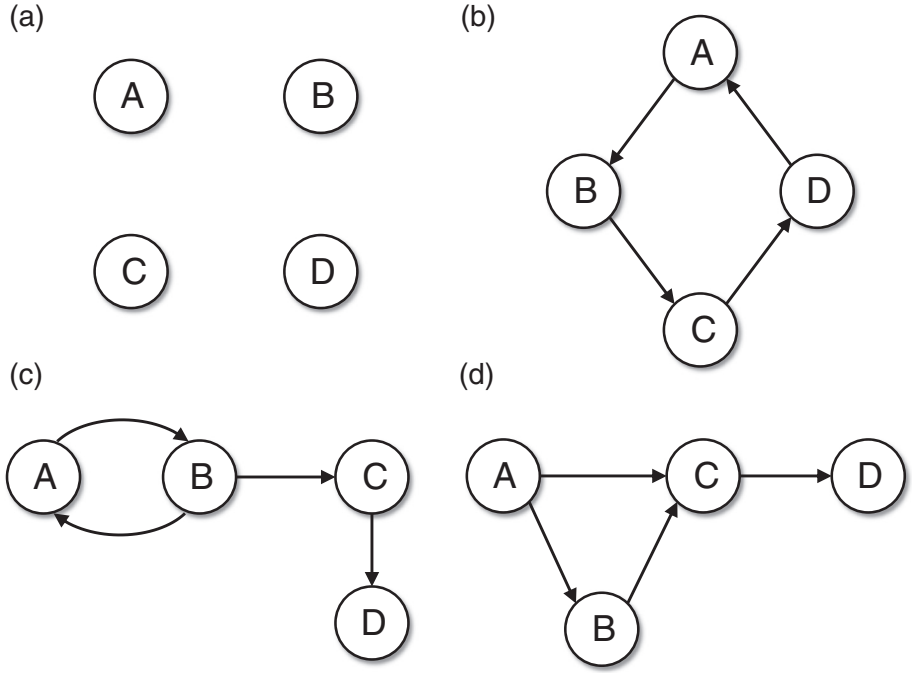


**Figure 4 (a) Four independent sub-networks without Granger causality between them; (b) Four sub-networks in a cyclic graph; (c) Feedback loop between sub-networks A and B; (d) A network between sub-networks, where sub-networks A only sends Granger causality and D only receives Granger Causality.**

Simulation 1:

$$
\begin{cases}
x^{\mathrm{A}}_{1,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{4,t-1} + \beta x^{\mathrm{A}}_{3,t-1} + \varepsilon_{1,t} \\
x^{\mathrm{A}}_{2,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{2,t} \\
x^{\mathrm{A}}_{3,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{3,t} \\
x^{\mathrm{A}}_{4,t} = \beta x^{\mathrm{A}}_{3,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{4,t} \\
x^{\mathrm{A}}_{5,t} = \beta x^{\mathrm{A}}_{2,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{5,t} \\
x^{\mathrm{B}}_{6,t} = \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{6,t} \\
x^{\mathrm{B}}_{7,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{7,t} \\
x^{\mathrm{B}}_{8,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{8,t} \\
x^{\mathrm{B}}_{9,t} = \beta x^{\mathrm{B}}_{7,t-1} - \beta x^{\mathrm{B}}_{8,t-1} + \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \varepsilon_{9,t} \\
x^{\mathrm{B}}_{10,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{9,t-1} + \varepsilon_{10,t} \\
x^{\mathrm{C}}_{11,t} = \beta x^{\mathrm{C}}_{12,t-1} - \beta x^{\mathrm{C}}_{15,t-1} + \varepsilon_{11,t} \\
x^{\mathrm{C}}_{12,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{13,t-1} + \varepsilon_{12,t} \\
x^{\mathrm{C}}_{13,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \varepsilon_{13,t} \\
x^{\mathrm{C}}_{14,t} = \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \varepsilon_{14,t} \\
x^{\mathrm{C}}_{15,t} = \beta x^{\mathrm{C}}_{15,t-1} - \beta x^{\mathrm{C}}_{12,t-1} + \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{14,t-1} + \varepsilon_{15,t} \\
x^{\mathrm{D}}_{16,t} = \beta x^{\mathrm{D}}_{19,t-1} - \beta x^{\mathrm{D}}_{20,t-1} + \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{16,t} \\
x^{\mathrm{D}}_{17,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{17,t-1} + \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{17,t} \\
x^{\mathrm{D}}_{18,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{18,t} \\
x^{\mathrm{D}}_{19,t} = \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{19,t} \\
x^{\mathrm{D}}_{20,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{19,t-1} + \varepsilon_{20,t}
\end{cases}
$$

Simulation 2:

$$
\begin{cases}
x^{\mathrm{A}}_{1,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{4,t-1} + \beta x^{\mathrm{A}}_{3,t-1} + \varepsilon_{1,t} \\
x^{\mathrm{A}}_{2,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{2,t} \\
x^{\mathrm{A}}_{3,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \gamma x^{\mathrm{D}}_{19,t-1} + \varepsilon_{3,t} \\
x^{\mathrm{A}}_{4,t} = \beta x^{\mathrm{A}}_{3,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{4,t} \\
x^{\mathrm{A}}_{5,t} = \beta x^{\mathrm{A}}_{2,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{5,t} \\
x^{\mathrm{B}}_{6,t} = \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \gamma x^{\mathrm{A}}_{3,t-1} + \varepsilon_{6,t} \\
x^{\mathrm{B}}_{7,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{7,t} \\
x^{\mathrm{B}}_{8,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{8,t} \\
x^{\mathrm{B}}_{9,t} = \beta x^{\mathrm{B}}_{7,t-1} - \beta x^{\mathrm{B}}_{8,t-1} + \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \varepsilon_{9,t} \\
x^{\mathrm{B}}_{10,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{9,t-1} + \varepsilon_{10,t} \\
x^{\mathrm{C}}_{11,t} = \beta x^{\mathrm{C}}_{12,t-1} - \beta x^{\mathrm{C}}_{15,t-1} + \varepsilon_{11,t} \\
x^{\mathrm{C}}_{12,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{13,t-1} + \varepsilon_{12,t} \\
x^{\mathrm{C}}_{13,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \varepsilon_{13,t} \\
x^{\mathrm{C}}_{14,t} = \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \gamma x^{\mathrm{B}}_{10,t-1} + \varepsilon_{14,t} \\
x^{\mathrm{C}}_{15,t} = \beta x^{\mathrm{C}}_{15,t-1} - \beta x^{\mathrm{C}}_{12,t-1} + \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{14,t-1} + \varepsilon_{15,t} \\
x^{\mathrm{D}}_{16,t} = \beta x^{\mathrm{D}}_{19,t-1} - \beta x^{\mathrm{D}}_{20,t-1} + \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{16,t} \\
x^{\mathrm{D}}_{17,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{17,t-1} + \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{17,t} \\
x^{\mathrm{D}}_{18,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{18,t} \\
x^{\mathrm{D}}_{19,t} = \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{19,t} \\
x^{\mathrm{D}}_{20,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{19,t-1} + \gamma x^{\mathrm{C}}_{15,t-1} + \varepsilon_{20,t}
\end{cases}
$$

Simulation 3:

$$
\begin{cases}
x^{\mathrm{A}}_{1,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{4,t-1} + \beta x^{\mathrm{A}}_{3,t-1} + \varepsilon_{1,t} \\
x^{\mathrm{A}}_{2,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{2,t} \\
x^{\mathrm{A}}_{3,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \gamma x^{\mathrm{B}}_{6,t-1} + \varepsilon_{3,t} \\
x^{\mathrm{A}}_{4,t} = \beta x^{\mathrm{A}}_{3,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{4,t} \\
x^{\mathrm{A}}_{5,t} = \beta x^{\mathrm{A}}_{2,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{5,t} \\
x^{\mathrm{B}}_{6,t} = \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{6,t} \\
x^{\mathrm{B}}_{7,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \gamma x^{\mathrm{A}}_{3,t-1} + \varepsilon_{7,t} \\
x^{\mathrm{B}}_{8,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{8,t} \\
x^{\mathrm{B}}_{9,t} = \beta x^{\mathrm{B}}_{7,t-1} - \beta x^{\mathrm{B}}_{8,t-1} + \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \varepsilon_{9,t} \\
x^{\mathrm{B}}_{10,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{9,t-1} + \varepsilon_{10,t} \\
x^{\mathrm{C}}_{11,t} = \beta x^{\mathrm{C}}_{12,t-1} - \beta x^{\mathrm{C}}_{15,t-1} + \varepsilon_{11,t} \\
x^{\mathrm{C}}_{12,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{13,t-1} + \varepsilon_{12,t} \\
x^{\mathrm{C}}_{13,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \varepsilon_{13,t} \\
x^{\mathrm{C}}_{14,t} = \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \gamma x^{\mathrm{B}}_{10,t-1} + \varepsilon_{14,t} \\
x^{\mathrm{C}}_{15,t} = \beta x^{\mathrm{C}}_{15,t-1} - \beta x^{\mathrm{C}}_{12,t-1} + \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{14,t-1} + \varepsilon_{15,t} \\
x^{\mathrm{D}}_{16,t} = \beta x^{\mathrm{D}}_{19,t-1} - \beta x^{\mathrm{D}}_{20,t-1} + \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{16,t} \\
x^{\mathrm{D}}_{17,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{17,t-1} + \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{17,t} \\
x^{\mathrm{D}}_{18,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{18,t} \\
x^{\mathrm{D}}_{19,t} = \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{19,t} \\
x^{\mathrm{D}}_{20,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{19,t-1} + \gamma x^{\mathrm{C}}_{15,t-1} + \varepsilon_{20,t}
\end{cases}
$$

Simulation 4:

$$
\begin{cases}
x^{\mathrm{A}}_{1,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{4,t-1} + \beta x^{\mathrm{A}}_{3,t-1} + \varepsilon_{1,t} \\
x^{\mathrm{A}}_{2,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{2,t} \\
x^{\mathrm{A}}_{3,t} = \beta x^{\mathrm{A}}_{1,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{3,t} \\
x^{\mathrm{A}}_{4,t} = \beta x^{\mathrm{A}}_{3,t-1} - \beta x^{\mathrm{A}}_{5,t-1} + \varepsilon_{4,t} \\
x^{\mathrm{A}}_{5,t} = \beta x^{\mathrm{A}}_{2,t-1} - \beta x^{\mathrm{A}}_{3,t-1} + \beta x^{\mathrm{A}}_{4,t-1} + \varepsilon_{5,t} \\
x^{\mathrm{B}}_{6,t} = \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{6,t} \\
x^{\mathrm{B}}_{7,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{7,t-1} + \gamma x^{\mathrm{A}}_{3,t-1} + \varepsilon_{7,t} \\
x^{\mathrm{B}}_{8,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{7,t-1} + \varepsilon_{8,t} \\
x^{\mathrm{B}}_{9,t} = \beta x^{\mathrm{B}}_{7,t-1} - \beta x^{\mathrm{B}}_{8,t-1} + \beta x^{\mathrm{B}}_{9,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \varepsilon_{9,t} \\
x^{\mathrm{B}}_{10,t} = \beta x^{\mathrm{B}}_{10,t-1} - \beta x^{\mathrm{B}}_{6,t-1} + \beta x^{\mathrm{B}}_{9,t-1} + \varepsilon_{10,t} \\
x^{\mathrm{C}}_{11,t} = \beta x^{\mathrm{C}}_{12,t-1} - \beta x^{\mathrm{C}}_{15,t-1} + \varepsilon_{11,t} \\
x^{\mathrm{C}}_{12,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{13,t-1} + \varepsilon_{12,t} \\
x^{\mathrm{C}}_{13,t} = \beta x^{\mathrm{C}}_{14,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \gamma x^{\mathrm{B}}_{8,t-1} + \varepsilon_{13,t} \\
x^{\mathrm{C}}_{14,t} = \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{11,t-1} + \gamma^{\mathrm{A}}_{5,t-1} + \varepsilon_{14,t} \\
x^{\mathrm{C}}_{15,t} = \beta x^{\mathrm{C}}_{15,t-1} - \beta x^{\mathrm{C}}_{12,t-1} + \beta x^{\mathrm{C}}_{13,t-1} - \beta x^{\mathrm{C}}_{14,t-1} + \varepsilon_{15,t} \\
x^{\mathrm{D}}_{16,t} = \beta x^{\mathrm{D}}_{19,t-1} - \beta x^{\mathrm{D}}_{20,t-1} + \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{16,t} \\
x^{\mathrm{D}}_{17,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{17,t-1} + \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{17,t} \\
x^{\mathrm{D}}_{18,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{18,t} \\
x^{\mathrm{D}}_{19,t} = \beta x^{\mathrm{D}}_{17,t-1} - \beta x^{\mathrm{D}}_{18,t-1} + \varepsilon_{19,t} \\
x^{\mathrm{D}}_{20,t} = \beta x^{\mathrm{D}}_{20,t-1} - \beta x^{\mathrm{D}}_{19,t-1} + \gamma x^{\mathrm{C}}_{15,t-1} + \varepsilon_{20,t}
\end{cases}
$$

where $\beta = 0.6$, $\gamma = 0.3$, $\varepsilon_{i,t} \sim N(0, \Sigma)$ with

$$\Sigma = \mathbf{I}_{(20\times 20)} \otimes \Gamma \tag{13}$$

and

$$\Gamma = \begin{pmatrix} 1 & 0.6 & \dots & 0.6 \\ 0.6 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.6 \\ 0.6 & \dots & 0.6 & 1 \end{pmatrix}_{(10\times 10)} \tag{14}$$

for $i = 1, \dots, 20$.

#### Actual biological data

In order to illustrate an application of the proposed approach, a dataset collected by [30] was used. The work presents whole genome gene expression data during the cell division cycle of a human cancer cell line (HeLa) characterized using cDNA microarrays. The dataset contains three complete cell cycles of ∼16 hours each, with a total of 48 time points distributed at intervals of one hour. The full dataset is available at: http://genome-www.stanford.edu/Human-CellCycle/HeLa/.

In order to evaluate our proposed approach, we chose to analyze the same gene set examined in Figure 5 of [10], which comprised a set of 50 genes.

## Results

#### Simulated data

In order to study the properties of the proposed functional clustering method and to check its consistency, we performed four simulations with distinct network characteristics in terms of structure and Granger causality.

Table 1 describes the frequency that each number of clusters was identified as optimal in each simulation and time series length. Notice that the accuracy of the method in identifying the correct number of clusters clearly converges to 100% as the time series length increases (the correct number of clusters is four for all the scenarios). The same result was obtained with varying numbers of sub-networks or when Granger causality within clusters increased, demonstrating the consistency of the method. Moreover, both the cluster indices value and the respective standard deviation for each simulation and time series length are described. The average cluster index value was calculated by using the value at the breakpoint as described in Figure 3 in 1,000 repetitions. By analyzing Table 1, it is possible to verify that the longer the time series length, the smaller are the standard deviations and the greater is the silhouette width demonstrating that the method is consistent.

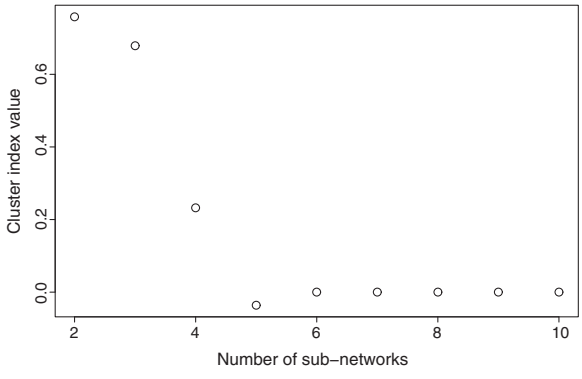


**Figure 5 The optimum number of sub-networks in the actual biological data is indicated by the breakpoint in the graph (the optimum number in this case is three).**

Table 2 describes the average of the frequency (in percentage) the time series were correctly clustered for each scenario and each time series length given the correct number of clusters. It is important to point out that the number of correctly classified time series increases as the time series length increases.

Table 3 represents the frequency (in percentage) each edge of the simulated network was identified when the estimated number of clusters were correctly identified as four. The correctly identified edges are in bold. Since the p-value threshold was set to 0.05, it is expected to identify $\approx$ 5% of false positive edges where there is indeed no Granger causality. In fact, where there is no Granger causality, the rate of false positives was controlled to 5%, and where there is Granger causality, the number of identified edges is clearly higher than where there is no Granger causality.

#### Biological data

By applying the method described in section *Functional clustering* to the biological dataset, the optimum number of sub-networks was identified as three. Notice in Figure 5 that there is a clear breakpoint when the number of clusters is three.

Once clusters were obtained, the cluster-cluster network (Figure 6) was modeled by applying the method described in [20,21]. Two of the depicted clusters, clusters one and two, provide interesting material for biological interpretation. Genes belonging to cluster two highlight expected interconnections in cell cycle regulation. For instance, aberrant activation of signal transcription factors NF-$\kappa$B or STAT3, and alterations in p53 status, have each been reported to affect cell survival individually. The presence of the three genes in the same cluster is in agreement with a recent study which examined the hypothesis that alterations in a signal network involving NF-$\kappa$B, STAT3 and p53 could modulate expression of proapoptotic BAX and antiapoptotic BCL-XL proteins, promoting cell survival [31]. The authors show that overexpression of p53 together with inhibition of NF-$\kappa$B or STAT3 induced greater increase in the BAX/BCL-XL ratio

**Table 1 Frequency of the selected number of clusters for each scenario and time series length**

| Time series length/Number of clusters | 1 | 2 | 3 | 4 | 5 | 6 | silhouette width |
|---|---|---|---|---|---|---|---|
| Scenario 1 | | | | | | | |
| 50 | 0 | 0 | 48 | **700** | 252 | 0 | 0.502 (0.098) |
| 75 | 0 | 0 | 1 | **785** | 214 | 0 | 0.582 (0.054) |
| 100 | 0 | 0 | 3 | **805** | 192 | 0 | 0.610 (0.042) |
| 200 | 0 | 0 | 4 | **825** | 171 | 0 | 0.641 (0.034) |
| Scenario 2 | | | | | | | |
| 50 | 0 | 0 | 65 | **713** | 222 | 0 | 0.479 (0.112) |
| 75 | 0 | 0 | 28 | **760** | 212 | 0 | 0.555 (0.071) |
| 100 | 0 | 0 | 9 | **834** | 157 | 0 | 0.587 (0.050) |
| 200 | 0 | 0 | 3 | **883** | 114 | 0 | 0.621 (0.029) |
| Scenario 3 | | | | | | | |
| 50 | 0 | 0 | 63 | **666** | 271 | 0 | 0.461 (0.123) |
| 75 | 0 | 0 | 18 | **784** | 198 | 0 | 0.552 (0.078) |
| 100 | 0 | 0 | 8 | **851** | 141 | 0 | 0.586 (0.050) |
| 200 | 0 | 0 | 6 | **883** | 111 | 0 | 0.618 (0.031) |
| Scenario 4 | | | | | | | |
| 50 | 0 | 0 | 53 | **686** | 261 | 0 | 0.465 (0.110) |
| 75 | 0 | 0 | 17 | **786** | 197 | 0 | 0.551 (0.075) |
| 100 | 0 | 0 | 11 | **815** | 174 | 0 | 0.581 (0.055) |
| 200 | 0 | 0 | 6 | **887** | 107 | 0 | 0.619 (0.033) |

In bold are the correct number of clusters. Between brackets is one standard deviation for the silhouette width calculated in the breakpoint. For each scenario and each time series length, the number of repetitions was set to 1,000.

than modulation of these transcription factors individually. As discussed earlier in this paper, this is a situation in which similar patterns of gene expression are not sufficient to comprehend the biological process.

In [10], a network depicting Granger interaction among genes from this same gene dataset was presented. The authors analyzed the network in the context of tumor progression and identified gene-gene connections associated with NF-$\kappa$B, p53, and STAT3. Here, cluster 1 groups not only NF-$\kappa$B, p53, and STAT3, but also the functionally associated gene BCL-XL, NF-$\kappa$B regulator A20 and targets IAP and i$\kappa$B$\alpha$. The presence of NF-$\kappa$B and fibroblast growth factors (FGFs) and receptors (FGFRs) in the same cluster is also in agreement with the previous work. Members of the FGF family and NF-$\kappa$B have been shown to interact in various contexts and, despite distinct roles, are involved in cell proliferation, migration and survival [32,33].

**Table 2 Average of the percentage of correctly clustered time-series in 1,000 repetitions given the correct number of clusters**

| Scenario/Time series length | 50 | 75 | 100 | 200 |
|---|---|---|---|---|
| 1 | 78.8 | 96.0 | 98.9 | 99.9 |
| 2 | 72.9 | 91.2 | 95.8 | 99.2 |
| 3 | 71.6 | 90.6 | 95.2 | 99.7 |
| 4 | 68.9 | 88.7 | 93.7 | 99.1 |

Even though MCL-1 and P21 play important roles in cell survival, and BAI1 is transcriptionally regulated by P53, the analysis run here clustered them separately from P53 containing cluster. This result suggests that, in the context of this dataset, their interaction is stronger with genes such as c-JUN, also functionally related to cell survival, proto-oncogene MET and tumor suppressor MASPIN, for instance. Also worth noticing is the interaction of this cluster with the two members of cluster 3: FGF5 and FOP. Like the other members of FGF family grouped in cluster 2, FGF5 is involved in cell survival activities, while FOP was originally discovered as a fusion partner with FGFR1 in oncoproteins that give raise to stem cell myeloproliferative disorders. It would be interesting to identify specific details regarding the intensity and direction of the information flow within this cluster for a clearer understanding of their relationship in the context of cell cycle progression.

## Discussions

Fujita *et al.* [20,21] suggested both a concept of Granger causality for sets of time series and a method for its

**Table 3 Percentage of edges with time series length equals to 50/75/100/200 when the estimated number of clusters were correctly identified as four**

| from/to | A | B | C | D |
|---|---|---|---|---|
| Scenario 1 | | | | |
| A | **100/100/100/100** | 6.7/6.3/5.2/5.4 | 8.9/6.0/5.0/5.3 | 4.8/5.7/5.4/4.5 |
| B | 6.9/7.1/5.5/6.8 | **99.9/100/100/100** | 7.8/6.2/6.3/4.6 | 5.6/6.9/4.9/5.6 |
| C | 7.6/5.9/6.5/5.6 | 6.9/7.7/4.7/5.1 | **100/100/100/100** | 4.9/5.4/5.7/5.8 |
| D | 6.2/5.3/5.1/4.7 | 5.3/5.2/5.3/5.7 | 7.0/5.2/5.2/5.6 | **100/100/100/100** |
| Scenario 2 | | | | |
| A | **100/100/100/100** | **28.9/59.8/80.4/99.7** | 8.0/6.4/6.8/5.2 | 6.4/6.6/5.0/5.0 |
| B | 5.4/5.3/5.5/4.6 | **100/100/100/100** | **29.6/60.9/82.1/99.9** | 6.4/6.3/5.7/5.7 |
| C | 7.5/5.4/6.7/4.5 | 8.8/6.6/6.6/6.3 | **100/100/100/100** | **23.0/50.4/71.2/99.1** |
| D | **17.6/35.5/51.2/95.4** | 6.5/4.2/3.4/5.0 | 12.5/10.4/7.5/5.0 | **100/100/100/100** |
| Scenario 3 | | | | |
| A | **100/100/100/100** | **29.6/61.9/82.1/100** | 7.8/7.3/4.5/5.0 | 7.4/6.8/4.6/5.2 |
| B | **28.5/53.0/78.0/99.9** | **100/100/100/100** | **31.8/61.1/82.9/99.9** | 7.0/7.1/6.2/4.7 |
| C | 8.4/6.9/6.4/5.6 | 7.6/7.8/7.3/5.2 | **99.9/100/100/100** | **25.5/46.8/70.6/99.3** |
| D | 6.8/5.6/5.8/5.0 | 5.5/4.5/5.7/4.3 | 13.9/8.2/6.1/5.4 | **100/100/100/100** |
| Scenario 4 | | | | |
| A | **100/100/100/100** | **25.1/52.6/75.8/99.6** | **22.9/41.8/59.5/96.0** | 6.8/5.8/5.2/4.7 |
| B | 6.7/5.9/5.7/5.9 | **100/100/100/100** | **28.6/58.4/81.9/100** | 7.9/6.0/6.1/5.2 |
| C | 9.3/8.8/6.1/6.2 | 8.8/6.2/6.3/4.5 | **100/100/100/100** | **26.5/53.2/75.4/99.2** |
| D | 5.4/5.8/5.1/4.7 | 5.8/5.0/4.2/5.2 | 14.9/11.9/7.9/5.4 | **100/100/100/100** |

The rows and columns represent the clusters A, B, C, and D. The rate of false positives was controlled to 5% (p-value $< 0.05$). The edges which actually exists in the network are shown in bold.

identification with a statistical test to control the rate of false positives. Although this method is useful for the identification of Granger causality between sets of time series in Bioinformatics and Neuroscience [22], the application was limited to pre-defined clusters, i.e., the time series composing each cluster needed to be previously known. We developed an objective method to define clusters based on the intuitive concept that a gene cluster should interact more intensely in terms of Granger causality within itself than with neighboring clusters.

Krishna *et al*. [34] proposed a Granger causality clustering method based on the structure of a pair-wise network. Their method consists in identifying pairwise Granger causality between gene expression time series and then, by applying the method proposed by Bader and Hogue (2003), to detect dense regions in the network. The difference between their approach and ours is that they take into account the number of edges, and the density of the network which is given by the number of estimated edges divided by the total number of possible edges. The presence of an edge is determined by the p-value's threshold. Notice that depending on the threshold, the results can change. In our framework, we take into account the weight of Granger causality between sets of time series in order to identify how close two sets are. Consequently, it is possible to obtain a notion of distance between two clusters based on Granger causality, i.e., a continuous measure (distance in terms of Granger causality) instead of a discrete measure (presence or absence of an edge). Moreover, by using the concept of Granger causality between sets of time series proposed by [20], the concept of density of a network can be easily defined in terms of Granger causality instead of a density based on the number of edges as proposed by [34].

A disadvantage of our method is that it cannot be applied for very large datasets. The larger is the number of time series (genes), or the higher the order of the autoregressive process to be analyzed, the higher the chance to generate non-invertible covariance matrices in the calculation of distance (definition 2) and degree (definition 4) between clusters. We believe that this drawback can be overcome through sparse canonical correlation analysis [35], recently proposed in the literature. However, this topic deserves further studies before it can be used in both clustering and identification of Granger causality between sets of time series, since penalized methods relying on L1 penalization [35] or kernel [36] may present biased estimators.

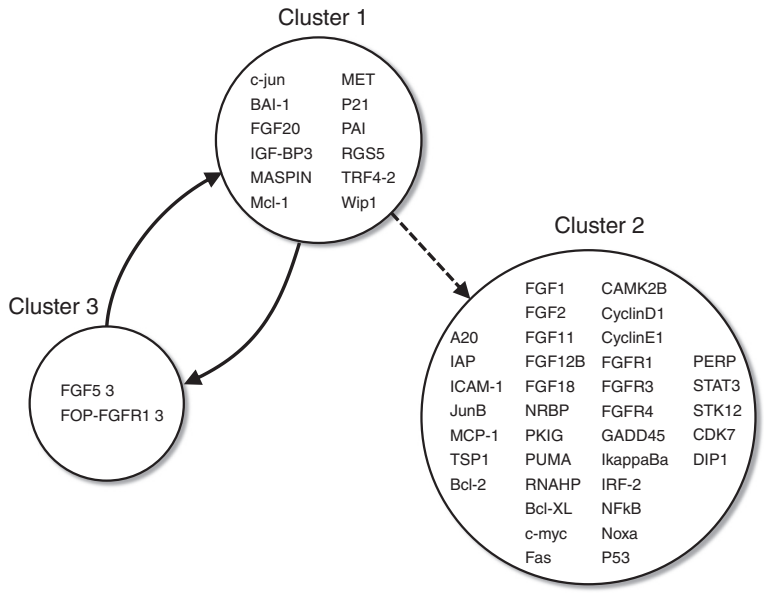


**Figure 6 The network obtained with three (*k*=3) sub-networks.** Solid arrows are significant Granger causality with p-value $< 0.05$ and dashed arrow is significant Granger causality with p-value $< 0.10$. The circles represent the clusters.

We only analyzed the autoregressive process of order one because gene expression time series data, possibly due to experimental limitations, are typically not large. However, if one is interested in analyzing greater orders, one minus the maximum canonical correlation analysis value among all the tested autoregressive orders can be used as the distance measure between two time series.

The clustering algorithm used here is based on the well-known spectral clustering. Although results were satisfactory, other graph clustering methods may be used. The normalized cuts algorithm proposed by [37], for instance, presents better results in non Gaussian data sets.

Finally, which biological process underlie time series datasets correlation, remains a difficult question to be answered. Studies suggest that correlated genes may belong to common pathways or present the same biological function. However, it is also known that methods based exclusively on correlation cannot reconstruct entire gene networks. Further studies in the field of systems biology might be able to answer this question in the future.

## Conclusions

We propose a time series clustering approach based on Granger causality and a method to determine the number of clusters that best fit the data. This method consists of (1) the definition of degree and distance, usually used in graph theory but now generalized for time series data analysis in terms of Granger causality; (2) a clustering algorithm based on spectral clustering and (3) a criterion to determine the number of clusters. We demonstrate, by simulations, that our approach is consistent even when the number of genes is greater than the time series' length.

We believe that this approach can be useful to understand how gene expression time series relate to each other, and therefore help in the functional interpretation of data.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

AF has made substantial contributions to the conception and design of the study, analysis and interpretation of data. KK, AGP and JRS contributed to the analysis and interpretation of mathematical results. PS contributed to the analysis and interpretation of biological data. AF and PS have been involved in drafting of the manuscript. SM directed the work. All authors read approved the final manuscript.

#### Acknowledgements

The supercomputing resource was provided by Human Genome Center (Univ. of Tokyo). This work was supported by FAPESP and CNPq - Brazil and RIKEN - Japan.

#### Author details

[1]Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo 05508-090, Brazil. [2]Center for Experimental Research, Albert Einstein Research and Education Institute, Av. Albert Einstein, 627 - São Paulo, 05652-000, Brazil. [3]Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan. [4]Center of Mathematics, Computation and Cognition, Universidade Federal do ABC, Rua santa Adélia, 166 - Santo André, 09210-170, Brazil.

Received: 14 October 2011 Accepted: 17 October 2012
Published: 30 October 2012

#### References

1. Ng SK, McLachlan GJ, Wang K, Jones LB-T, Ng S-W: **A mixture model with random-effects components for clustering correlated gene-expression profiles.** *Bioinformatics* 2006, **22:**1745–1752.

2. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34:**166–176.
3. Shiraishi Y, *et al*: **Inferring cluster-based networks from differently stimulated multiple time-course gene expression data.** *Bioinformatics* 2010, **26:**1073–1081.
4. Stuart JM, Segal E, Koller D, Kim SK: **A gene co-expression network for global discovery of conserved genetics modules.** *Science* 2003, **302:**249–55.
5. Yamaguchi R, Yoshida R, Imoto S, Higuchi T, Miyano S: **Finding module-based networks with state-space models - mining high-dimensional and short time-course gene expression data.** *IEEE Signal Process Mag* 2007, **24:**37–46.
6. Granger CWJ: **Investigating causal relationships by econometric models and cross-spectral methods.** *Econometrica* 1969, **37:**424–438.
7. Ahmed HA, Mahanta P, Bhattacharyya DK, Kalita JK: **GERC: tree based clustering for gene expression data.** *11th IEEE Int Conference Bioinf Bioeng* 2011:299–302 .
8. Bandyopadhyay S, Bhattacharyya M: **A biologically inspired measure for coexpression analysis.** *IEEE/ACM Trans comput biol bioinf* 2011, **8:**929–942.
9. Fujita A, Sato JR, Garay-Malpartida HM, Morettin PA, Sogayar MC, Ferreira CE: **Time-varying modeling of gene expression regulatory networks using wavelet dynamic vector autoregressive method.** *Bioinformatics* 2007a, **23:**16253–1630.
10. 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 Syst Biol* 2007b, **1:**39.
11. Fujita A, Sato JR, Garay-Malpartida HM, Sogayar MC, Ferreira CE, Miyano S: **Modeling nonlinear gene regulatory networks from time-series gene expression data.** *J Bioinf Comput Biol* 2008, **6:**961–79.
12. Fujita A, Patriota AG, Sato JR, Miyano S: **The impact of measurement error in the identification of regulatory networks.** *BMC Bioinf* 2009, **10:**412.
13. Guo S, Wu J, Ding M, Feng J: **Uncovering interactions in the frequency domain.** *PLoS Comput Biol* 2008, **4:**e1000087.
14. 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* 2008, **20:**37–51.
15. Lozano AC, Abe N, Liu Y, Rosset S: **Grouped graphical Granger modeling for gene expression regulatory networks discovery.** *Bioinformatics* 2009, **25:**i110—i118.
16. Mukhopadhyay ND, Chatterjee S: **Causality and pathway search in microarray time series experiments.** *Bioinformatics* 2007, **23:**442–449.
17. Nagarajan R: **A note on inferring acyclic network structures using Granger causality tests.** *Int J Biostatistics* 2009, **5:**10.
18. Shojaie A, Michailidis G: **Discovering graphical Granger causality using the truncating lasso penalty.** *Bioinformatics* 2010, **26:**i517—i523.
19. Baccala LA, Sameshima K: **Partial directed coherence: A new concept in neural structure determination.** *Biol Cybernetics* 2001, **84:**463–474.
20. Fujita A, Sato JR, Kojima K, Gomes LR, Nagasaki M, Sogayar MC, Miyano S: **Identification of Granger causality between gene sets.** *J Bioinf Comput Biol* 2010a, **8:**679–701.
21. Fujita A, Kojima K, Patriota AG, Sato JR, Severino P, Miyano S: **A fast and robust statistical test based on Likelihood ratio with Bartlett correction to identify Granger causality between gene sets.** *Bioinformatics* 2010b, **26:**2349–2351.
22. Sato JR, Fujita A, Cardoso EF, Thomaz CE, Brammer MJ, Amaro E: **Analyzing the connectivity between regions of interest: An approach based on cluster Granger causality for fMRI data analysis.** *NeuroImage* 2010, **52:**1444–1455.
23. Tononi G, McIntosh AR, Russel DP, Edelman GM: **Functional clustering: identifying strongly interactive brain regions in neuroimaging data.** *NeuroImage* 1998, **7:**133–149.
24. Edachery J, Sen A, Brandenburg F: **Graph clustering using distance-k cliques.** In *Proceedings of the Seventh International Symposium on Graph Drawing. Lecture Notes in Computer Science. vol. 1731*. Edited by Smith Y. Berlin, Heidelberg, Germany: Springer-Verlag GmbH:1999.
25. Ng A, *et al*: *Advances in Neural Information Processing Systems*. New York: MIT Press; 2002.
26. Bhattacharya A, De RK: **Divisive correlation clustering algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles.** *Bioinformatics* 2008, **24:**1359–1366.
27. Ihmels J, Bergmann S, Berman J, Barkai N: **Comparative gene expression analysis by differential clustering approach: applications to the Candida albicans transcription program.** *PLoS Genet* 2005, **1:**e39.
28. Rousseeuw PJ: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** *J Comput Appl Math* 1987, **20:**53–65.
29. Celeux G, Govaert G: **A classification EM algorithm for clustering and two stochastic versions.** *Comput Stat Data Anal* 1992, **14:**315–332.
30. 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.** *Mol Biol Cell* 2002, **13:**1977–2000.
31. Lee TL, Yeh J, Friedman J, Yan B, Yang X, Yeh NT, Waes CV, Chen Z: **A signal network involving coactivated NF-$\kappa$B and STAT3 and altered p53 modulates BAX/BCL-XL expression and promotes cell survival of head and neck squamous cell carcinomas.** *Int J Cancer* 2008, **122:**1987–1998.
32. Lungu G, Covaleda L, Mendes O, Martini-Stoica H, Stoica G: **FGF-1-induced matrix metalloproteinase-9 expression in breast cancer cells is mediated by increased activities of NF-$\kappa$B and activating protein-1.** *Mol Carcinog* 2008, **47:**424–435.
33. Drafahl KA, McAndrew CW, Meyer AN, Haas M, Donoghue DJ: **The Receptor Tyrosine Kinase FGFR4 Negatively Regulates NF-kappaB Signaling.** *PLoS ONE* 2010, **5:**e14412.
34. Krishna R, Li CT, Buchanan-Wollaston V: **A temporal precedence based clustering method for gene expression microarray data.** *BMC Bioinf* 2010, **11:**68.
35. Witten DM, Tibshirani R, Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.** *Biostatistics* 2010, **10:**515–534.
36. Hardoon DR, Shawe-Taylor J: **Sparse canonical correlation analysis.** *Machine Learning* 2011, **83:**331–353.
37. Shi J, Malik J: **Normalized cuts and image segmentation.** *IEEE Trans Pattern Anal Machine Intelligence* 2000, **22:**888–905.

doi:10.1186/1752-0509-6-137
**Cite this article as:** Fujita *et al.*: Functional clustering of time series gene expression data by Granger causality. *BMC Systems Biology* 2012 **6**:137.