ABSTRACT

Motivation: Interaction among time series can be explored in many ways. All the approach has the usual problem of low power and high dimensional model. Here we attempted to build a causality network among a set of time series. The causality has been established by Granger causality, and then constructing the pathway has been implemented by finding the Minimal Spanning Tree within each connected component of the inferred network. False discovery rate measurement has been used to identify the most significant causalities.

Results: Simulation shows good convergence and accuracy of the algorithm. Robustness of the procedure has been demonstrated by applying the algorithm in a non-stationary time series setup. Application of the algorithm in a real dataset identified many causalities, with some overlap with previously known ones. Assembled network of the genes reveals features of the network that are common wisdom about naturally occurring networks.

Contact: nitai@lilly.com; chatterjee@stat.umn.edu

1 INTRODUCTION

Recent studies in cell cycle regulation experiments have resulted in considerable quantity of data on multiple gene expressions over time, see Cho et al. (1998), (2001) Spellman et al. (1998), Whitfield et al. (2002). Such data has lead to studies on several aspects of high dimensional time series analysis using gene expressions, like synchronization, periodicity, gene association, correlation and co-expression detection, clustering of genes and so on.

A fundamental objective of tracking gene expressions across time is to study causality. Informally, gene $G_1$ is a cause for gene $G_2$ if expression of $G_1$ is predictive of expression of $G_2$, possibly at a future time period. A more precise definition of causality, borrowed from economics literature, is presented in Section 2 for use in this paper. A direct causal relationship $G_1 \rightarrow G_2$ implies expression of gene $G_1$ predicts expression of gene $G_2$. An indirect causal relationship $G_1 \rightarrow G_{i_1} \cdots \rightarrow G_{i_k} \rightarrow G_2$ is a link from $G_1$ to $G_2$ through a sequence of direct causal relationships involving one or more intermediate genes $G_{i_1}, \ldots, G_{i_k}$. The above causality network may be represented by a graph $G = (V, E)$ whose vertices $V$ are the genes $G_{i_1}, \ldots, G_{i_k}$ under study, and a directed graph edge exists between $G_i$ and $G_j$ if expression of $G_i$ acts as a direct cause for expression of $G_j$.

This paper proposes a methodology for identifying causality and pathway map for gene expressions over time. We apply our pathway detection technique to the human cancer cell (HeLa S3) cycle data available at http://genome-www.stanford.edu/Human-CellCycle/Hela. Identifying causality and pathway of causality has several important ramifications. Causality is a study of interaction among genes, but unlike traditional interaction studies like association, correlation or clustering it establishes a directional pattern in which gene action may trigger/suppress and be triggered/suppressed by actions of other genes in the network. Since indirect causality is also identified by segments of the pathway, our method can be used to describe non-linear and complex dependence structures.

Our elicited pathway graph for the HeLa cell cycle data reveals several interesting features. The degree distribution of the graph imitates a power law, and there is also strong evidence of modularity of network or existence of hubs. This may be compared with several studies on metabolic pathways, robustness studies, complex network evolution in ecology, internet. See Jeong et al. (2000), Albert and Barabasi (2002), Ma and An-Ping (2003b), (2003a), Kitano (2004), Mason (2000), Verowerd (2006), Schuster et al. (2000) and Wagner and Fell (2001), http://arxiv.org/pdf/q-bio.MN/0604006 for related studies and more references.

Statistical research on the Hela cell cycle and similar high dimensional biological time series datasets has primarily focussed on periodicity and phase detection (Filkov et al., 2002; Wichert et al., 2004). A number of papers on detection of gene interaction and related topics use the correlation coefficient and its variations as a measure of interaction. Partial correlations, empirical Bayes and bootstrap methods are used in Schafer and Strimmer (2005) to obtain gene networks. In Zhu et al. (2005a) and (2005b) correlation is used as a primary tool for constructing network and pathways among genes and also for clustering genes. Correlation is an effective tool for computing direction free linear dependence when a sample of independent data is available.

In frequency domain time series analysis causality and interrelationship among the components is studied using coherence and partial coherence. Graphical models based on such analysis have been studied by Brillinger (1996) and Dahlhaus (2000) and applied to biological time series by Butte et al. (2001) and Salvador et al. (2005). However, Albo et al. (2004) showed that partial coherence based causality measures are sensitive to measurement noise.

In a commentary on Albo et al. (2004), Baccala and Sameshima (2006) argue that partial coherence is only able to detect the factor with the least amount of additive noise, and has nothing to do with time precedence, hence it is not a measure of causality. Winderhalter et al. (2005) report a detailed comparative study of techniques for directed interactions in multivariate time series. Their simulations show that Granger causality is excellent for
(1) detection of causalities and their directions in stationary time series, (2) detecting lack of relationship in independent series, (3) is capable of reproducing non-stationary interdependence structure, including abrupt changes, but (4) is sensitive to non-linearity, if not properly modeled. The partial directed coherence (Baccala and Sameshima, 2001), which attempts at combining properties of partial coherence and Granger causality in the frequency domain, is the closest in scope to Granger causality, but has a higher chance of detecting false relationships unless mathematical fine tunings are in place. Also, Kaminski et al. (2001) report that the frequency characteristics of the partial directed coherence are not consistent with that of the signal. They also show that the directed transfer function, another interdependence detection technique, can be interpreted in terms of Granger causality. Based on the above considerations, in this paper we adopt Granger causality as the gene interdependence detection tool.

In Section 2 we present in detail the methodology employed for causality detection and pathway search. Two small scale simulation experiments are reported in Section 3 to illustrate the performance of the proposed technique. The study of HeLa cell cycle data is described in Section 4. Discussion on some broader issues related to our algorithm, its limitations and open questions is available in Section 5.

2 METHODS

For two genes \( G_i \) and \( G_j \) with their corresponding weakly stationary time series \( G_i^t \) and \( G_j^t \), assume the following autoregressive model holds:

\[
G_i^t = \epsilon + \alpha_i G_i^{t-1} + \ldots + \alpha_{q_i} G_i^{t-q_i} + \beta_i G_j^{t-1} + \ldots + \beta_{q_j} G_j^{t-q_j} + \epsilon_t, \tag{1}
\]

where \( \epsilon \sim N(0, \sigma^2) \) and \( q \) is the autoregressive lag length. The gene \( G_j \) is said to Granger cause \( G_i \) if \( \beta_i \neq 0 \) for at least one \( i \). This is determined by an F-test of the null hypothesis

\[
H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0. \tag{2}
\]

When a higher number of genes \( G_1, \ldots, G_n \) are available, Granger causality may be tested using a vector autoregressive (VAR) framework. See Hamilton (1994) for further details on Granger causality and VAR.

In the present context the number of genes to consider is in several hundreds. In any microarray experiment this number generally runs into thousands, thereby making it impossible to accommodate the most general form of Granger causality test. Hence, we employ the simplified version here: we adopt the pairwise comparison study given by Equations (1) and (2) with \( q = 1 \). For each pair of genes \( G_i \) and \( G_j \) we test if \( G_i \) Granger causes \( G_j \) and vice versa. We assume that both these Granger causality relations cannot hold, hence we retain only the one with a lower \( p \)-value. This leads to \( \binom{p}{2} = p(p - 1)/2 \) \( p \)-values obtained from \( p(p - 1) \) tests, where \( p \) is the number of genes.

In considering hypotheses involving all possible pairs, a control for the number of false positives in the set of positive results has to be introduced. We use Benjamini and Hochberg (1995) false discovery rate (FDR) to control for false discoveries. Suppose the set of the \( p \)-values arranged in increasing order is \( p_{(1)}, p_{(2)}, \ldots, p_{(p)} \). Let \( k \) be the largest \( i \) such that \( p_{(i)} \leq \frac{q}{i} \). Then we accept that Granger causality exists [alternatively, reject the null hypothesis in Equation (2)] for all hypotheses where the \( p \)-value is less than or equal to \( p_{(i)} \). This ensures a false discovery rate of \( \rho_0 \).

Next, we construct a directed, weighted graph using the genes \( G_1, \ldots, G_p \) as the vertices, and with an edge of weight \( p_{ij} \) from \( G_i \) to \( G_j \) if the test for Granger causality of \( G_j \) on \( G_i \) has a \( p \)-value \( p_{ij} \). The previous steps ensure that no edge goes from \( G_i \) to \( G_j \) and \( p_{ij} > p_{(i)} \) above.

The graph constructed in the above fashion may have cycles in it. Also, it may have components that are not connected. In order to remove cycles and multiple paths between a pair of genes, we construct a minimal spanning tree from the above graph using Kruskal’s algorithm on each connected component. A graph that is a minimal spanning tree on each connected component is often referred to as a minimal spanning forest, hence our final outcome is a minimal spanning forest. Details of graph theoretic terminology and concepts may be found in Harary (1969).

Note that the presence or absence of Granger causality does not ensure the presence or absence of a causal relationship. Our approach is purely statistical, hence not guaranteed to ensure functional causality. However, our method proposes a very strong starting point for research in functional and metabolic relationship and biological causation models for temporal gene expressions. This is especially relevant given that modern studies involve thousands of genes with complex dependence structure, and in vitro study for all functional and metabolic relationship pathways is generally not feasible.

Certain statistical and biological assumptions on the \( p \)-dimensional microarray time series \( \{G_i^t, i = 1, \ldots, p\} \) are implicit in the above methodology. We assume the series is weakly stationary linear autoregressive for the Granger causality computations. In assuming the absence of bi-directional Granger causality and in constructing a minimal spanning forest, we exclude the possibility of existence of feedback cycles and multiple pathways of interaction between genes. Biological networks may have such cycles and multiple paths at various time lags. However, the reduced set of interactions available, after controlling for false discoveries and constructing the minimal spanning forest, provides a good starting point for an in vitro study. In applying statistical techniques, we make the tacit assumption that the available data is representative of the underlying biological process with probabilistically random noise terms.

The concept of Granger causality rests on the fact that predictability can be tested by determining whether one time series is related to past or current values of another time series, in addition to its own past values. In the present paper, a restrictive definition of Granger causality is employed assuming linear autoregression, possible dependence only on immediate past, with only bivariate relationships being tested. Such restrictive framework is in order to make computations feasible in available biological time series datasets, where typically number of genes are in thousands and time instances in tens.

3 SIMULATIONS

We first test our method of causality detection and pathway construction in two simple examples. A multidimensional time series process may be stationary or non-stationary. One of the most common causes of non-stationarity is the presence of trend or periodicity. We separately study our techniques for a small stationary and non-stationary network.

3.1 Stationary network

We simulate a network of 14 genes consisting of one complex and one simple causality relationship network. The two parts are disconnected, and represented in Figure 1. The independent genes in the network, namely \( x_1, x_7, x_8, x_9, x_{11}, x_{14} \) are all AR(1) process with autocorrelation \( < 1 \). The Granger caused time series \( x_2, x_3, x_6, x_4, x_5 \) and \( x_{10} \) are generated with one or more of the causing series, as indicated in Figure 1, all with a lag of 1 and autocorrelation \( < 1 \). The first column of Table 1 gives the formulae for the independent series for this experiment, while the third column gives the Granger caused series. All the series are generated for 100 equidistant time points, however keeping in mind the usual lack of multiple time points in genome scans, we apply our algorithm at time points \( t = 5, 10, 15, 20, 40, 60, 80 \) and 100. Note that edges present in Figure 1 represent direct causation only; and indirect causation can be easily read from it.
The above network is small, but contains several complexities that might help us assess performance of our strategy in realistic and difficult problems. For example, $x_1 \rightarrow x_3$, $x_3 \rightarrow x_4$ or $x_3 \rightarrow x_5$ are indirect causalities; there are disconnected components in the network; there is simultaneous causality of $(x_7-x_8)$ on $x_6$ and there is a parent node $(x_{11})$ with multiple children nodes $(x_{12}, x_{13})$.

In the above simulation, our algorithm follows the following steps:

1. For each pair of series, say, $G^i$ and $G^j$, test if $G^i$ Granger causes $G^j$ as well as $G^j$ Granger causes $G^i$.

2. Retain the causality with lower $P$-value if it is $< 0.01$. This corresponds to an adaptive choice of FDR. In real data, an overall fixed FDR is used instead.

3. Put all the significant causality together in one directed graph with all the genes as vertices and each significant causality as a directed edge. Identify each connected component of the constructed graph.

4. Construct the minimal spanning tree of each component with the corresponding $P$-values as the edge weights.

We replicate the simulation 100 times to study the general behavior of the constructed graph and the minimal spanning tree. Note that there are $\binom{14}{2} = 91$ possible edges to consider. The edges that join genes with direct or indirect causality relations have a direction component also, while the edges that join genes with no relation are directionless.

For each one of these 91 possible edges and each time point $t = 5, 10, 15, 20, 40, 60, 80$ and 100 when the data was analyzed, the ‘inclusion percentage’ is defined as the percentage of times an edge was included in the final minimal spanning tree. Since there were 100 replications of the experiment, this is simply the number of times each edge is included.

We plot the inclusion percentage over time to understand how our algorithm performs. This is displayed in Figure 2. Out of the 91 edges, some are present in the original graph Figure 1, and are denoted by black solid lines in Figure 2. Some of the edges join genes that only have an indirect causal relationship and hence do not have an edge in Figure 1. Such edges are denoted by grey dotted lines. Other edges that join genes that are not related directly or indirectly are in solid grey color. In each of the 100 replications of the experiments, if the direction of a true edge or a secondary edge was not picked correctly by the algorithm, they are considered undetected and do not contribute to the inclusion percentage.

It can be readily seen that overwhelming majority of the true ‘solid black’ edges from Figure 1 are selected by our algorithm; even for very short time runs of $t = 10$ or $t = 15$. By $t = 40$, all the true edges are widely separated from the other ones. By $t = 80$, a number of ‘secondary’ edges which denote indirect causation are separated out from the ‘grey’ edges which denote no relationship. At $t = 100$, the minimum inclusion percentage for a true edge is $\sim 40\%$, while the maximum inclusion percentage for a secondary edge is $\sim 10\%$. The inclusion percentage of wrong edges (solid grey) are $< 5\%$.

To numerically quantify the performance of our algorithm, we define a measure of accuracy as follows: Let $G^\text{true}$ be the complete graph where the true causalities are denoted by directed edges, as shown in Figure 1 in the present example. Let the elicited minimal spanning forest at time $t$ be denoted by $\hat{G}_t$. For each edge $e_i$, $i = 1, \ldots, p(p-1)$, define

$$l(e_i) = \begin{cases} 0 \quad &\text{if } e_i \in G^\text{true} \text{ and } e_i \in \hat{G}_t, \text{ directions consistent}, \\ 0 \quad &\text{if } e_i \not\in G^\text{true} \text{ and } e_i \not\in \hat{G}_t, \\ 1 \quad &\text{otherwise}. \end{cases}$$

Thus $l(e_i)$ is the loss at edge $e_i$ which is zero if $e_i$ appears with the same direction in $G^\text{true}$ and $\hat{G}_t$, or does not appear in either graph. The loss is one otherwise. Summing up over all possible edges, we define:

$$A(\hat{G}_t, G^\text{true}) = 100 \left( 1 - \frac{2}{p(p-1)} \sum_{e_i} l(e_i) \right)$$

as a measure of accuracy of estimating $G^\text{true}$ with $\hat{G}_t$.

In the first two rows of Table 2 we report the average $A(\hat{G}_t, G^\text{true})$ over the 100 replications and its standard error. It can be seen that $\sim 90\%$ of the edges, with direction if they are present in $G^\text{true}$ are correctly detected, with a standard error typically of the order of $0.2$. Since these experiments are a novelty in Statistics we do not have a benchmark to compare these figures, but they suggest that our algorithm works well for stationary data.

### 3.2 Non-stationary network

Real life time series data, including microarray time series data rarely have all components to be stationary. The cell cycle regulation translates to periodicity in gene expressions, which itself is a non-stationary phenomenon. This subsection is a study of the performance of our algorithm under non-stationary.

We consider the same 14 vertex graph with causality relations as depicted in Figure 1, but this time we add periodic deterministic trend components in the profile of genes 1, 8 and 11. The series $x_1$ now has a $\sin(\pi t/40)$ periodic trend, while $x_9$ and $x_11$ have $\cos(\pi t/40)$ periodic trend. The second column of Table 1 gives the formulae for the independent series for this experiment, while the third column gives the Granger caused series. The presence of non-stationarity in $x_1$, $x_9$ and $x_{11}$ results in all the Granger caused series to be non-stationary. We plot the inclusion percentage for one of these 100 times and compute the inclusion percentages and the accuracy measure $A(\hat{G}_t, G^\text{true})$ at different time points $t$. 

![Fig. 1. Causality set up of the simulation.](image-url)
Table 1. The actual causality relations used in simulation of stationary and non-stationary time series

<table>
<thead>
<tr>
<th>Independent series</th>
<th>Non-stationary</th>
<th>Dependent series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stationary</td>
<td></td>
<td>Granger caused</td>
</tr>
<tr>
<td>(x_{t1} = 0.7x_{1(t-1)} + \epsilon)</td>
<td>(x_{t1} = \sin \frac{t}{20} + 0.7x_{1(t-1)} + \epsilon)</td>
<td>(x_{20} = 0.29x_{2(t-1)} + 0.65x_{1(t-1)} + \epsilon)</td>
</tr>
<tr>
<td>(x_{t7} = 0.8x_{7(t-1)} + \epsilon)</td>
<td>(x_{t7} = 0.8x_{7(t-1)} + \epsilon)</td>
<td>(x_{21} = 0.15x_{3(t-1)} + 0.29x_{2(t-1)} + 0.65x_{4(t-1)} + \epsilon)</td>
</tr>
<tr>
<td>(x_{t8} = 0.7x_{8(t-1)} + \epsilon)</td>
<td>(x_{t8} = \cos \frac{t}{20} + 0.7x_{8(t-1)} + \epsilon)</td>
<td>(x_{20} = 0.12x_{8(t-1)} + 0.3x_{7(t-1)} + 0.3x_{6(t-1)} + \epsilon)</td>
</tr>
<tr>
<td>(x_{t9} = 0.77x_{9(t-1)} + \epsilon)</td>
<td>(x_{t9} = 0.77x_{9(t-1)} + \epsilon)</td>
<td>(x_{20} = 0.3x_{6(t-1)} + \epsilon)</td>
</tr>
<tr>
<td>(x_{t11} = 0.7x_{11(t-1)} + \epsilon)</td>
<td>(x_{t11} = \cos \frac{t}{20} + 0.7x_{11(t-1)} + \epsilon)</td>
<td>(x_{20} = 0.17x_{11(t-1)} + 0.4x_{3(t-1)} + 0.4x_{6(t-1)} + \epsilon)</td>
</tr>
<tr>
<td>(x_{t14} = 0.65x_{14(t-1)} + \epsilon)</td>
<td>(x_{t14} = 0.65x_{14(t-1)} + \epsilon)</td>
<td>(x_{20} = 0.6x_{14(t-1)} + 0.8x_{4(t-1)} + \epsilon)</td>
</tr>
</tbody>
</table>

The inclusion percentages are presented in Figure 3. This time the graphs are less satisfactory, although there is a general separation from the black (true) edges, and the grey dotted and solid grey ones. The true edges \(x_9 \rightarrow x_4\) and \(x_9 \rightarrow x_6\) generally sit among dotted or solid `grey’ edges. The edge \(x_{11} \rightarrow x_{10}\) has decreasing inclusion percentage with time. However, after \(t = 60\), all true edges have inclusion percentage \(\geq 20\%\), while certain spurious edges also have inclusion percentage between 20 and 30\%. Thus, our algorithm seem to err on the side of including some spurious edges in the non-stationary case.

The accuracy measure \(A(G_t, G_{true})\) for this experiment is reported in the last two rows of Table 2. This shows that even in the presence of trend, generally \(\geq 85\%\) edges are correctly detected. The important message from this experiment is that accuracy of \(G_t\) improves if de-trending is carried out, but not by a huge amount. De-trending leads to more significant improvement at high values of \(t\). However, currently available microarray time series data are of short lengths of time, with typical time lengths from \(t = 20\) to \(t = 60\). The improvement for high values of \(t\) may play an important role in future as longer run data becomes available.

Note that all the above analysis is with \(G_t\), a minimal spanning tree on disjoint components, or a minimal spanning forest. In reality, \(G_{true}\) need not be a tree or a forest, and may have cycles and multiple paths connecting the genes. We construct the minimal spanning forest for data reduction purposes, and it essentially represents the statistically strongest relationship structure.

Figure 4 present the originally elicited graph by our algorithm and the resulting minimal spanning forest for \(t = 100\) from the first simulation on the stationary network. The graph on the left is the elicited initial graph, while the one on the right is the minimal spanning forest. The corresponding figures of the originally elicited graph and the resulting minimal spanning forest from the first simulation on the non-stationary network are given in Figure 5. Similar graph at earlier time points are included in the supplementary material. Upon varying the coefficients and trend factors in Table 1, we found that a large autoregressive co-efficient improves chance of causality detection.

4 APPLICATION AND RESULTS

We apply our pathway detection technique to the human cancer cell (HeLa S3) cycle data available at http://genome-www.stanford.edu/Human-CellCycle/Hela. The data contains results from five different experiments involving three different cell synchronization method. Each experiment uses HeLa cell line arrested in S-phase. We focus on the first three experiments which use a common double thymidine block for cell synchronization and consist of 12, 26 and 48 time points respectively. A list of 1134 genes are identified as periodic or cell cycle regulators in Whitfield et al. (2002). From this list, only 802 are not missing for experiment 1, consequently we identify the pathway of causality only in this collection of 802 genes.

At time point \(t = 0\), the expression value is the average of periodic or cell cycle regulators in Whitfield et al. (2002). From this list, only 802 are not missing for experiment 1, consequently we identify the pathway of causality only in this collection of 802 genes. The actual causality relations used in simulation of stationary and non-stationary time series. The true edges are the wrong edges. Solid black edges represent the direct edges, dotted grey are the secondary edges and solid grey are the wrong edges.

The inclusion percentages are presented in Figure 3. This time the graphs are less satisfactory, although there is a general separation from the black (true) edges, and the grey dotted and solid grey ones. The true edges \(x_9 \rightarrow x_4\) and \(x_9 \rightarrow x_6\) generally sit among dotted or solid ‘grey’ edges. The edge \(x_{11} \rightarrow x_{10}\) has decreasing inclusion percentage with time. However, after \(t = 60\), all true edges have inclusion percentage \(\geq 20\%\), while certain spurious edges also have inclusion percentage between 20 and 30\%. Thus, our algorithm seem to err on the side of including some spurious edges in the non-stationary case.

The accuracy measure \(A(G_t, G_{true})\) for this experiment is reported in the last two rows of Table 2. This shows that even in the presence of trend, generally \(\geq 85\%\) edges are correctly detected. The important message from this experiment is that accuracy of \(G_t\) improves if de-trending is carried out, but not by a huge amount. De-trending leads to more significant improvement at high values of \(t\). However, currently available microarray time series data are of short lengths of time, with typical time lengths from \(t = 20\) to \(t = 60\). The improvement for high values of \(t\) may play an important role in future as longer run data becomes available.

Note that all the above analysis is with \(G_t\), a minimal spanning tree on disjoint components, or a minimal spanning forest. In reality, \(G_{true}\) need not be a tree or a forest, and may have cycles and multiple paths connecting the genes. We construct the minimal spanning forest for data reduction purposes, and it essentially represents the statistically strongest relationship structure.

Figure 4 present the originally elicited graph by our algorithm and the resulting minimal spanning forest for \(t = 100\) from the first simulation on the stationary network. The graph on the left is the elicited initial graph, while the one on the right is the minimal spanning forest. The corresponding figures of the originally elicited graph and the resulting minimal spanning forest from the first simulation on the non-stationary network are given in Figure 5. Similar graph at earlier time points are included in the supplementary material. Upon varying the coefficients and trend factors in Table 1, we found that a large autoregressive co-efficient improves chance of causality detection.

4 APPLICATION AND RESULTS

We apply our pathway detection technique to the human cancer cell (HeLa S3) cycle data available at http://genome-www.stanford.edu/Human-CellCycle/Hela. The data contains results from five different experiments involving three different cell synchronization method. Each experiment uses HeLa cell line arrested in S-phase. We focus on the first three experiments which use a common double thymidine block for cell synchronization and consist of 12, 26 and 48 time points respectively. A list of 1134 genes are identified as periodic or cell cycle regulators in Whitfield et al. (2002). From this list, only 802 are not missing for experiment 1, consequently we identify the pathway of causality only in this collection of 802 genes. At time point \(t = 0\), the expression value is the average of the same measurement obtained from two biological replicates. There is considerable overlap between the cell cycle regulating genes obtained by Whitfield et al. (2002) and by Schafer and Strimmer (2005), consequently we expect our results to be reasonably robust to periodicity detection techniques.

We do not attempt trend removal because of the short length of the data and the wide variability of the periodicity of the cell division cycle. We concentrate on experiment one for obtaining the causality relationships, and then follow up the significant causalities in the other two experiments.

In each experiment, the last few time points are spaced at irregular intervals. The influence of each time point in the final results is low owing to the large sample nature of the entire statistical procedure,
hence we pretend as if all the time points are equally spaced. Correcting for unequally spaced time points is mathematically cumbersome, and in the present problem leads to no significant gain. However, this might be a necessary correction for other problems, depending on the inequality of spacings of the times at which the microarrays were visited.

As described in previous sections, Granger causality tests were performed between all pairs among the 802 genes. From the two tests between each pair of genes, only the one with lower $P$-value is retained as a plausible causality case. After that, thresholds for $P$-values are determined by the Benjamini–Hochberg FDR implementation. The results we obtain are quite sensitive to the FDR threshold used. The distribution of the size of connected components in the elicited graph is presented in Table 3 for two different FDR cutoff values.

The threshold of 0.25 and 0.30 on FDR translates to a threshold of 0.0000099 and 0.00095 on $P$-values. The extremely restrictive threshold of FDR = 0.25 returns only 11 components with only 1 involving three genes and 10 others involving two each. We retain these results since these appear to be the strongest causation relations, and might deserve in vitro study for functional relationship. In Figure 6 we present these networks.

In Figure 7 we display the time courses of the 12 pairs of ‘causing-caused’ gene pairs depicted in Figure 6. The paths from Experiment 1 are in thin black, those from Experiment 2 are in bold black and those from Experiment 3 are in bold grey. We artificially shifted the paths from Experiments 2 and 3 by 1 and 2 units respectively, to increase visibility. In each figure the ‘causing’ gene is the solid curve, while the ‘caused’ gene is represented by the dotted line.

The $P$-values of Granger causality tests on data from Experiments 2 and 3 are displayed alongside. Note that several interesting causality patterns are evident in these graphical displays. We see cases of overexpression of one gene leading to overexpression of another (e.g. 144-HMGE $\rightarrow$ 114-FKBP1A), as well as overexpression of one causing diminished expression of the other (e.g., 737-NS1-BP $\rightarrow$ 629-CDC25B). Time lag effects are also evident in Figure 7.

At 30% threshold on FDR we obtained 10 pairs of isolated two gene components, 2 groups each of size 3 and 1 network each of size 4, 6, 7, 8 and 575. Once a small connected component is identified by Granger causality, we can embark on a full VAR analysis of it. When large datasets are available, a VAR analysis is better than a pairwise Granger causality test, since there the relationship between a pair of genes are studied in the presence of other genes, hence complex relationships are more likely to be revealed. However, VAR modeling of a connected network of $q$ genes requires...
the minimal spanning tree is computed, the degree distribution obtains that gene 280-ZNF265, with a degree of 17, is the single most connected gene, while 30-NXF1, 95-ZNF42, 594-KIAA0135 of degree 10 each and several ESTs have high degree too. Figure 9 shows the biggest module around 280-ZNF265. Other smaller modules in the network are provided in the website.

Existing gene–gene interaction studies are typically on a few pre-selected genes. For example, in Li et al. (2006) gene regulatory networks are obtained on a subset of 20 genes using the third experiment from the above HeLa cell cycle data. Eighteen of these 20 are included in the 802 genes we studied. A close look at the interactions between these 18 genes reveals interesting structures.

Our analysis is graphically represented in Figure 10. No interaction was detected for genes CDC25A, BRCA1 and TYMS. We obtain several edges emanating from CCNE1, CCNF and CCNA2; several edges ending at and starting from STK15, E2F1, NPAT, BUB1B; while several edges terminate at CDC20, PLK, CKS2 and PCNA. This structure suggests the presence of a ‘hub-and-spoke’ structure in the network of genes, with STK15 and E2F1 being hubs in the hub. The regulation of genes CCNF, CCNE1, CDC20 and PLK have been found significant in several in vitro and in vivo experiments, see additional file 4 of Li et al. (2006) for some documentation. We find that CCNF and CCNE1 are strong ‘regulating’ genes, STK15 and E2F1 are intermediate causes that serve as traffic hubs, while CDC20 and PLK are important ‘regulated’ genes. Such structure may help, for example, in controlling activity in cancerous cell cycles. The entire gene network is effectively broken down if the expression of CCNF, STK15 or CDC20 can be controlled.

Among these 18 genes, we obtain a few interesting gene–gene relations that have been documented earlier. The regulation of PLK by STK15 and CCNF has been documented in Li et al. (2006). We obtain that PLK is ‘caused’ by STK15, CCNF, CCNA2 and CCNE1, and STK15 is caused by CCNF. The lack of interaction for BRCA1 is corroborated in the studies of Li et al. (2006) and Whitfield et al. (2006). The gene CCNB1 has high periodicity score, hence is thought of as an important cell cycle regulatory gene. Both Li et al. (2006) and Whitfield et al. (2002) failed to assign its relationship with other genes. We obtain that CCNB1 is regulated by STK15. We corroborate the finding of Li et al. (2006) that STK15 regulates CDC20. However, unlike Li et al. (2006), we did not find a relationship in which CDC20 regulates other genes at different lags of time units. This points to two important limitations with our present algorithm: we do not allow for bi-directional and cyclical dependencies, and we only study causality of one time unit lag. Bi-directional and cyclical relationships and causality at lags of several time units can be studied when more data is available. However, our present algorithm obtains a network involving 802 genes, with several previously undiscovered gene–gene relations. Our method also eliminates the need for pre-selection of a handful of genes for study.

5 DISCUSSION AND CONCLUSION

Prior research on naturally occurring networks suggest an evolutionary development of these networks from procaryotic organisms to higher level animals. In view of this postulate, it is of interest to
infer about the level of interaction within a small set of genes. This paper is on development of methodology to address this issue. Pathway detection in microarray time series experiments are important for understanding genetic interaction and causation. Microarray experiments involve several factors where chance fluctuations and random processes play a significant role. This is unlike the traditional framework of biochemistry where metabolic pathways are commonly constructed from known reactions among metabolites and enzymes. We account for the randomness using time series techniques here. Our analysis also departs from the tradition of using correlation as a measure of relationship between genes. Apart from the limitation of correlation as a measure, its
computation requires independent observations while microarray time series are dependent among themselves.

The pairwise Granger causality studies are put together in a graph. Some of the steps of our algorithm are intended to reduce the number of edges in the graph, so that major relationships are highlighted. Removal of cyclic trend, if the period is known, is recommended. However in microarray time series, that is generally not feasible, since data on only few time points are available, and periodicity seems to be gene specific.

It must be emphasized that Granger causality does not imply, nor is implied by functional causality, but is merely indicative of it. For a rich discussion on the philosophical and other angles on causality in economics, see Engle and White (1999). Our application of Granger causality further assumes that the time points are equi-spaced, which may not be the case in microarray experiments. However, this assumption can be removed easily with a slightly different formulation.

One concern, shared by all microarray related methods, is the dependence among the multiple comparisons carried out. This is addressed by using a standard false discovery rate control technique here. The available data limits the extent of Granger causality tests we can do, nevertheless a large number of tests among related objects need to be carried out. A possible approach to reduce the huge number of tests can be using cluster analysis first and then search for causality within the tightly linked clusters. Another approach is to incorporate prior knowledge about gene relationships in a Bayesian causal formulation. These issues deserve further study.

ACKNOWLEDGEMENTS

We benefitted from a discussion with Professor Sounak Mishra about construction of minimal spanning trees. The detailed feedback from two anonymous reviewers helped in revising the manuscript. The research of the second author is partially supported by a grant from the University of Minnesota.

Conflict of Interest: none declared.

REFERENCES