*bioengineering*

MDPI

*Article*

# Stable Gene Regulatory Network Modeling From Steady-State Data †

**Joy Edward Larvie [1], Mohammad Gorji Sefidmazgi [1,‡], Abdollah Homaifar [1,*], Scott H. Harrison [2], Ali Karimoddini [1] and Anthony Guiseppi-Elie [3]**

[1] Department of Electrical and Computer Engineering, North Carolina A&T State University, 1601 E. Market Street, Greensboro, NC 27411, USA; jelarvie@aggies.ncat.edu (J.E.L.); mgorjise@email.arizona.edu (M.G.S.); akarimod@ncat.edu (A.K.)

[2] Department of Biology, North Carolina A&T State University, 1601 E. Market Street, Greensboro, NC 27411, USA; scotth@ncat.edu

[3] Department of Biomedical Engineering, Texas A&M University, 5045 ETB, College Station, TX 77843, USA; guiseppi@tamu.edu

* Correspondence: homaifar@ncat.edu; Tel.: +1-336-285-3271; Fax: +1-336-334-7716

† This paper is an extended version of our paper published in 41st Annual Northeast Biomedical Engineering Conference (NEBEC).

‡ Current affiliation: School of Information, The University of Arizona, 1103 E. 2nd St, Tucson, AZ 85721, USA

Academic Editor: Aldo R. Boccaccini

Received: 19 November 2015; Accepted: 6 April 2016; Published: 19 April 2016

**Abstract:** Gene regulatory networks represent an abstract mapping of gene regulations in living cells. They aim to capture dependencies among molecular entities such as transcription factors, proteins and metabolites. In most applications, the regulatory network structure is unknown, and has to be reverse engineered from experimental data consisting of expression levels of the genes usually measured as messenger RNA concentrations in microarray experiments. Steady-state gene expression data are obtained from measurements of the variations in expression activity following the application of small perturbations to equilibrium states in genetic perturbation experiments. In this paper, the least absolute shrinkage and selection operator-vector autoregressive (LASSO-VAR) originally proposed for the analysis of economic time series data is adapted to include a stability constraint for the recovery of a sparse and stable regulatory network that describes data obtained from noisy perturbation experiments. The approach is applied to real experimental data obtained for the SOS pathway in *Escherichia coli* and the cell cycle pathway for yeast *Saccharomyces cerevisiae*. Significant features of this method are the ability to recover networks without inputting prior knowledge of the network topology, and the ability to be efficiently applied to large scale networks due to the convex nature of the method.

**Keywords:** gene regulatory network; reverse engineering; sparse network; stable network; convexity

## 1. Introduction

A number of technological advances, such as oligonucleotide arrays, serial analysis of gene expression (SAGE) and cDNA microarrays [1], have enabled biomedical researchers to expeditiously and simultaneously collect large amounts of metabolomic, transcriptomic, proteomic data [2] in a single experiment, providing a wealth of information for elucidating gene regulation, functions and interactions [3,4]. Over time, repositories such as the Gene Expression Omnibus (GEO) [5] and the Biological General Repository for Interaction Datasets (BioGRID) [6] are mapping functional information and ontologies to expression data sets [2,7]. Gene expression measurement data acquired from microarray experiments typically occur in two contexts: steady-state data which provides

information on interaction directions, and temporal data that allows for the investigation of temporal patterns in biological networks [8,9].

Owing to their inherent ability to encapsulate the high dimensional data of biological processes and pathways, networks have become an important tool in functional genomics [10,11]. Researchers refer to any such network that provides a system level interaction among genes as a gene regulatory network (GRN) [12,13]. GRNs are usually represented by directed graphs with nodes as genes, and edges depicting either an inhibition (negative regulation) or an activation (positive regulation) imposed by a gene over another through the production of a protein [14,15].

The process of identifying genetic interactions from measured gene expression data is referred to as reverse engineering or network inference or recovery [7]. Inferring the topology of GRNs and isolating functional subnetworks are computationally challenging tasks in contemporary functional genomics, and these efforts are valuable for advancing scientific insight and for capitalizing on the time and costs associated with experimental data [16–19]. GRNs typically contain information about the pathway to which a gene belongs and the genes it interacts with [16], and this helps to reveal potential pathway initiators and drug targets [8]. Further analysis, to map interactions among phenotypic and genotypic characteristics, can provide a framework for the identification of biomarkers for medical diagnosis and prognosis [20,21].

A plethora of modeling approaches such as co-expression clustering [22], Boolean network [23,24], Bayesian network [25] and ordinary differential equation (ODE) [8] models have been proposed for recovering genetic networks. Cluster analysis and the sequential search for patterns of gene expression related with some pathological state of interest usually provide only indirect information about the structure of the network [7]. Alternatively, grouping of co-expressed genes may be achieved using information-theoretic methods. Both approaches, however, lack causality [9]. Causality may be recovered through Bayesian networks which can handle directed graphs [9,26]. However, Bayesian networks typically do not accommodate cycles, and, hence, are unable to handle feedback motifs that are common in gene regulatory networks [26]. Causality and feedback motifs, however, are no longer a problem when the network is modeled as a set of differential equations [26]. Excellent as they are at modeling causality and feedback motifs, differential equations are only suitable for small-scale networks [9].

These existing techniques, however, rely heavily on temporal expression data which can be very difficult to acquire, and also require high computational effort [8,26]. Major considerations of sparsity, stability and causality must be captured in the biological network recovery process [2]. In this paper, the least absolute shrinkage and selection operator-vector autoregressive (LASSO-VAR) model, originally proposed for the analysis of economic time series data in [27], is adapted to include a stability constraint defined and used by [26] for the recovery of sparse and stable regulatory networks that describe steady-state data obtained from noisy perturbation experiments. The fact that LASSO-VAR is a vector autoregressive process implies that Granger causality can be inferred. The technique only requires one tuning parameter, which works to penalize non-sparse networks. The selection of this parameter is based on its mean square forecast error. The identification algorithm proposed is applicable for the identification of regulatory roles of individual genes and control genes in the network. It is also applicable for identifying genes that directly impact the bioactivity of a compound in the cell. The approach is applied to real experimental data obtained for the SOS pathway in *Escherichia coli* and the cell cycle pathway for yeast *Saccharomyces cerevisiae*. The significant features of this method are the ability to recover networks without *a priori* knowledge of the network topology, and to be efficiently applied to large scale networks due to the convex nature of the method.

## 2. Methodology

This section introduces the stable LASSO-VAR, the identification technique being adapted for reverse engineering gene regulatory networks from steady-state data [28]. In its original form, the LASSO-VAR technique described in [27] finds applications in the analysis and prediction of economic

and financial time series. It is an extension of the VAR model to include a selection and shrinkage operator known as the LASSO. The inherent advantages of the LASSO-VAR are the ability to perform dimension reduction and variable selection, as well as being able to test Granger causality. For this reason, this paper adapts the LASSO-VAR concept and incorporates a stability constraint as a convex constraint to allow for the inference of a stable genetic network from steady-state data.

### *2.1. Network Identification Approach*

The vector autoregressive (VAR) model is known to be one of the most flexible and easy to use models for analyzing multivariate time series [27]. It has found applications in neurosciences for the estimation of functional connectivity between several brain areas [29], and most recently in system biology for the reconstruction of gene regulatory networks [2].

In the general case, an $N$-dimensional multiple time series gene expression data $y_1, ..., y_T$ with $y_t = (y_{1t}, ..., y_{Nt})'$ can be assumed to be generated by a stationary, stable VAR($p$) process as [30]:

$$y_t = v + A_1 y_{t-1} + ... + A_p y_{t-p} + u_t, \tag{1}$$

where $p$ denotes the order of the vector autoregressive process (*i.e.*, the vector autoregressive lag length), $y_t$ is an $(N \times 1)$ random vector, $A_i$ is a fixed $(N \times N)$ coefficient matrix, $v = (v_1, ..., v_N)'$ is a fixed $(N \times 1)$ vector of intercept terms allowing for the possibility of a nonzero mean $E(y_t)$, $u_t = (u_{1t}, ..., u_{Nt})'$ is a $N$-dimensional white noise, thus, $E(u_t) = 0$, $E(u_t u_t') = \Sigma_u$ and $E(u_t u_s') = 0$ for $s \neq t$. The covariance matrix $\Sigma_u$ is assumed to be nonsingular. The framework of the general VAR($p$) allows for the testing of Granger causality [31]. The concept of Granger causality is founded on the idea that a cause must precede an effect. This concept was originally proposed by Granger in [32].

In the present context, however, the number of genes considered in most microarray experiments generally runs from several thousands to millions, thereby making it impossible to accommodate the most general form of the Granger causality test [31]. Thus, a VAR(1) (VAR of order one) model is usually employed to allow for a pairwise comparison study as seen in [31] and [29]. Stated simply, in the case of a VAR(1), if gene $b$ at time $t$ is affected by a gene $a$ at time $(t-1)$, the latter should help to predict the target gene expression [29].

A first order VAR model is defined as [30]:

$$y_t = v + A y_{t-1} + u_t. \tag{2}$$

For convenience, (2) is usually expressed in compact matrix notation as [30]:

$$\mathbf{Y} = \mathbf{v} + \mathbf{AZ} + \mathbf{U}, \tag{3}$$

where $\mathbf{Y} = (y_1, y_2, ..., y_T)$ is an $N \times T$ data or response matrix, $\mathbf{A}$ is an $N \times N$ unknown coefficient matrix, $\mathbf{Z} = (Z_0, ..., Z_{T-1})$ is an $N \times T$ covariate matrix and $\mathbf{U} = (u_1, ..., u_T)$ is an $N \times T$ matrix.

The solution to (3) is given as follows [30]:

$$A = ((ZZ^T)^{-1} Z \otimes I_N) Y, \tag{4}$$

where $I_N$ is an $N \times N$ identity matrix, and $\otimes$ is the Kronecker product or direct product.

In high dimensional space, the VAR processes become computationally intractable [27]. As such, the model usually contains unwanted parameters which leads to less efficient parameter estimates [33]. The LASSO-VAR model addresses the intractability issue by zeroing some elements of the coefficient matrix, which removes unnecessary variables [27,33]. The requirement that the coefficient matrix, $A$, be sparse is due to the loose connectivity that biological networks generally exhibit [8,26].

This requirement is addressed by applying an $L_1$ penalty to the convex least squares objective function, resulting in [27]:

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{v} - \boldsymbol{AZ}\|_F^2 + \lambda\|\boldsymbol{A}\|_1, \tag{5}$$

where $\|\mathbf{X}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}|^2$ is the square of the Frobenius norm of $\mathbf{X}$ (*i.e.*, the sum of the absolute squares of its elements), $\|\mathbf{X}\|_1 = \Sigma_{jk}|X_{jk}|$ is the sum of the absolute values of $\mathbf{X}$, and $\lambda \geq 0$ is a penalty parameter.

According to [30], if all eigenvalues of the coefficient matrix $A$ of the VAR(1) process have absolute values less than 1, the sequence $A^i, i = 0, 1, ...$ is absolutely summable; as such, the infinite sum $\sum_{i=1}^{\infty}$ exists in mean square. Hence, in general, a VAR(1) is said to be stable $iff$ all eigenvalues of $A$ have absolute value less than one. It is mathematically equivalent to [30]:

$$\mathbf{det}(\boldsymbol{I_N} - A\tau) \neq 0 \text{ for } |\tau| \leq 1. \tag{6}$$

The original formulation of the LASSO-VAR technique by [27] lacks the ability to infer a stable network. This setback means that stability cannot be inferred from gene perturbation experiments. To solve this inadequacy, a stability constraint that relies on the theorem by Geršgorin as discussed in [34] is incorporated into the LASSO-VAR objective function (5).

Geršgorin's theorem states that all the eigenvalues of an $n \times n$ matrix $A = (a_{ij})$ are in the union of the discs whose boundaries are circles $C_1, C_2, ..., C_n$ with centers at the points $a_{11}, a_{22}, ..., a_{nn}$ and the radii are: $r_i = \sum_{j=1, j\neq i}^{n}|a_{ij}|$. Stated more compactly, every eigenvalue $\tau$ must be contained in at least one of the circles characterized by the rows of $A$ for an $n \times n$ matrix $A$. In essence, the eigenvalues of a square matrix can not be too far from its diagonal entries. It follows that [34]:

$$|\tau - a_{ii}| \leq \sum_{j=1, j\neq i}^{n}|a_{ij}|, \; i = 1, 2, ..., n. \tag{7}$$

As such, the real part of each eigenvalue must satisfy one of the conditions [34]

$$\mathrm{Re}[\tau] \leq a_{ii} + \sum_{j=1, j\neq i}^{n}|a_{ij}|, \quad i = 1, 2, ..., n. \tag{8}$$

Since $V^{-1}AV$ and $A$ have the same eigenvalues for all invertible matrix $V$, it is possible to apply Geršgorin's theorem to $V^{-1}AV$. For a good choice of $V$, one can find some tighter bounds for the eigenvalues. [26]. A particularly convenient choice is $V \triangleq \mathrm{diag}(v_1, ..., v_n)$, with $v_i > 0$ for all $i = 1, ..., n$. Then, $V^{-1}AV = (v_j a_{ij}/v_i)$. It follows therefore that, $\forall V \in \mathbf{V}$, the real part of an eigenvalue of $A$ must satisfy [26]:

$$\mathrm{Re}[\tau] \leq a_{ii} + \sum_{j=1, j\neq i}^{n}\frac{v_j}{v_i}|a_{ij}|, \quad i = 1, 2, ..., n. \tag{9}$$

The stability requirement of the algorithm stems from the steady-state nature of the gene expression data adopted. Stability of the network simply refers to the robustness of the network to topology and parameter changes, as well as instrumental and biological noise [35]. The inherent stability description of the VAR model therefore allows the incorporation of a stability constraint that helps to address the specification of a stable gene regulatory network from steady-state data.

Incorporating this concept as a constraint retains the convex nature of the objective function; hence, it has the associated properties of scalability and global optimality. The resulting overall optimization problem is given as:

$$\begin{aligned} &\textbf{minimize } \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{AZ}\|_F^2 \\ &\textbf{subject to } \|\boldsymbol{A}\|_1 < \lambda,\ 0 \le \lambda \le 1 \\ &\qquad a_{ii} \le -\sum_{j=1, j\neq i}^{n} \frac{v_j}{v_i}|a_{ij}|,\ i = 1, ..., n,\ v_i,\ v_j > 0 \end{aligned} \quad . \tag{10}$$

The choice of $v_i$ is dependent on the stability requirements defined in the problem formulation. Zavlanos *et al.* [26] provide a convenient way to choose the weights $v_i$.

Define the deleted absolute sum for row $i$ as $R_i(A) \triangleq \sum_{j\neq i} |a_{ij}|$. Then, for $\beta \triangleq \frac{1}{n}\sum_{i=1}^{n}(|a_{ii}| - R_i(A))$ the weights $v_i$ are chosen given Figure 1 as follows [26]:

$$v_i \triangleq \begin{cases} 1 + \dfrac{|a_{ii}| - R_i(A) - \beta}{\delta + (|a_{ii}| - R_i(A) - \beta)}, & \text{if } |a_{ii}| - R_i(A) > \beta \\ \dfrac{\delta}{\delta - (|a_{ii}| - R_i(A) - \beta)}, & \text{if } |a_{ii}| - R_i(A) \le \beta \end{cases} . \tag{11}$$

Solving the constrained optimization problem in (10) iteratively yields a sparse, stable coefficient matrix that models the causal interactions among the genes under observation as desired.

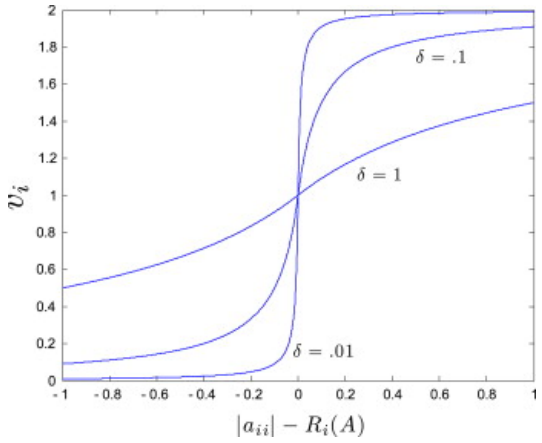



**Figure 1.** Plot of $v_i$ as a function of the entries $|a_{ii}| - R_i(A)$, for average $\beta = 0$ and different values of the parameter $0 < \delta \le 1$. Taken from [26].

In shrinkage problems, a formalized approach for the selection of an optimal penalty parameter value is achieved by employing either a $k$-fold or a leave-one-out cross validation [27]. Due to time-dependence, however, traditional cross-validation techniques are not well-suited for the problem formulation. The optimal penalty parameter is selected by minimizing the one-step ahead mean square forecast error (MSFE) [27]. This process starts by dividing the data into three periods: one for initialization, one for training, and one for forecast evaluation. Two time indices are defined as: $T_1 = \lfloor \frac{T}{3} \rfloor$, and $T_2 = \lfloor \frac{2T}{3} \rfloor$, where $\lfloor x \rfloor$ is the largest integer less than or equal to $x$ . The period $T_1 + 1$ through $T_2$ is used for training and $T_2 + 1$ through $T$ for the evaluation of forecast accuracy in a rolling

manner. In addition, for the one-step ahead forecast based on all observations from $1, ..., t$ is defined as $\hat{y}^{\lambda}_{t+1}$ [27]. The objective therefore is to minimize [27]:

$$MSFE(\lambda) = \frac{1}{T_2 - T_1 - 1} \sum_{t=T_1}^{T_2-1} \|\hat{y}^{\lambda}_{t+1} - y_{t+1}\|_F^2. \tag{12}$$

MSFE represents the most appropriate criterion given the use of the least squares objective function. Instead of parallelizing the cross-validation procedure, this approach uses the result from the previous period as an initialization, substantially reducing computational time [27].

The algorithm was implemented in MATLAB® R2014a using the `cvx 2.1` toolbox for convex optimization problems [36].

### 2.2. Performance Evaluation

In order to evaluate the performance of the proposed network identification algorithm for reconstructing stable gene regulatory networks from datasets, statistical measures are employed. For predictive analysis, the confusion matrix (Table 1), represents a table with two rows and two columns that report the number of True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs). Since the parameter, $\lambda$, regulates the weight imposed on sparsity, the terms "Positives" and "Negatives" here refer to non-zero and zero interactions between genes, respectively.

**Table 1.** Confusion matrix.

| | | **True Network** | **Total** |
|---|---|---|---|
| Inferred Network | True Positive | False Postive | **P′** |
| | False Negative | True Negative | **N′** |
| Total | **P** | **N** | |

TP represents the interaction that exists in both the true network and inferred network, FP denotes the interaction that does not exist in the true network but was falsely inferred, TN is the interaction that does not exist in either the true network or the inferred network, while FN represents the interaction that does exist in the actual network but is not recovered by the network identification method.

Three other criteria *Sensitivity* (sen), *Specificity* (spc) and *Precision* (not commonly used) are also employed as evaluation methods. *Sensitivity* (sen), is the fraction of the number of recovered true regulations to all regulations in the model. *Specificity* (spc), is the ratio of correctly found no-interactions to all no-interactions in the model. *Precision* (pre), measures the fraction of the number of correctly found regulations to all found regulations in the inferred network. These three performance criteria are defined as follows [37]:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{13a}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{13b}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{13c}$$

For the purposes of comparison with other network identification techniques, the performance evaluation graphs are restricted to sensitivity and specificity.

Overall, the following specific steps are performed for modeling the GRN using LASSO-VAR. The time series of gene expressions are converted to a matrix where the rows are expression of various genes and the columns are observations at different time points. Assuming the model of Equation (2), the optimization problem as Equation (10) can be generated. Solving this optimization

using the iterative approach yields the matrix A with sparse and stable structure. The optimal values of hyperparameters ( $\lambda$ and $v_i$ ) should also be found. The structure of gene regulatory network is found using the matrix A. The results of our method are then compared with target GRNs that are accepted or inferred by other means.

Data for GRN recovery of the *Escherichia coli* SOS network was from a perturbation experiment for relative RNA expression changes from Table S6 of [8]. Data for GRN recovery of the cell cycle pathway in yeast *Saccharomyces cerevisiae* was based on the alpha time series of [38], including 18 time points at 7 min interval over 119 min. Data came from the yeast cell cycle analysis database [39] with the analysis conducted on a set of 14 genes. Standard (and systematic) gene names for these genes are: FUS3 (YBL016W), SIC1 (YLR079W), FAR1 (YJL157C), CDC6 (YJL194W), CDC20 (YGL116W), CDC28 (YBR160W), CLN1 (YMR199W), CLN2 (YPL256C), CLN3 (YAL040C), CLB5 (YPR120C), CLB6 (YGR109C), SWI4 (YER111C), SWI6 (YLR182W) and MBP1 (YDL056W).

## 3. Results and Discussion

In this section, the efficiency of the proposed network identification is analyzed by studying networks for which the experimental data as well as the ground truth is available. The studied datasets consist of real experimental dataset in a known subnetwork of the SOS pathway in *Escherichia coli*, provided in [26], and the cell cycle pathway in yeast *Saccharomyces cerevisiae*.

With datasets that have a known network, it is possible to evaluate the performance of the algorithm, allowing for the measurement of the false positives, false negatives, *etc*. The effect of different values of the penalty parameter $\lambda$ on the performance of the algorithm is also investigated.

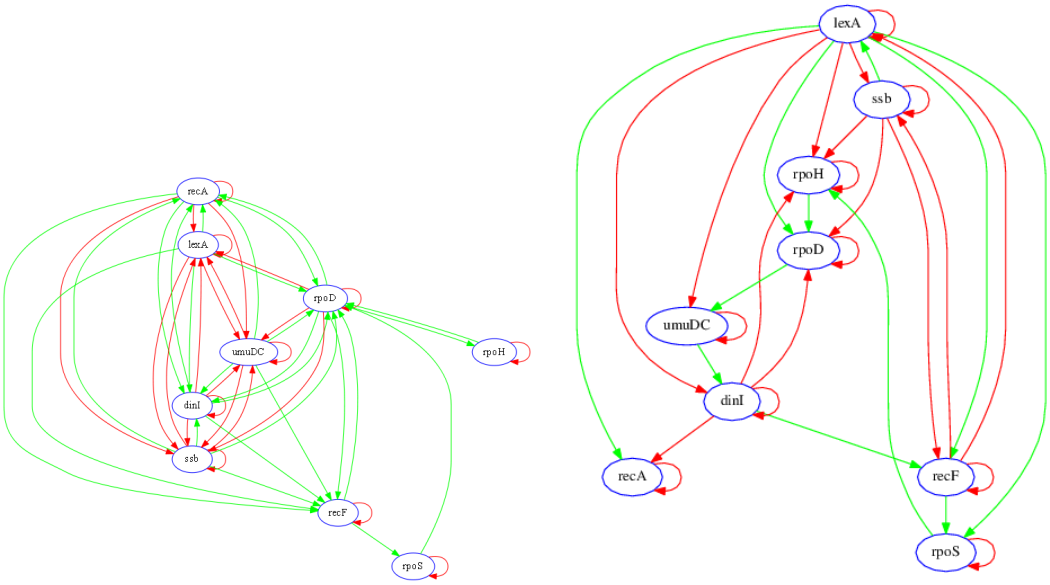
### *3.1. SOS Pathway in* Escherichia coli

The proposed identification algorithms is first applied to a sub-network of the SOS pathway in *Escherichia coli*, using the gene perturbation experimental data set provided in [8]. The SOS pathway is known to control the survival and repair of cells after DNA damage [8]. This pathway typically involves the genes *recA* and *lexA* directly regulating over 30 genes, and indirectly controlling over 100 genes [8].

The sub-network considered, shown in Figure 2a, consists of nine genes and several transcription factors and metabolites [8] whose expression levels are measured over nine different perturbations. According to Gardner *et al.* [8], the nine transcripts in the test network (Figure 2a) were chosen to enable evaluation of the performance of their proposed algorithm. These nine transcripts include the key mediators of the SOS response (*lexA* and *recA*) and sets of genes with known regulatory roles (*ssb*, *recF*, *dinI*, and *umuDC*) and unknown regulatory roles (*rpoD*, *rpoH*, and *rpoS*). The presence of genes with regulatory roles that are already known allows this network to be used to validate an inference algorithm [8].

#### 3.1.1. Network Recovery

In order to evaluate the performance of the proposed algorithm, those links that are correctly recovered in the model are determined based on knowledge of the true network. An inferred connection is regarded to be accurate if there exists a known RNA, protein, or metabolite pathway between the two transcripts, and if the sign of the net effect of regulatory interaction (*i.e.*, inhibition or activation) is correct, as determined by the known network in Figure 2a. In general, since RNA concentrations (*i.e.*, expressions) were measured and not metabolite or protein species, the recovered regulatory network model does not necessarily depict physical connections; rather, the links show effective functional associations between transcripts.



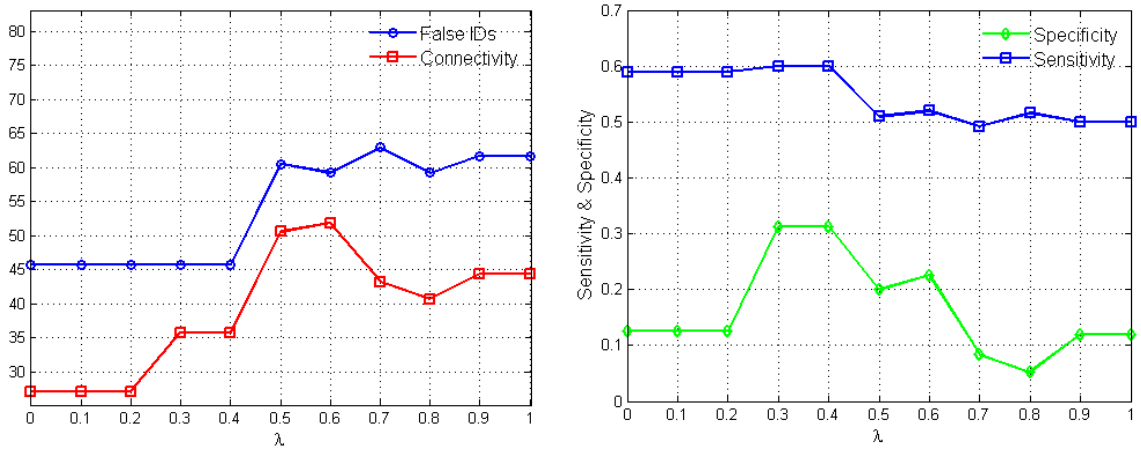

(a) (b)

**Figure 2.** Known and recovered GRN for SOS pathway in *E. coli*. (**a**) Diagram of interactions in the SOS network. DNA lesions caused by mitomycin C (MMC) (**blue hexagon**) are converted to single-stranded DNA during chromosomal replication. Upon binding to ssDNA, the RecA protein is activated (RecA*) and serves as a coprotease for the LexA protein. The LexA protein is cleaved, thereby diminishing the repression of genes that mediate multiple protective responses. Green arrows denote positive regulation, while red arrows denote negative regulation. *Adapted from* [8]. (**b**) Diagram of the recovered gene regulatory network of the SOS pathway in *Escherichia coli*. Green arrows denote positive regulation, while red arrows denote negative regulation.

Figure 2b shows the inferred gene regulatory network from the steady-state data. The identification algorithm accurately identified the key regulatory associations in the network. For instance, the model correctly shows that *lexA* activates *recA* while negatively regulating its own transcription, whereas *recA* negatively regulates its own transcription. In addition, the model identified *lexA* as having the greatest regulatory influence on the other genes in the network. Due to the differences in network topology (e.g., *recA*, *lexA* and *CDC20*), inaccuracies are expected from either the current published model, the LASSO-VAR GRN recovery, or both. Some of these potential differences may alternatively be dependent on the dynamic state of the system as inferred from the temporal context.

The plots in Figure 3 show the variations in algorithm performance as the penalty parameter $\lambda$ varies between 0 and 1. In Figure 3a, the total number of false identifications (*i.e.*, false activations, inhibitions and no-interactions) and the net connectivity of the network are measured against the different $\lambda$ values. The net connectivity provides a measure of the total number of interactions inferred by the algorithm. As such since sparsity is required, lower values in the net connectivity is desired. Figure 3b shows the variations in the performance metrics, sensitivity and specificity, as $\lambda$ changes.

The choice of the "best" penalty parameter for the given application represents the $\lambda$ value that produces the best trade-off between the number of false identification, net connectivity, sensitivity and specificity. In this regard, values of 0.3 and 0.4 produce 46% false identification, 37% net connectivity, sensitivity of 60% and specificity of 31%. Eventually, $\lambda = 0.3$ is selected based on its superior MSFE value of 5.20 as opposed to 5.22 for $\lambda = 0.4$ and satisfies the desired constraints of stability, sparsity and causality.




(a) (b)

**Figure 3.** Variations in $\lambda$ and the corresponding algorithm performance. (**a**) plot of $\lambda$ *versus* total number of false identifications and net connectivity in percentages. (**b**) plot of $\lambda$ *versus* sensitivity and specificity.

Table 2 shows how the proposed network identification algorithm without *a priori* knowledge of the network structure compares with that proposed by Zavlanos *et al.* [26] with 30% *a priori* knowledge of the network.

**Table 2.** Comparisons of the inferred network for the SOS pathway in *E. coli* using LASSO-VAR and Zavlanos' method.

| | TP | FP | TN | FN | Sensitivity | Specificity | Precision |
|---|---|---|---|---|---|---|---|
| LASSO-VAR | 39 | 11 | 5 | 26 | 60% | 31% | 78% |
| ZAVLANOS [26] | 40 | 10 | 15 | 16 | 71% | 60% | 80% |

Table 2 shows how the proposed algorithm compares to that proposed by Zavlanos *et al.* In all, the recovered network has four false activations, eight false inhibitions, 25 false no-interactions, and 37 false identifications in total, while satisfying the desired constraints of stability, sparsity and causality. The penalty parameter, $\lambda$ selected is 0.4 based to its MSFE value.

### 3.2. *Yeast* Saccharomyces Cerevisiae *Cell Cycle*

From 1998, when Spellman *et al.* published the yeast *Saccharomyces cerevisiae* (*i.e.*, budding yeast) cell cycle microarray expression levels [38], many computational methods have been applied to these data. To demonstrate the applicability of the proposed algorithm in this study, a subset from yeast *Saccharomyces cerevisiae* microarray time series dataset including 14 genes, FUS3, SIC1, FAR1, CDC6, CDC20, CDC28, CLN1, CLN2, CLN3, CLB5, CLB6, SWI4, SWI6 and MBP1, is perturbed and used. The details of *S. cerevisiae* cell cycle control are well known, as shown in Figure 4a.

The 14 genes are known to be involved in the early cell cycle of the yeast *Saccharomyces cerevisiae*. The cell cycle describes the series of events that precedes its division and duplication [40]. The mitotic cell cycle in yeast is accomplished through a reproducible sequence of events: DNA replication (S phase) and mitosis (M phase) separated temporally by gaps, G1 and G2 phases. At the G1 phase, CDC28 associates with CLN1, CLN2 and CLN3, while CLB5 and CLB6 controls CDC during S, G2, and M phases [41]. Cell cycle progression begins upon the activity of CLN3/CDC28. When the levels of CLN3/CDC28 accumulate more than a certain threshold, SWI4/SWI6 and MBF1/SWI6 are activated, promoting transcription of CLN1 and CLN2 [41]. CLN1/CDC28 and CLN2/CDC28 promote

activation of other associated kinase, which drives DNA replication. SIC1 and FAR1 are the substrates and inhibitors of CDC28. CDC6 and CDC20 affect the cell division control proteins. Mitogen-activated protein kinase affect this progression through FUS3 [41]. The dataset generated by Spellman *et al.* [38] contains three time series measured using different cell synchronization methods: $\alpha$ factor-based arrest (referred to as alpha, includes 18 time points at 7 min interval over 119 min), size-based (*elu*, 14 time points at 30 min interval over 390 min), and arrest of a cdc15 temperature-sensitive mutant (*cdc15*, 24 time points, the first four and last three of which are at 20 min interval and the rest are at 10 min interval over 290 min). The *alpha* dataset is used and then studied in more detail as has been explored in literature [42].

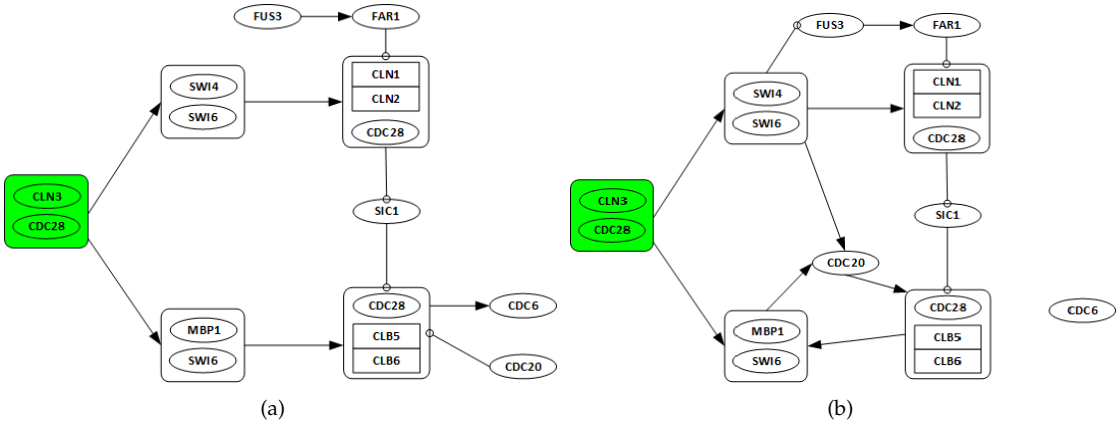



**Figure 4.** Known and recovered GRN for cell cycle pathway in yeast *Saccharomyces cerevisiae*. (**a**) target pathways of the 14 genes. CDC28 associates with cyclin CLN3 at the start of mitosis to cause the activation of SBF (SWI4/SWI6) and MBF (MBP1/SWI6), promoting the transcription of CLN1, CLN2. At G1 phase CDC28 associates with G1-cyclins CLN1 to CLN3, while B-type cyclins CLB1 to CLB6 regulate CDC28 during S, G2, and M phases. CLN1 and CLN2 interacting with CDC28 promote activation of B-type cyclin associated Cyclin-dependent kinase (CDK), which drives DNA replication and entry into mitosis. *Adapted from* [41]. (**b**) recovered yeast cell cycle pathway. The arrows show the direction of regulation. Some key regulations like activation (positive regulation) of the SBF (SWI4/SWI6) and MBF (MBP1/SWI6) complexes by the starter complex (CDC28/CLN3) are recovered.

Data for the expected topology of this network were extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [41], which is a major collection of knowledge for molecular and genetic pathways and includes information on experimental observations in organisms. The KEGG regulatory pathway represents current knowledge on the protein and gene interaction networks.

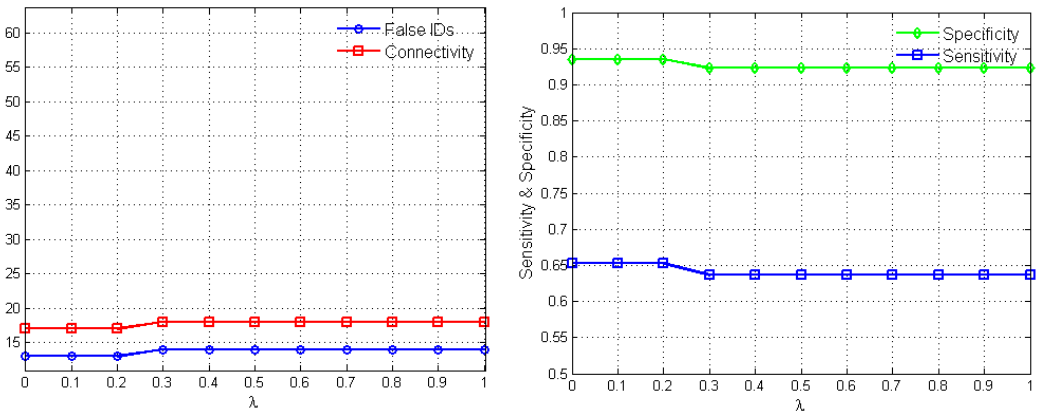
#### 3.2.1. Network Recovery

As discussed in the preceding section, the KEGG pathway is considered as the target network for comparison. Complexes including one or several genes are considered as a '*gene*' in the network. There are 10 complexes, including CLN3/CDC28, SWI4/SWI6, MBP1/SWI6, CLN1/CLN2/CDC18, and CLB5/CLB6/CDC28. Other nodes that are made of one single gene only, CDC20, CDC6, SIC1, FAR1, and FUS. The following assumptions are made:

1. Genes CLN3 and CDC28 are only considered as possible regulators, as they are starters of the cell cycle network.
2. All discovered links from any gene in one complex to any other genes in a different complex are considered as a single regulation.
3. All regulations among genes in the same complex are ignored.

Figure 4b is the recovered network from the perturbed *alpha* gene expression data for the yeast *Saccharomyces cerevisiae* cell cycle pathway.

The recovered network has seven true positives and five false positives. The algorithm recovers key regulations. For instance, the activation (positive regulation) of the SWI4/SWI6 and MBP1/SWI6 complexes by the starter complex are identified. The graphs in Figure 5 show the changes in algorithm performance for varying $\lambda$ values. In Figure 5a, the relationship between $\lambda$ and total false identification as well as the net connectivity of the network. The plot in Figure 5b show the variations in $\lambda$ and the corresponding effects on the performance of the algorithm. Again, the terms false identifications and net connectivity have the same meanings as discussed in Section 3.1.1. The results in this application are quite uniform due to the assumptions made in grouping some genes into complexes as required. $\lambda \in [0, 0.2]$ provides the best trade-off between the number of false identification, net connectivity, sensitivity and specificity. They produce sensitivity and specificity values of 65% and 94%, respectively. Based on its lower MSFE value, $\lambda = 0.2$ is chosen.




(a) (b)

**Figure 5.** Variations in $\lambda$ and the corresponding algorithm performance. (**a**) plot of $\lambda$ *versus* total number of false identifications and net connectivity in percentages; (**b**) plot of $\lambda$ *versus* sensitivity and specificity.

## 4. Conclusion and Discussion

In this paper, the least absolute shrinkage and selection operator—vector autoregressive (LASSO-VAR) model–has been adapted in solving the problem of identifying a minimal model that best explains genetic perturbation data obtained at a network's equilibrium state. The fact that the network identification algorithm is an autoregressive technique means it has the inherent ability to model Granger causality, making it possible to identify regulatory roles among the genes under consideration. Additionally, the technique handles the sparsity constraint imposed by the loose-connectivity restriction of biological networks through the application of the $L_1$ penalty term. Due to the steady-state nature of the expression data, a stability constraint is imposed on the original LASSO-VAR objective function, which allows robustness of the inferred networks to slight variations in the input.

To evaluate the reliability and efficiency of LASSO-VAR for recovering stable, sparse and causal regulatory interactions from steady-state gene expression data, data from the SOS pathway in *E. coli* was first used. The performance of the algorithm was measured and compared with results obtained in literature. This comparison was based on two statistical evaluation criteria, sensitivity and specificity, which allowed the accuracies of the inferred network structure using the LASSO-VAR technique to be quantified. LASSO-VAR performed without prior knowledge at a roughly comparable level to the alternative Zavanos method that requires some prior knowledge. The efficiency of the stable LASSO-VAR for learning the network structure was then evaluated using the perturbed gene

expression data of 14 genes in yeast *Saccharomyces cerevisiae* cell cycle reported in [41]. The network inferred from the yeast data by LASSO-VAR is compared with the known network from the cell cycle pathway of the yeast *Saccharomyces cerevisiae* using the evaluation criteria: sensitivity and specificity. Results showed the ability of the identification algorithm to infer the regulatory network for the cell cycle pathway.

The surge of large biological data sets [43,44] and the aggressive efforts at methods for ontology-driven annotation and data modeling [45,46] have helped to provide opportunities for a more objective and reproducible basis for analysis. The formulation of our pathway recovery analysis model fulfills those criteria necessary for it to be highly scalable with data for biological networks by: (1) avoiding the context-limiting aspects of *a priori* knowledge and presumptions of frequentist-type statistics which are difficult to implement based on the inherent sparsity of biological networks; and (2) being strictly data-driven in matrix-based numerical forms with one controlling parameter—*i.e.*, being outside of subjective standards of knowledge and curation. The criterion for stability as we investigate it as a necessary condition for steady-state data, although seemingly trivial, is nonetheless fundamental for how structural and functional robust aspects for pathways are identified. In general, scoring outcomes from our methodology identified key mechanisms of pathways through a scoring gradient. We expect there to be significant dividends from this approach that go beyond our initial usage of a reference database for tallying of true *versus* false results. The quantitative support underlying false positive and false negative results for the target model may aid in the development and testing of new hypotheses, or quality control measures, agendas that are important in large-scale investments into empirical data collection such as for perturbation studies.

Canonical pathways provide well-vetted models that have a deep legacy in empirical published studies [46,47]. It is still the case, however, that uncharted dynamics and natural variation go beyond the limited range of studied organisms and would furthermore impact predictive modeling even for the two chosen models of this study for which knowledge remains limited [48,49]. For instance, synchronization of expression across multiple genes is likely lost following its initial measurement at a starting point of an experimental assay. The gain or loss of known interactions as input data would furthermore be expected to impact the predictive power of LASSO-VAR. There is, therefore, a need to more systematically study LASSO-VAR across these frontier contexts involving a larger number of simulated and empirical data sets. This would in particular aid the empirical study of computational performance, beyond theoretical expectations for restricting the parameter space through the penalty parameter [27]. This would also provide a robust capacity for selecting which genes to base a model upon and for constructing the analysis in a way that separates initialization and training from forecast evaluation, to guide contrasting modes of usage for how LASSO-VAR could be used in supervised *versus* unsupervised modes of analytical evaluations. As multiple gene expression data resources and annotations may be harnessed for this effort [50–53], it remains an essential next step to identify computational and theoretical limits for objectively inferring GRN configurations based on input data complexity and sizes. The overall outcome for such an effort would help resolve the lack of overlap between pathway databases and approaches to analytical treatment, both with respect to content [50,54] and foundational criteria such as how to define start and stop points for individual pathways [50]. Future usage of this approach could identify pathways for assembly and disassembly of differentially constructed multicomponent cellular objects recovered at the same point of time. Such usage of cellular component data to infer subunit associations would add to the explanatory potential of this algorithm, and help to guard against the *post hoc, ergo propter hoc* fallacy for how changes in cellular composition would otherwise be inferred from analyses conducted only upon time series.

**Acknowledgments:** The authors wish to acknowledge Michael Zavlanos for making the source codes to his genetic network identification algorithm available. His source package has been modified and used in this work. This work is partially supported by the National Science Foundation (NSF) under Cooperative Agreement No. CCF-1029731. In addition, the third and the fifth authors would like to acknowledge the support from the Air Force Research Laboratory and Office of the Secretary of Defense (OSD) for sponsoring this research under agreement number FA8750-15-2-0116. The views and conclusions contained herein are those of the authors and

should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, OSD, NSF or the U.S. Government.

**Author Contributions:** Joy Edward Larvie has made substantial contributions to the conception and design of the study, analysis and interpretation of biological data. Mohammad Gorji Sefidmazgi has made substantial contributions to the analysis and interpretation of mathematical results. Joy Edward Larvie and Mohammad Gorji Sefidmazgi have been involved in drafting the manuscript. Abdollah Homaifar and Ali Karimoddini have critically read and revised the manuscript for important mathematical content. Scott Harrison and Anthony Guiseppi-Elie have provided critical biological interpretation to the results and have critically read and revised the manuscript for analytical utility. All authors read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2001**, *2*, 418–427.
2. Michailidis, G.; d'Alché Buc, F. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Math. Biosci.* **2013**, *246*, 326–334.
3. Hasty, J.; McMillen, D.; Isaacs, F.; Collins, J.J. Computational studies of gene regulatory networks: In numero molecular biology. *Nat. Rev. Genet.* **2001**, *2*, 268–279.
4. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R.D.; Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* **2006**, *7*, S7.
5. Stark, C.; Breitkreutz, B.J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34*, D535–D539.
6. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210.
7. Tegnér, J.; Yeung, M.K.S.; Hasty, J.; Collins, J.J. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *PNAS* **2003**, *100*, 5944–5949.
8. Gardner, T.S.; Di Bernardo, D.; Lorenz, D.; Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **2003**, *301*, 102–105.
9. Chai, L.E.; Loh, S.K.; Low, S.T.; Mohamad, M.S.; Deris, S.; Zakaria, Z. A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **2014**, *48*, 55–65.
10. Lopes, F.M.; de Oliveira, E.A.; Cesar, R.M. Inference of gene regulatory networks from time series by Tsallis entropy. *BMC Syst. Biol.* **2011**, *5*, 61.
11. Wang, Y.X.R.; Huang, H. Review on statistical methods for gene network reconstruction using expression data. *J. Theor. Biol.* **2014**, *362*, 53–61.
12. Polynikis, A.; Hogan, S.J.; di Bernardo, M. Comparing different ODE modelling approaches for gene regulatory networks. *J. Theor. Biol.* **2009**, *261*, 511–530.
13. De Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **2002**, *9*, 67–103.
14. Hecker, M.; Lambeck, S.; Toepfer, S.; van Someren, E.; Guthke, R. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **2009**, *96*, 86–103.
15. Ahmad, J.; Bourdon, J.; Eveillard, D.; Fromentin, J.; Roux, O.; Sinoquet, C. Temporal constraints of a gene regulatory network: Refining a qualitative simulation. *Biosystems* **2009**, *98*, 149–159.
16. Someren, E.V.; Wessels, L.; Backer, E.; Reinders, M. Genetic network modeling. *Pharmacogenomics* **2002**, *3*, 507–525.
17. Hartemink, A.J. Reverse engineering gene regulatory networks. *Nat. Biotechnol.* **2005**, *23*, 554–555.
18. Yeung, M.S.; Tegnér, J.; Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6163–6168.
19. Wang, Y.; Joshi, T.; Zhang, X.S.; Xu, D.; Chen, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* **2006**, *22*, 2413–2420.
20. Karlebach, G.; Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 770–780.

21. Kordmahalleh, M.M.; Sefidmazgi, M.G.; Homaifar, A.; Karimoddini, A.; Guiseppi-Elie, A.; Graves, J.L. Delayed and Hidden Variables Interactions in Gene Regulatory Networks. In Proceedings of the 2014 IEEE International Conference on Bioinformatics and Bioengineering (BIBE), Boca Raton, FL, USA, 10–12 November 2014; pp. 23–29.
22. D'haeseleer, P.; Liang, S.; Somogyi, R. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **2000**, *16*, 707–726.
23. Kauffman, S. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **1969**, *22*, 437–467.
24. Bornholdt, S. Less is more in modeling large genetic networks. *Science* **2005**, *310*, 449.
25. Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **2000**, *7*, 601–620.
26. Zavlanos, M.M.; Julius, A.A.; Boyd, S.P.; Pappas, G.J. Inferring stable genetic networks from steady-state data. *Automatica* **2011**, *47*, 1113–1122.
27. Nicholson, W.; Matteson, D.; Bien, J. *Structured Regularization for Large Vector Autoregression*; Technical report; Cornell University: Ithaca, NY, USA, 2014.
28. Larvie, J.E.; Gorji, M.S.; Homaifar, A. Inferring stable gene regulatory networks from steady-state data. In Proceedings of the 2015 41st Annual Northeast Biomedical Engineering Conference (NEBEC), Troy, NY, USA, 17–19 April 2015; pp. 1–2.
29. Fujita, A.; Sato, J.R.; Garay-Malpartida, H.M.; Yamaguchi, R.; Miyano, S.; Sogayar, M.C.; Ferreira, C.E. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Syst. Biol.* **2007**, *1*, 39.
30. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin, Germany, 2007.
31. Mukhopadhyay, N.D.; Chatterjee, S. Causality and pathway search in microarray time series experiment. *Bioinformatics* **2007**, *23*, 442–449.
32. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438.
33. Hsu, N.J.; Hung, H.L.; Chang, Y.M. Subset selection for vector autoregressive processes using Lasso. *Comput. Stat. Data Anal.* **2008**, *52*, 3645 – 3657.
34. Chen, J. Sufficient conditions on stability of interval matrices: connections and new results. *IEEE Trans. Autom. Control* **1992**, *37*, 541–544.
35. Rajapakse, J.C.; Mundra, P.A. Stability of building gene regulatory networks with sparse autoregressive models. *BMC Bioinform.* **2011**, *12*, S17.
36. CVX Research Inc. CVX: Matlab Software for Disciplined Convex Programming, Version 2.0, 2012. Available online: http://cvxr.com/cvx (accessed on 15 October 2015).
37. De Muth, J. *Basic Statistics and Pharmaceutical Statistical Applications*, 2nd ed.; Pharmacy Education Series, Taylor & Francis: Boca Raton, FL, USA, 2006.
38. Spellman, P.T.; Sherlock, G.; Zhang, M.Q.; Iyer, V.R.; Anders, K.; Eisen, M.B.; Brown, P.O.; Botstein, D.; Futcher, B. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell* **1998**, *9*, 3273–3297.
39. The Yeast Cell Cycle Analysis Database. Available online: http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt (accessed on 15 November 2015).
40. Lodish, H. *Molecular Cell Biology*, 5th ed.; W. H. Freeman: New York, NY, USA, 2003.
41. KEGG PATHWAY: map04111. Available online: http://www.genome.jp/dbget-bin/www_bget?map04111 (accessed on 27 April 2015).
42. Kim, S.Y.; Imoto, S.; Miyano, S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* **2003**, *4*, 228–235.
43. Schatz, M.; Langmead, B. The DNA data deluge. *IEEE Spectr.* **2013**, *50*, 28–33.
44. Barrett, T. Gene Expression Omnibus (GEO). Available online: http://www.ncbi.nlm.nih.gov/geo/ (accessed on 15 October 2015).
45. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **2015**, *43*, D1049–D1056.

46. Fehrmann, R.S.N.; Karjalainen, J.M.; Krajewska, M.; Westra, H.J.; Maloney, D.; Simeonov, A.; Pers, T.H.; Hirschhorn, J.N.; Jansen, R.C.; Schultes, E.A.; *et al*. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **2015**, *47*, 115–125.
47. Zhou, H.; Jin, J.; Zhang, H.; Yi, B.; Wozniak, M.; Wong, L. IntPath–an integrated pathway gene relationship database for model organisms and important pathogens. *BMC Syst. Biol.* **2012**, *6* (Suppl. 2), S2.
48. Michel, B. After 30 Years of Study, the Bacterial SOS Response Still Surprises Us. *PLoS Biol.* **2005**, *3*, e255.
49. Alberghina, L.; Mavelli, G.; Drovandi, G.; Palumbo, P.; Pessina, S.; Tripodi, F.; Coccetti, P.; Vanoni, M. Cell growth and cell cycle in Saccharomyces cerevisiae: Basic regulatory design and protein-protein interaction network. *Biotechnol. Adv.* **2012**, *30*, 52–72.
50. Caspi, R.; Dreher, K.; Karp, P.D. The challenge of constructing, classifying, and representing metabolic pathways. *FEMS Microbiol. Lett.* **2013**, *345*, 85–93.
51. Krämer, A.; Green, J.; Pollard, J.; Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014**, *30*, 523–530.
52. Croft, D.; Mundo, A.F.; Haw, R.; Milacic, M.; Weiser, J.; Wu, G.; Caudy, M.; Garapati, P.; Gillespie, M.; Kamdar, M.R.; *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **2014**, *42*, D472–D477.
53. Morgat, A.; Coissac, E.; Coudert, E.; Axelsen, K.B.; Keller, G.; Bairoch, A.; Bridge, A.; Bougueleret, L.; Xenarios, I.; Viari, A. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **2012**, *40*, D761–D769.
54. Karp, P.D. Pathway databases: a case study in computational symbolic theories. *Science* **2001**, *293*, 2040–2044.