BMC Systems Biology

Methodology article

Modeling gene expression regulatory networks with the sparse vector autoregressive model

André Fujita[1,2], João R Sato[1], Humberto M Garay-Malpartida[2,3], Rui Yamaguchi[4], Satoru Miyano[4], Mari C Sogayar[2] and Carlos E Ferreira*[1]

Address: [1]Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010 – São Paulo, 05508-090, SP, Brazil, [2]Chemistry Institute, University of São Paulo, Av. Lineu Prestes, 748 – São Paulo, 05513-970, SP, Brazil, [3]School of Arts, Science and Humanities, University of São Paulo, Av. Arlindo Bettio, 1000 – São Paulo, 03828-000, SP, Brazil and [4]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

Email: André Fujita - fujita@ime.usp.br; João R Sato - jsato@ime.usp.br; Humberto M Garay-Malpartida - hmgaray@usp.br; Rui Yamaguchi - ruiy@ims.u-tokyo.ac.jp; Satoru Miyano - miyano@ims.u-tokyo.ac.jp; Mari C Sogayar - mcsoga@iq.usp.br; Carlos E Ferreira* - cef@ime.usp.br

* Corresponding author

Published: 30 August 2007

BMC Systems Biology 2007, **1**:39 doi:10.1186/1752-0509-1-39

Received: 25 April 2007
Accepted: 30 August 2007

This article is available from: http://www.biomedcentral.com/1752-0509/1/39

Abstract

Background: To understand the molecular mechanisms underlying important biological processes, a detailed description of the gene products networks involved is required. In order to define and understand such molecular networks, some statistical methods are proposed in the literature to estimate gene regulatory networks from time-series microarray data. However, several problems still need to be overcome. Firstly, information flow need to be inferred, in addition to the correlation between genes. Secondly, we usually try to identify large networks from a large number of genes (parameters) originating from a smaller number of microarray experiments (samples). Due to this situation, which is rather frequent in Bioinformatics, it is difficult to perform statistical tests using methods that model large gene-gene networks. In addition, most of the models are based on dimension reduction using clustering techniques, therefore, the resulting network is not a gene-gene network but a module-module network. Here, we present the Sparse Vector Autoregressive model as a solution to these problems.

Results: We have applied the Sparse Vector Autoregressive model to estimate gene regulatory networks based on gene expression profiles obtained from time-series microarray experiments. Through extensive simulations, by applying the SVAR method to artificial regulatory networks, we show that SVAR can infer true positive edges even under conditions in which the number of samples is smaller than the number of genes. Moreover, it is possible to control for false positives, a significant advantage when compared to other methods described in the literature, which are based on ranks or score functions. By applying SVAR to actual HeLa cell cycle gene expression data, we were able to identify well known transcription factor targets.

Conclusion: The proposed SVAR method is able to model gene regulatory networks in frequent situations in which the number of samples is lower than the number of genes, making it possible to naturally infer partial Granger causalities without any *a priori* information. In addition, we present a statistical test to control the false discovery rate, which was not previously possible using other gene regulatory network models.

Background

In order to understand cell functioning as a whole, it is necessary to describe, at the molecular level, how gene products interact with each other. This could help to identify new target genes and to design new drugs for treatment of several diseases [1-3]. Due to the high number of genes involved in these networks, activating or suppressing feedback loops, the dynamics of their interactions is very complex and difficult to infer.

With the development of high-throughput technologies, such as DNA microarrays, it is possible to simultaneously analyze the expression of up to thousands of genes and to construct gene networks based on inferences over gene expression data.

Several methods to model genetic networks were proposed in the last few years, such as the Bayesian networks [4-8], Structural Equation Models [9], Probabilistic Boolean Networks [10-12], Graphical Gaussian Models [13], Fuzzy controls [14], and Differential Equations [15].

Although these methods allow modeling several regulatory networks for which biological information is available, it is difficult to determine the flow of information when there is no *a priori* knowledge.

In addition, all of these methods face the same problem, i.e., the number of samples (microarrays) is very small, when compared to the high number of variables (genes) (ill posed problems, related to the "curse of dimensionality") [16]. Therefore, it is difficult to infer large scale networks using traditional statistical methods, limiting this inference to only a few genes. As a consequence, modeling and simulating large networks becomes a field of intensive and challenging research. At this point, it is important to define what is considered a "large" network. We consider as "large" a network in which the number of genes is larger than the number of microarrays experiments, implying in a large number of parameters to be estimated.

Some methods have been developed to overcome this problem. For example, Barrera *et al*. use mutual information for dimension reduction [17], with mutual information between genes being computed and then, the highest mutual informations selected. However, this approach is not founded on a statistical test, rendering it very difficult to interpret and identify the actual edges of the network. Therefore, the choice of the threshold parameter to determine whether there is or not a connection, becomes quite subjective. An alternative to model the large number of genes is to construct modules (clusters), where each module is composed by several genes, and then, to construct the module-module networks [18-20]. A limitation of these methods is that they still are not a gene-gene network, therefore, interpretation of the meaning of each module is difficult, varying with each cluster.

Here we present the Sparse Vector Autoregressive model to approach these problems. This method was first applied, with success, in neurosciences, to estimate functional connectivity between several brain areas [21]. Here, we present the Sparse Vector Autoregressive model based on LASSO penalized regression for variable selection to reduce the dimensionality on large gene networks.

In cases of multiple time series, a first approach to infer connectivity would be to apply techniques such as multivariate autoregressive modeling (VAR), which allows identification of connectivity by combining graphical modeling methods with the concept of Granger causality [22]. This is an attractive approach since it does not require *a priori* network information. Unfortunately, the current time series methods can only be applied only for cases in which the length of the time-series T is much larger than n, the number of genes, which is exactly the reverse of the situation commonly found in microarray experiments, for which relatively short time-series are measured over tens of thousands of genes. The Sparse Vector AutoRegressive model (SVAR), on the other hand, estimates the network in a two-stage process involving (i) penalized regression with LASSO regression [23] and (ii) pruning of unlikely connections by means of the False Discovery Rate (FDR) developed by [24]. Extensive simulations were performed with artificial gene networks having scale-free like topologies [25] and stable dynamics. These simulations show that the detection efficiency of connections of the proposed procedure is quite high. An application of the method to actual HeLa cell line data was illustrated by the identification of well known transcription factor targets and circuitries involving important genes in cancer development.

Results and discussion

In order to measure the performance of SVAR, intensive simulations were carried out. For this purpose, we simulated hundreds of networks with scale-free like topology since the metabolic network was described as scale-free graphs by [25]. In our case, the graph nodes represent the genes whereas the edges represent the Granger-causal relationships. For details of these artificial regulatory networks, see the Methods section.

The number of genes was kept at n = 100 and we varied the sample size, i.e., the time-series length (time-series length T = 25, 50, 75, 100, 125, 150, 175 and 200 for SVAR and T = 110, 125, 150, 175 and 200 for VAR). Notice that, for VAR of order one, $m = T - 1$ must be larger than n. For each time-series length, we performed 100 simulations, i.e., 100 different scale-free like graphs were

generated. The starting conditions of the scale-free like graphs were two fully connected genes ($z_0 = 2$, $z_{edges} = 2$, where z_0 is the initial number of genes and z_{edges} is the initial number of edges), in other words, two nodes with two edges, one pointing to the other. The number of edges added at each iteration is $z = 1$, therefore, each network is composed by 100 genes and 100 edges out of 10,000 possible edges (the maximum number of possible edges is n^2). Notice that since the goal is to construct a network with $n = 100$ genes, we set the number of iterations $T_{step} = n - z_0 = 98$. In Figure 1, an example of the artificially generated gene expression regulatory network is illustrated.

It is important to highlight that SVAR was able to identify true positive edges even when the time-series length was lower than the number of genes. Figures 2, 3 and 4 show, respectively, the number of true positives inferred by SVAR and VAR for controlled false positives rate, i.e., q-value (error type I rate within rejected hypotheses) thresholds lower than 0.01, 0.05 and 0.10. Since the estimated β's standard error is proportional to the time series' length (the greater the time series, the lower is the β's standard error) we varied only the time series' length.

Analyzing figures 2, 3 and 4, we obtained the following results:

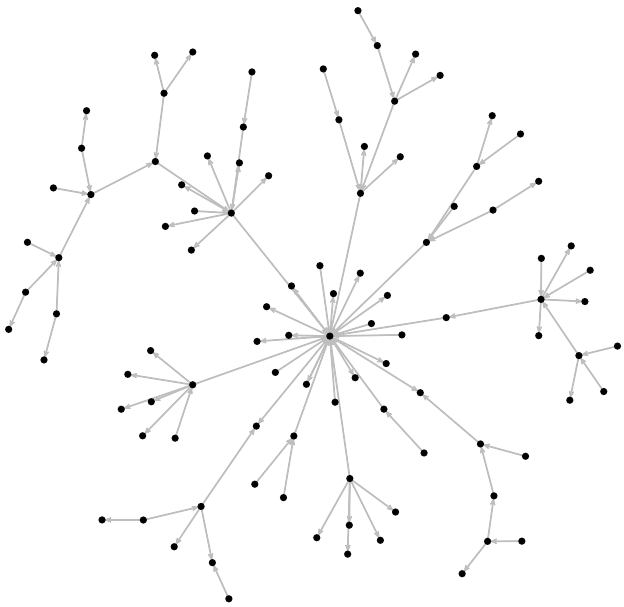

Figure 1
Artificial gene regulatory network. Example of a simulated sparse gene regulatory network with $n = 100$ genes and 100 connections. The arrows indicate the Granger-causal relationships.

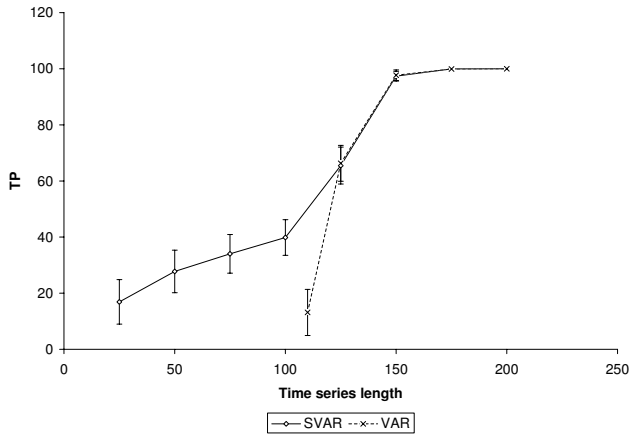

Figure 2
Comparison between SVAR and VAR. The simulations were performed in a scale-free like network composed of 100 nodes and 100 edges. VAR was performed only for experiments with the length of the time-series of up to 110. TP: True positives. The number of false positives is controlled using q-value < 0.01. The error bar is representing one standard error.

1. The capacity of SVAR to identify true positives even when the number of samples is lower than the number of genes is satisfactory. This was found when comparing the performance between SVAR, with the time-series length equal to 50, and VAR, with time-series length equal to 110. Also, in this case, SVAR has identified more true pos-

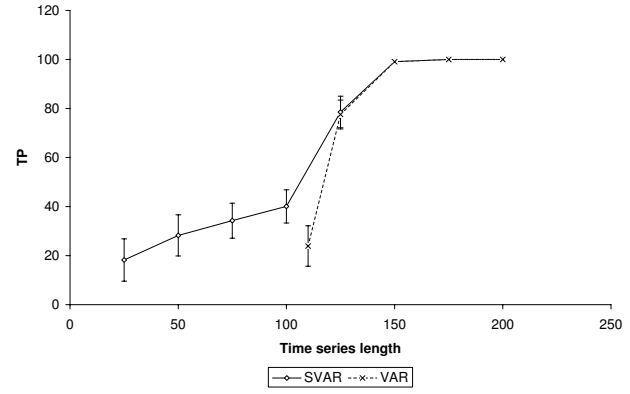

Figure 3
Comparison between SVAR and VAR. The simulations were performed in a scale-free like network composed of 100 nodes and 100 edges. VAR was performed only for experiments with the length of the time-series of up to 110. TP: True positives. The number of false positives is controlled using q-value < 0.05. The error bar is representing one standard error.

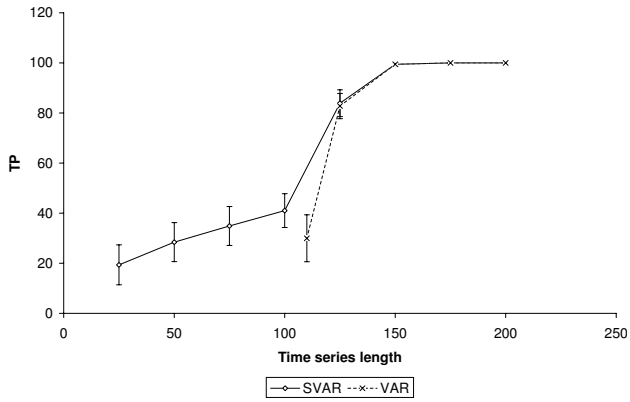

Figure 4
Comparison between SVAR and VAR. The simulations were performed in a scale-free like network composed of 100 nodes and 100 edges. VAR was performed only for experiments with the length of the time-series of up to 110. TP: True positives. The number of false positives is controlled using q-value < 0.10. The error bar is representing one standard error.

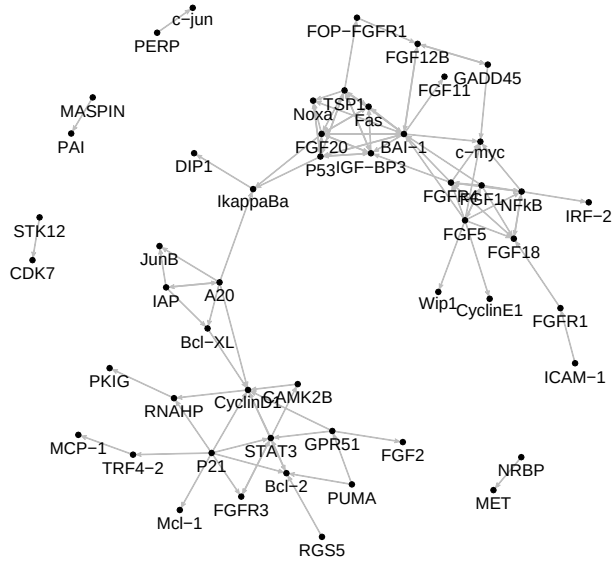

Figure 5
HeLa gene expression regulatory network. Gene regulatory network inferred from HeLa cell cycle gene expression data. The arrows represent the Granger-causal associations with q-value < 0.05. Genes with no Granger-causal links identified by SVAR were not plotted.

itive edges than VAR (the proportion of the quantity of true positives inferred by SVAR is about 75% higher than the number of true positives inferred by VAR).

2. By comparing SVAR and VAR when the number of genes is lower than the number of samples, in general, SVAR is slightly more powerful than VAR, since the number of connectivities is larger than the number of samples.

3. When $m \gg n$, where $m = T - 1$ and n is the number of genes, there is no statistical difference between SVAR and VAR. This could be explained, in this context, because the best λ which minimizes the GCV (Generalized Cross-Validation) is near to zero. When $\lambda = 0$, the SVAR model becomes the traditional VAR model.

We have also analyzed the expression profile of a set of 94 cell cycle-regulating genes represented by 48 microarrays, i.e., the number of genes n is approximately 2 times larger than the time-series length T. Figure 5 shows the genes that display any connectivity under a false-positive rate (FDR) of 5% (q-value < 0.05). Genes with no connectivity were excluded.

The SVAR method reveals at least three gene regulatory networks related to cell transformation and tumor progression, namely: NFκB, p53, and STAT3 transcriptional modules [26-28], which is in agreement with already well known cell cycle-regulated pathways in several cellular models and in Hela cells themselves.

It is important to highlight that the out-degree (number of edges with the gene as their initial vertex) of genes encoding proteins that act as well-known transcriptional factors (p53, NFκB and STAT3) or important genes for cell proliferation control (*p21*, *bai1*, *tsp1*, *a20*) is higher than that of other genes. In a similar analysis, the in-degree (number of edges with the gene as their terminal vertex) of the FGFs (*fgf18*, *fgf20*, *fgfr4*) and of genes involved in cell cycle regulation and apoptosis (*cyclin d1*, *c-myc*, *bcl-2*, *noxa*, *fas*) is also higher, demonstrating the association between their key role in cell homeostasis and their in-degree and/or out-degree values [29].

NFκB is an inducible transcription factor complex formed by heterodimeric association between *relA* and *c-rel* gene products, whose transcriptional activity is regulated by interaction with the inhibitory IκBα protein. It has already been demonstrated that activation of NFκB controls cell-cycle progression in HeLa cells by several mechanisms [30]. The SVAR method was not able to identify the relationship between NFκB and its natural targets, such as *A20*, *iap*, *bclx* and *iκBα* genes. However, SVAR is showing that NFκB directly regulates several fibroblast growth factors (FGFs) and the c-Myc protein, which are key regulators of cell proliferation. Indeed, it is noticed that the majority of NFκB transcriptional activity is mediated by interaction with FGFs-related proteins, at the upstream and/or downstream levels. These results support the

hypothesis that some of the multiple aspects of tumorigenesis in Hela cells may be related to NFκB -mediated transcription of FGFs-related proteins.

As discussed above, the positive NFκB regulation of several well-known natural targets was not detected by SVAR. However, these regulatory processes appear to be present, even in the absence of an evident direct link with NFκB, since all of these transcriptionally regulated genes form a highly related network (Figure 5). A20, a zinc finger protein, which is transcriptionally regulated by NFκB in several cell types [31], appears to orchestrate the genes relationship in this network, activating the transcription of well-known anti-apoptotic genes, such as *iap*, *bclx* and *junB* – NFκB target genes themselves [32-34] – towards transduction of the proliferative transcriptional activity of NFκB. The A20 protein is also involved in NFκB regulation, blocking its activity, in a negative feedback mechanism [35]. Although this control is operated at the post-transcriptional level, results obtained using the SVAR method suggest that this process could also be controlled by A20-mediated positive regulation of *iκBα* (Figure 5). These results confirm the reliability of SVAR for predicting gene relationship, since *iκBα*, the natural NFκB inhibitor has a key role in controlling the NFκB -regulated cell cycle events in Hela cells, as referred to in literature [30]. Moreover, SVAR showed that this role of *iκBα* in Hela cell cycle progression also appears to be regulated through p53-mediated activation of *iκBα* (Figure 5), in agreement with data reported in the literature [36]. In summary, these data support the hypothesis that *iκBα* may be involved in attenuation of tumor progression and be responsible for the mildly invasive phenotype displayed by Hela cells.

The p53 protein is a transcription factor that binds to the enhancer/promoter elements of downstream target genes and regulating their transcription and initiating cellular programs that account for most of its tumor-suppressor functions, namely: cell cycle arrest, inhibition of angiogenesis and metastasis, apoptosis induction and DNA repair [37]. The SVAR method was capable of identifying the interactions of several members of the p53 network. IGF-BP3 (IGF-binding protein 3), an inhibitor of insulin-like growth factor, and NOXA, a BCL-2 homology domain 3-only (BH3-only) protein, are transcriptionally activated by p53 in activation of apoptosis in several cell types [38,39]. Our *in silico* results showed that this regulation is also present in Hela cells. Although the *fas* gene is not a known target of p53, its activation could be mediated by other p53 targets, leading to increased apoptosis rate and cell proliferation control. On the other hand, SVAR showed that *bai-1* and *tsp-1* genes are induced by the *p53* gene product in Hela cells. It is known that the *bai-1* gene codes for a member of the secretin receptor family, which contains at least one functional p53-binding site within an intron, and its product is postulated to be an inhibitor of angiogenesis and a tumor growth suppressor [40]. Similarly, the *tsp-1* gene codes for an adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions and has been shown to play a role in platelet aggregation, angiogenesis, and tumorigenesis [41]. Taken together, the p53-mediated upregulation of *bai-1* and *tsp-1* genes may be a mechanism to evade cell migration and angiogenesis, features which are commonly absent in Hela cells. We noticed that the classical p53 targets, such as *gadd45* and *p21*, do not appear to be directly regulated by p53 in the SVAR analysis (Figure 5). This may be explained by the fact that the time-series length is not large enough. It is important to note that our previous study applying DVAR (Dynamic Vector AutoRegressive) [42], it was possible to identify these connectivities.

The observed p53-independent transcriptional regulation of the *p21* gene (Figure 5), appears to be unrelated to cell cycle arrest, as discussed below.

The STAT3 protein is a member of the STAT protein family. In response to cytokines and growth factors, it forms both homo- or heterodimers with other STAT proteins and the complex translocates to the nucleus, where they act as transcriptional activators. STATs mediate the cell response to different stimuli, playing a key role in several cellular processes, such as cell growth and apoptosis [43]. As shown, using the SVAR method (Figure 5), STAT3 regulates the expression of the cycle positive regulator Cyclin D1 and of the anti-apoptotic protein Bcl-2. It has already been reported that constitutive activation of STAT 3 correlates with *cyclin d1* and *bcl-2* gene overexpression, thus providing a novel prognostic marker for head and neck squamous cell carcinoma [44]. Moreover, repression of *p53* gene expression by STAT3 is likely to have an important role in development of tumors [45]. These evidence point to an involvement of STAT3 in cell cycle progression and transformation of Hela cells.

Our *in silico* analysis also highlighted an unexpected behavior for the *p21* gene, independently of p53 regulation. This alternative regulation has already been described for other cell types [46], but still remains unclear in the case of Hela cells. Although p21 is not a transcription factor, it is conceivable that indirect effects of p21 on cellular gene expression of well-known cell cycle progression promoters, such as Cyclin D1 and apoptosis inhibitors, such as Bcl-2 may mediate some unexpected functions in Hela cells. These functions appear to be unrelated to growth inhibition and cell cycle arrest, supporting the hypothesis that p53-independent regulation of p21 could be one of the signaling pathways activated during tumorogenesis and/or tumor progression in Hela cells as well as in other cancer types [47,48]. Future

efforts directed to evaluate this hypothesis include gene transfection of p21 mutants lacking the p53 and STAT3-binding sites and subsequent, analysis of the newly identified p21 targets gene expression and changes in Hela cells phenotype and tumorigenicity.

It is interesting that, even using a small dataset, the SVAR method allowed identification of actual regulations, as detailed above, illustrating the power of this technique. In general, the methods reported in the literature are not based on a statistical test due to difficulties generated by the fact that the number of samples is lower than the number of parameters to be estimated, consequently, they do not provide an objective control for false-positives.

The main advantage of the sparse vector autoregressive model (SVAR), compared with other connectivity models, is that it models a Granger-causal network with a number of genes that is larger than the number of samples, in other words, it is useful to model "large" networks with a statistical test for each one of the edges. To the best of our knowledge, the approach taken here is the only one that combines these two advantages since other methods which model "large" networks usually do not present statistical tests for the edges. Moreover, "large" gene-gene networks are commonly dealt with in pairwise comparisons. Using SVAR, it is possible to infer partial Granger-causalities resulting in a lower number of spurious edges than pairwise comparisons.

Since SVAR deals with the multivariate case, the definition of Granger causality becomes complex, because of the existence of multi-steps connectivities. In the present report, identification of Granger causality using the SVAR model is related to the definition of partial Granger-causality given by [49]. By definition of Granger's causality [49] the SVAR model allows analysis of cycles containing networks. Therefore, there is no *a priori* assumption that the network must be a DAG (Directed Acyclic Graph), as assumed by other methods [5,9]. As a consequence, the SVAR method can be used to model networks with cycles. This is of extreme importance, since it is well-known that genetic regulatory networks maintain their control and balance by a number of positive/negative feedback cycles.

There is a class of Bayesian network with MCMC algorithm which may integrate expression data with multiple sources of information [8]. The advantages of integrating multiple sources of information, i.e., adding *a priori* knowledge, is speculative. Integration of *a priori* knowledge maybe interesting to recover more realistic connections and to increase the power of the test. However, it also lead to a bias depending on the kind of information assumed in the model. In this actual stage of development of SVAR, integration of different information is not possible since only gene expression levels are used to estimate Granger causality. Further studies may be focused on integrating biological information to improve the power of SVAR.

The experimental comparison between SVAR and other methods is difficult since SVAR is the only one which has a statistical test for gene-gene networks comprising a notion of Granger-causality. The Graphical Gaussian Models reported by Schäfer and Strimmer, which apply partial correlations in the context of ($n > m$) is the closest one to SVAR, presenting a statistical test, however, the edges obtained by this approach represent instantaneous associations (correlations), failing to provide a notion of Granger-causality, i.e., the edges have no direction.

Differently from score functions, which pose difficult interpretations or subjective choices of the threshold to determine where there is (or not) an edge, a statistical test is an objective way to determine whether there is an edge and what is the rate of type I error.

In this work, we considered only lags of first order, but it is relatively straightforward to generalize this method to analyze SVAR models with orders higher than one. However, this issue depends on the number of parameters to be estimated and the time series length.

The complexity of the proposed inference is linear to the number of genes, since only one regression is performed for each gene.

There are other approaches for variable selection based on stepwise methods. Unfortunately, these methods are not consistent when $n > m$ [50], i.e., even increasing the sample size ($T \to \infty$), there is no guarantee that the set of non-zero coefficients is the correct one. This result does not change even if all subsets of variables are explored.

In contrast to LASSO, one may choose to use other penalized regressions, such as the more popular Ridge [51] or the non-negative Garrote [52]. Ridge does not set the variables to zero, resulting in models with difficult interpretations. Comparing LASSO to non-negative Garrote, the latter is worse than LASSO when multicolinearity is present in the data [23]. Therefore, LASSO seems to be the most appropriate in identifying gene regulatory networks.

Another advantage of SVAR is the fact that it does not require model pre-specification; therefore, this method is unbiased and makes it possible to infer new connections, not just quantifying the dependence level measured by already known edges. Furthermore, it is not necessary to discretize gene expression values to Boolean variables, as

in the Boolean network models [17]; therefore, there is no loss of information.

In the SVAR approach, to render the application of statistics when ($n > m$) feasible, we used the fact that the metabolic networks are sparsely connected as part of the solution. Therefore, the number of variables to be analyzed decreases significantly, resulting only in variables whose estimated coefficients are large enough to be tested and rejected as being different from zero.

Conclusion

In summary, here we introduce the SVAR method to model gene regulatory networks in the present context, where the number of samples is often lower than the number of genes. With this method, it is possible to naturally model networks with feedback loops and to infer partial Granger causalities without any *a priori* information, which minimizes the number of spurious causalities. Moreover, we present a statistical test to control for the false discovery rate, a task which was not previously possible in several other proposed gene regulatory network models.

Methods

Firstly, we describe the classical vector autoregressive model (VAR) and, then, we explore the feasibility of using LASSO regression as part of a technique for variable selection, by introducing the sparse vector autoregressive model (SVAR). The statistical test for the edges is also presented followed by the control of the false positives. To simplify the description of these methods, we describe both the SVAR and the VAR of order one, but they could easily be generalized to higher orders. After this description, we present the algorithm to construct artificial regulatory networks based on scale-free topology, since metabolic networks were described to have power-law distributions in the nodes' degrees [25]. We use this artificial network to evaluate the performance of our proposed model. Finally, the SVAR model is applied to actual biological data.

Statistical background

Granger (1969) [53] defined a concept of causality, which is easy to deal with in the context of VAR models; therefore, it has become quite popular in recent years [54]. The idea is that a cause cannot come after the effect. Thus, in the case of VAR(1) (VAR of order one) [54], if a gene i at time (t - 1) affects another gene j at time t, the former should help to predict the target gene expression.

A first order VAR model is described as shown:

$$\gamma_t = A_1\gamma_{t-1} + \varepsilon_t \; t = 2,..., T \qquad (1)$$

where T is the time-series' length (number of microarrays) γ_t is an $n \times 1$ vector of gene expression (where n is the number of genes), the normally distributed disturbance ε_t is an $n \times 1$ vector with mean zero and covariance matrix Ω, and A_1 is an $n \times n$ matrix of parameters (connectivities). The disturbances ε_t are serially uncorrelated, but may be contemporaneously correlated. Thus $E(\varepsilon_t\varepsilon_t') = \Omega$, where Ω is an $n \times n$ matrix. It is important to highlight that, in this multivariate model, each gene may depend not only on its own past values, but, also, on the past values of the other genes. Thus if γ_{it} denotes the ith element in γ_t, the ith row yields

$$\gamma_{it} = a_{i1}\gamma_{1,t-1} + a_{i2}\gamma_{2,t-1} + ...+ a_{iN}\gamma_{N,t-1} + \varepsilon_{it},\; i = 1,...,n \qquad (2)$$

This model can be estimated by Ordinary Least Squares (OLS), simply by regressing each variable on the lags of itself and the other variables.

Therefore, we can re-write it as

$$Z = X\beta + EE_i \sim N(0, \Omega)\; i = 1,..., n \qquad (3)$$

where E_i follows a multivariate Gaussian distribution $N(0, \Omega)$, with zero mean $0_{(n\times 1)}$ and covariance matrix Ω.

We define $m = T$ - 1 and introduce the notation:

$$Z_{(m\times n)} = [\gamma_2,...,\gamma_t,...,\gamma_T]' = [z_1,...,z_i,...,z_n],$$

$$\beta_{(n\times n)} = A_1' = [\beta_1,..., \beta_n],$$

$$X_{(m\times n)} = [\gamma_1...\gamma_m]',$$

$$E_{(m\times n)} = [\varepsilon_2,...,\varepsilon_t,...,\varepsilon_T]'$$

The explicit solution of the OLS estimator is

$$\beta = (X'X)^{-1}X'Z \qquad (4)$$

Therefore, one can carry out separate regression analyses for each gene. In other words, it is possible to separately estimate each column β_i of β:

$$\hat{\beta}_i = (X'X)^{-1}X'z_i \quad i = 1,..., n \qquad (5)$$

where z_i is the i-th column of Z.

In order to specify the distribution of the *j*-th element of $\hat{\beta}$, let us denote the *j*-th diagonal element of $(X'X)^{-1}$ by w_{jj}. Then, we may assert the statistical test as

$$\frac{\hat{\beta}_{ij}}{\sqrt{\hat{\sigma}^2 w_{jj}}} \sim t(m-n) \quad i = 1,...,n \tag{6}$$

under the null hypothesis, where $t(m - n)$ denotes a *t* distribution of $(m - n)$ degrees of freedom and

$$\hat{\sigma}^2 = \frac{1}{m-n}(Z - X\hat{\beta})'(Z - X\hat{\beta}) = \frac{E'E}{m-n} \tag{7}$$

It is to point out that these definitions will work only if $m > n$. Additionally, it is also well known that OLS does not ensure sparse connectivity patterns for *A*.

To overcome these problems, in the next section, we introduce the sparse vector autoregressive model.

Sparse Vector AutoRegressive (SVAR)

Consider *Z*, β, *X* and *E* as described above.

According to [55-58], the LASSO (Least Absolute Shrinkage and Selection Operator) regression [23] can be carried out by iterative application of:

$$\hat{\beta}_i^{k+1} = (X'X + \lambda^2 D(\hat{\beta}_i^k))^{-1} X' z_i \quad i = 1,...,n \text{ and } k = 1,...,N_{it} \tag{8}$$

where N_{it} is the number of iterations (we set $N_{it} = 30$ to our analysis), λ is the regularization parameter which determines the amount of penalization enforced, $D(\beta_i^k)$ is a diagonal matrix defined by

$$D(\theta) = diag(p'_\lambda(\theta)/\theta) \quad k = 1,...,n \tag{9}$$

and

$$p'_\lambda(\theta) = \lambda sign(\theta) \tag{10}$$

At each iteration, the regression coefficients of each gene with all others are weighted according to their current size and several coefficients are successively down-weighted and set to zero.

The covariance matrix of the estimators may then be approximated by:

$$(X'X + \lambda^2 D(\beta))^{-1} X'X (X'X + \lambda^2 D(\beta))^{-1} \sigma^2 \tag{11}$$

where σ^2 is an estimate of the error variance

$$\hat{\sigma}^2 = \frac{1}{m-n-c}(Z - X\hat{\beta})'(Z - X\hat{\beta}) = \frac{E'E}{m-n-c} \tag{12}$$

and *c* is the number of variables β set to zero by LASSO regression.

When $\hat{\sigma}^2$ replaces σ^2, we get the result that the statistical test is

$$\frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2 w_{jj}}} \sim t(m-n-c) \quad i = 1,...,n \tag{13}$$

under the null hypothesis, where $t(m - n - c)$ denotes a *t* distribution of $(m - n - c)$ degrees of freedom and w_{jj} is the *j*-th diagonal element of

$$(X'X + \lambda^2 D(\beta))^{-1} X'X (X'X + \lambda^2 D(\beta))^{-1} \tag{14}$$

It is important to emphasize that the number of variables set to zero in this method will depend on the value of the regularization parameter λ, with higher values implying on the selection of fewer variables.

In our work, the value of the tuning parameter λ was selected as the value that minimizes the generalized cross validation criterion (GCV).

Let $q(\lambda) = tr\{X(X'X + \lambda^2 D(\beta))^{-1}X'\}$ and $rss(\lambda)$ be the residual sum of squares for the constrained fit with constraint λ, the generalized cross-validation statistic can be written as:

$$GCV = \frac{1}{m}\frac{rss(\lambda)}{\{1 - q(\lambda)/m\}^2} \tag{15}$$

The minimum value for GCV was achieved by the L-BFGS-B algorithm [59], which was implemented in the function *optim* of the R statistical environment.

For more details on the statistical properties of LASSO in autoregressive models see [60].

Controlling the number of false-positives

To control the type I error in cases of multiple tests of hundreds of edges, we applied the FDR method [24].

Firstly, assume that of the *n* hypotheses tested $\{H_1^0, H_2^0, ..., H_n^0\}$, where H_j^0 is the null hypothesis of the *j*-th test and $\{p(1), p(2), ..., p(n)\}$ their corresponding p-

values, n_0 are the number of true null hypotheses and the other (n - n_0) hypotheses are false.

Let $p(1) \leq p(2) \leq ... \leq p(n)$ be the ordered observed p-values of each test. Define

$$l - max\{i : p(i) \leq \frac{i}{n} q\} \quad (16)$$

and reject $H^0_{(1)} ... H^0_{(l)}$. If no such i exists, reject all null hypothesis.

FDR is defined as the expected proportion (q) of incorrectly rejected null hypotheses (type I error) in a list of all rejected hypotheses.

Artificial regulatory networks

The description that many networks in nature have a power-law degree distribution was first addressed by [61]. In their random graph model, called scale-free graph, it is described how these networks grow and expand, being based on two generic mechanisms, which are common to several networks in the real world. Several networks in the real world start from a small number of nodes and grow by continuous addition of new nodes, therefore, the number of nodes increases throughout the lifetime of the network. When a new node is added to the network, its attachment is preferential, i.e., the probability of a new node connects to the existing nodes is not uniform as in a random graph [62]. There is a higher probability to be linked to a node that already has a large number of connections, resulting in a power-law degree distribution. In other words, the probability $P(v)$ that a node in the network is connected to v other nodes decays as a power-law. Therefore, the degree distribution has a power-law tail $P(v) \sim v^{-\gamma}$, where γ is a scalar which represents the rate of decayment of the degree distribution. In our case, the nodes are representing the genes and the connections are the Granger-causal relationships.

This scale-free graph can be constructed as below:

1. Growth: Starting with a small number z_0 of genes, at each iteration, a new gene with $z \leq (z_0)$ edges are added. This new gene is connected to the genes already present in the network with a preferential attachment.

2. Preferential attachment: The gene with which the new gene will connect is selected in a non-deterministic fashion. Assume that the probability π that a new gene will be connected to gene i depends on the degree d_i of that gene which is already in the network. Therefore:

$$\pi(d_i) = \frac{d_i}{\sum_j d_j} \quad (17)$$

Since we are interested in causal relationships, we need to define a direction for each edge. Therefore, there is a third step in our graph construction. In our simulations, the probability attributed to add an edge from i to j is the same from j to i, i.e., 0.5.

After T_{step} iterations, the constructed random scale-free like network is composed of $n = T_{step} + z_0$ genes and $z * T_{step} + z_{edges}$ Granger-causal relationships, where z_{edges} is the initial number of edges.

The graph constructed using the algorithm described above may be represented by its adjacency matrix A, i.e., where there is an edge from gene i to gene j it was set to $A[i, j] = 0.8$, and 0 otherwise, in our simulations. This adjacency matrix A corresponds to the matrix A described in equation 1. The time-series' lag was set to one in our simulations, therefore, set $m = T - 1$.

To construct the corresponding time-series for each gene, firstly, generate normally distributed random numbers with zero mean and unit variance for each gene $i = 1,...,n$ for the time step $t = 1$, $\gamma_{i1} = \varepsilon_i$. Then, use equation 2 to generate the time-series for each gene $i = 1,...,n$, time step $t = 2,...,T$.

Implementation

We implemented our program using R [63], a statistical computing environment. Computation was conducted under a Pentium IV CPU 3.06 GHz, 2.5 GB of RAM.

Application to real data

We applied the SVAR approach to HeLa cell cycle gene expression data collected by Whitfield *et al.* (2002) [64]. Gene expression was measured using microarrays manufactured in the Stanford Microarray Facility. The data used contain 48 time points distributed at one hour intervals with one reading at each time point, synchronized by double thymidine block (described as Experiment 3 in the web page [65]). The 94 genes were selected from actual biological microarray data on the basis of there association with cell cycle regulation and tumor development. The HeLa cell cycle lasts 16 hours. These data were downloaded from: [65].

Authors' contributions

AF has made substantial contributions to the conception and design of the study, analysis and interpretation of data. JRS has made substantial contributions to the analysis and interpretation of mathematical results. HMGM has made substantial contributions to the analysis and inter-

pretation of biological data. AF, JRS and HMGM have been involved in drafting of the manuscript. RY and SM have discussed the mathematical results. MCS has discussed the biological results. CEF has directed the work. RY, SM, MCS and CEF critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by JICA, FAPESP, CAPES, CNPq, FINEP and PRP-USP.

References

1. Gardner T, di Bernardo D, Lorenz D, Collins J: **Inferring genetic networks and identifying compound mode of action via expression profiling.** *Science* 2003, **301:**102-105.
2. di Bernardo D, Thompson M, Gardner T, Chobot S, Eastwood E, Wojtovich A, Elliott S, Schaus S, Collins J: **Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.** *Nature Biotechnology* 2005, **23:**377-383.
3. Faith J, Hayete B, Thaden J, Mogno I, Wierzbowski J, Cotterel G, Kasif S, Collins J, Gardner T: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a Compedium of expression profiles.** *PLoS Biology* 2007, **5:**e8.
4. Imoto S, Goto T, Miyano S: **Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression.** *Pac Symp Biocomput* 2002:175-186.
5. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S: **Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection.** *Bioinformatics* 2003, **19:**227-236.
6. Friedman N: **Inferring cellular networks using probabilistic graphical models.** *Science* 2004, **303:**799-805.
7. Dojer N, Gambin A, Mizera A, Wilczynski B, Tiuryn J: **Applying dynamic Bayesian networks to perturbed gene expression data.** *BMC Bioinformatics* 2006, **7:**249.
8. Werhli A, Husmeier D: **Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge.** *Stat Appl Genet Mol Biol* 2007, **6:**15.
9. Xiong M, Li J, Fang X: **Identification of genetic networks.** *Genetics* 2004, **166:**1037-1052.
10. Akutsu T, Miyano S, Kuhara S: **Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.** *J Comput Biol* 2000, **7:**331-343.
11. Shmulevich I, Dougherty E, Zhang W: **Gene perturbation and intervention in probabilistic Boolean networks.** *Bioinformatics* 2002, **18:**1319-1331.
12. Pal R, Datta A, Bittner M, Dougherty E: **Intervention in context-sensitive probabilistic Boolean networks.** *Bioinformatics* 2005, **21:**1211-1218.
13. Schäfer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks.** *Bioinformatics* 2005, **21:**754-764.
14. Woolf P, Wang Y: **A fuzzy logic approach to analyzing gene expression data.** *Physiol Genomics* 2000, **3:**9-15.
15. Mestl T, Plahte E, Omholt S: **A mathematical framework for describing and analyzing gene regulatory networks.** *J theor Biol* 1995, **176:**291-300.
16. Vapnik V: *The nature of statistical learning theory* New York: Springer; 1995.
17. Barrera J, Cesar RJ, Martins DJ, Merino E, Vêncio R, Leonardi F, Yamamoto M, Pereira C, del Portillo H: **A new annotation tool for malaria based on inference of probabilistic genetic networks.** *Critical Assessment of microarray data analysis: 10–12 November 2004; Durham* 2004:36-40.
18. 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.
19. Xu X, Wang L, Ding D: **Learning module networks from genome-wide location and expression data.** *FEBS Lett* 2004, **578:**297-304.
20. Yamaguchi R, Yoshida R, Imoto S, Higuchi T, Miyano S: **Finding module-based gene networks in time-course gene expression data with state space models.** *IEEE Signal processing magazine* 2007.
21. Valdes-Sosa P, Sanchez-Bornot J, Lage-Castellanos A, Vega-Hernandez M, Bosch-Bayard J, Melie-Garcia L, Canales-Rodriguez E: **Estimating brain functional connectivity with sparse multivariate autoregression.** *Phil Trans R Soc B* 2005, **360:**969-981.
22. Eichler M: **A graphical approach for evaluating effective connectivity in neural systems.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360:**953-967.
23. Tibshirani R: **Regression shrinkage and selection via the Lasso.** *Journal of the Royal Statistical Society Series B* 1996, **58:**267-288.
24. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc Ser B* 1995, **57:**289-300.
25. Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A: **The large-scale organization of metabolic networks.** *Nature* 2000, **65:**651-654.
26. Inoue J, Gohda J, Akiyama T, Semba K: **NF-kappaB activation in development and progression of cancer.** *Cancer Sci* 2007, **98:**268-274.
27. Soussi T: **p53 alterations in human cancer: more questions than answers.** *Oncogene* 2007, **26:**2145-2156.
28. Yu H, Kortylewski M, Pardoll D: **Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment.** *Nat Rev Immunol* 2007, **7:**41-51.
29. Albert R, Jeong H, Barabasi A: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-385.
30. Chen F, Castranova V, Shi X: **New insights into the role of nuclear factor-kappaB in cell growth regulation.** *Am J Pathol* 2001, **159:**387-397.
31. Krikos A, Laherty C, Dixit V: **Transcriptional activation of the tumor necrosis factor alpha-inducible zinc finger protein, A20, is mediated by kappa B elements.** *J Biol Chem* 1992, **267:**17971-17976.
32. You M, Ku P, Hrdlickova R, Bose HJ: **ch-IAP1, a member of the inhibitor-of-apoptosis protein family, is a mediator of the antiapoptotic activity of the v-Rel oncoprotein.** *Mol Cell Biol* 1997, **17:**7328-7341.
33. Chen M, Ghosh G: **Regulation of DNA binding by Rel/NF-kappaB transcription factors: structural views.** *Oncogene* 1999, **377:**6845-6852.
34. Brown R, Ades I, Nordan R: **An acute phase response factor/NF-kappa B site downstream of the junB gene that mediates responsiveness to interleukin-6 in a murine plasmacytoma.** *J Biol Chem* 1995, **270:**31129-21135.
35. Storz P, Doppler H, Ferran C, Grey S, Toker A: **Functional dichotomy of A20 in apoptotic and necrotic cell death.** *Biochem J* 2005, **387:**47-55.
36. Dreyfus D, Nagasawa M, Gelfand E, Ghoda L: **Modulation of p53 activity by IkappaBalpha: evidence suggesting a common phylogeny between NF-kappaB and p53 transcription factors.** *BMC Inmunol* 2005, **6:**12.
37. Jin S, Levine A: **The p53 functional circuit.** *J Cell Sci* 2001, **114:**4139-4140.
38. Buckbinder L, Talbott R, Velasco-Miguel S, Takenaka I, Faha B, Seizinger B, Kley N: **Induction of the growth inhibitor IGF-binding protein 3 by p53.** *Nature* 1995, **377:**646-649.
39. Yakovlev A, Di Giovanni S, Wang G, Liu W, Stoica B, Faden A: **BOK and NOXA are essential mediators of p53-dependent apoptosis.** *J Biol Chem* 2004, **279:**28367-28374.
40. Fukushima Y, Oshika Y, Tsuchida T, Tokunaga T, Hatanaka H, Kijima H, Yamazaki H, Ueyama Y, Tamaoki N, Nakamura M: **Brain-specific angiogenesis inhibitor 1 expression is inversely correlated with vascularity and distant metastasis of colorectal cancer.** *Int J Oncol* 1998, **13:**967-970.
41. Dameron K, Volpert O, Tainsky M, Bouck N: **Control of angiogenesis in fibroblasts by p53 regulation of thrombospondin-1.** *Science* 1994, **265:**1582-1584.
42. Fujita A, Sato J, Garay-Malpartida H, Morettin P, Sogayar M, Ferreira C: **Time-varying modeling of gene expression regulatory net-**

works using the wavelet dynamic vector autoregressive method. *Bioinformatics* 2007, **23:**1623-1630.
43. Jing N, Tweardy D: **Targeting Stat3 in cancer therapy.** *Anticancer Drugs* 2005, **16:**601-607.
44. Masuda M, Suzui M, Yasumatu R, Nakashima T, Kuratomi Y, Azuma K, Tomita K, Komiyama S, Weinstein I: **Constitutive activation of signal transducers and activators of transcription 3 correlates with cyclin D1 overexpression and may provide a novel prognostic marker in head and neck squamous cell carcinoma.** *Cancer Res* 2002, **62:**3351-3355.
45. Niu G, Wright K, Ma Y, Wright G, Huang M, Irby R, Briggs J, Karras J, Cress W, Pardoll D, Jove R, Chen J, Yu H: **Role of Stat3 in regulating p53 expression and function.** *Mol Cell Biol* 2005, **25:**7432-7440.
46. Roninson I: **Oncogenic functions of tumour suppressor p21(Waf1/Cip1/Sdi1): association with cell senescence and tumour-promoting activities of stromal fibroblasts.** *Cancer Lett* 2002, **179:**1-14.
47. Gartel A: **Is p21 an oncogene?** *Mol Cancer Ther* 2006, **5:**1385-1386.
48. De la Cueva E, Garcia-Cao I, Herranz M, Lopez P, Garcia-Palencia P, Flores J, Serrano M, Fernandez-Piqueras J, Martin-Caballero J: **Tumorigenic activity of p21Waf1/Cip1 in thymic lymphoma.** *Oncogene* 2006, **25:**4128-4132.
49. Hosoya Y: **Elimination of third-series effect and defining partial measures of causality.** *Journal of time series analysis* 2001, **22:**537-554.
50. Hastie T, Tibshirani R, Friedman J: **The elements of statistical learning: data mining, inference, and prediction.** *Econometrica* 1969, **37:**424-438.
51. Hoerl A, Kennard R: **Ridge regression: biased estimation for non-orthogonal problems.** *Technometrics* 1970, **12:**55-67.
52. Breiman L: **Better subset regression using the nonnegative garrote.** *Technometrics* 1995, **37:**373-384.
53. Granger C: **Investigating causal relation by econometric and cross-sectional method.** *Econometrica* 1969, **37:**424-438.
54. Mukhopadhyay N, Chatterjee S: **Causality and pathway search in microarray time series experiment.** *Bioinformatics* 2007, **23:**442-449.
55. Fan J, Li R: **Variable selection via nonconcave penalized likelihood and its oracle properties.** *J Am Stat Assoc* 2001, **96:**1348-1360.
56. Fan J, Peng H: **Nonconcave penalized likelihood with a diverging number of parameters.** *Ann Stat* 2004, **32:**928-961.
57. Hunter D: **MM algorithms for generalized Bradley-Terry models.** *Ann Stat* 2004, **32:**384-406.
58. Hunter D, Lange K: **A tutorial on MM algorithms.** *Am Stat* 2004, **58:**30-37.
59. Bryd R, Peihuang L, Nocedal J, Ciyou Z: **A limited memory algorithm for bound constrained optimization.** *SIAM J Scientific Computing* 1995, **16:**1190-1208.
60. Wang H, Li G, Tsai C: **Regression coefficient and autoregressive order shrinkage and selection via the lasso.** *J R Statist SocB* 2007, **69:**63-78.
61. Barabási A, Albert R: **Emergence of scaling in randomnetworks.** *Science* 2000, **286:**509-512.
62. Erdös P, Rényi A: **On random graphs.** *Publicationes Mathematicae* 1959, **6:**290-297.
63. **The R project for statistical computing** [http://www.r-project.org]
64. Whitfield M, Sherlock G, Saldanha A, Murray J, Ball C, Alexander K, Matese J, Perou C, Hurt M, Brown P, Botstein D: **Identification of genes periodically expressed in the human cell cycle and their expression in tumors.** *Molecular Biology of the Cell* 2002, **13:**1977-2000.
65. **Human cell cycle: HeLa cells** [http://genome-www.stanford.edu/Human-CellCycle/HeLa/]