Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Gene networks as a tool to understand transcriptional regulation

Diogo Fernando Veiga, Fábio Fernandes da Rocha Vicente and Gustavo Bastos

Laboratório de Bioinformática, Centro de Informática,
Universidade Federal de Pernambuco, Caixa Postal 7851,
50732-970 Recife, PE, Brasil
The present address of D.F. Veiga and F.F.R. Vicente is
Laboratório Nacional de Computação Científica,
Laboratório de Bioinformática, Av. Getúlio Vargas, 333,
Petrópolis, RJ, Brasil
Corresponding author: D.F. Veiga
E-mail: dfv@cin.ufpe.br

Genet. Mol. Res. 5 (1): 254-268 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. Gene regulatory networks, or simply gene networks (GNs), have shown to be a promising approach that the bioinformatics community has been developing for studying regulatory mechanisms in biological systems. GNs are built from the genome-wide high-throughput gene expression data that are often available from DNA microarray experiments. Conceptually, GNs are (un)directed graphs, where the nodes correspond to the genes and a link between a pair of genes denotes a regulatory interaction that occurs at transcriptional level. In the present study, we had two objectives: 1) to develop a framework for GN reconstruction based on a Bayesian network model that captures direct interactions between genes through nonparametric regression with B-splines, and 2) to demonstrate the potential of GNs in the analysis of expression data of a real biological system, the yeast pheromone response pathway. Our framework also included a number of search schemes to learn the network. We present an intuitive notion of GN theory as well as the detailed mathematical foundations of the model. A comprehensive anal-

ysis of the consistency of the model when tested with biological data was done through the analysis of the GNs inferred for the yeast pheromone pathway. Our results agree fairly well with what was expected based on the literature, and we developed some hypotheses about this system. Using this analysis, we intended to provide a guide on how GNs can be effectively used to study transcriptional regulation. We also discussed the limitations of GNs and the future direction of network analysis for genomic data. The software is available upon request.

Key words: Gene networks, Bayesian networks, Transcriptional regulation, Pheromone response pathway, *Saccharomyces cerevisiae*

INTRODUCTION

Gene regulatory networks, or simply gene networks (GNs), have shown to be a promising approach that the bioinformatics community has been developing for studying regulatory mechanisms in biological systems. These GNs are built from the genome-wide high-throughput gene expression data (Figure 1) that are often available from DNA microarray experiments (Friedman, 2004). Transcriptome data have become increasingly important because they provide an appropriate source for studies of the systemic behavior of complex biological systems, and the GNs are one of the computational tools that have been employed to this end.

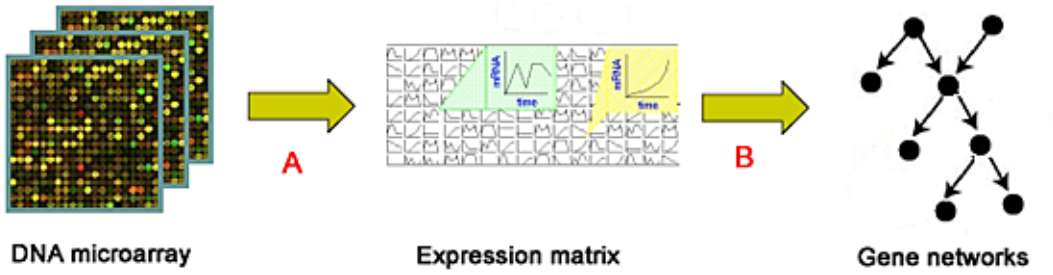


Figure 1. General steps in recovering gene networks (GNs) from DNA microarray data. **A.** Image processing and normalization of data. **B.** Inference of the network using expression data.

A DNA microarray experiment, which measures the levels of genes being expressed, can be both designed to generate time series expression data (sampling at different times under the same conditions) or it can be designed in a static manner, in which one takes snapshots from cells under a variety of physiological conditions (i.e., from different tissues). The microarray generates the mRNA levels of transcripts, and these data can be used to reveal the hidden "program" that controls gene expression in that biological system.

The expression program of a cell determines which group of genes is going to shift from the basal to active transcription, and which genes are going to be inhibited. Several physical and

mechanical systems are modeled as stochastic processes susceptible to noisy measures (Roweis and Ghahramani, 1999). We can also borrow this approach to analyze biological systems, as we consider that the system that we want to model is a program that controls gene expression at the molecular level. In this system, we have a number of regulatory mechanisms, represented by the transcriptional control strategies of the cell, such as the synthesis of specialized proteins as transcription factors. We also know that this expression program works differently according to external stimuli and to intracellular conditions, such as the concentration of some metabolites. Furthermore, the output of this system is the mRNA levels of the genes that are involved. These measures are, in turn, highly noisy, due to experimental and computational factors. In this scenario, the question that arises is "What is the structure of the system that has generated this expression data?" or equivalently, "How do genes interact among themselves to generate this output?" The GNs aim to recover these interactions and therefore, the underlying system that works at the transcriptional level, through so-called probabilistic graphical models.

Conceptually, GNs are (un)directed graphs, in which the nodes correspond to the genes and a link between a pair of genes denotes a regulatory interaction between them (Figure 2). Since we are assessing mRNA profiles, a link between two genes represents the relationship that occurs at the transcriptional level, i.e., a transcription factor that is activating its target gene.

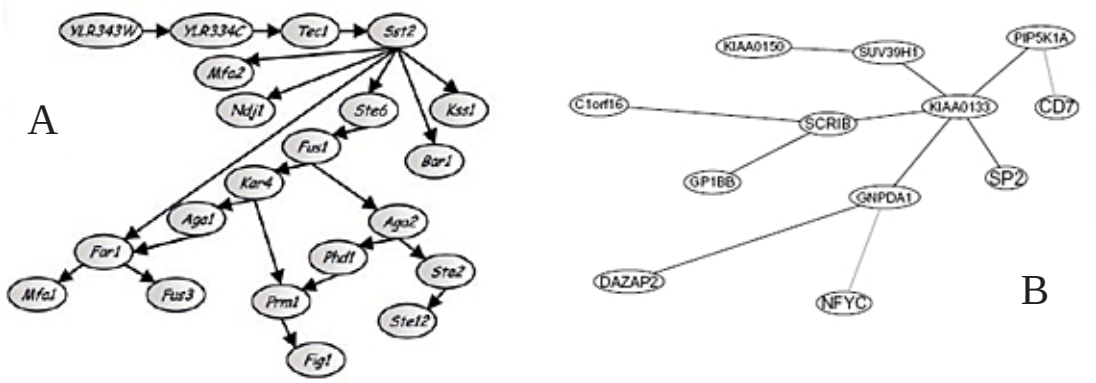


Figure 2. Graphical model representations of gene networks. In **A**, the gene network is an acyclic directed graph (Friedman, 2004), Bayesian network, which denotes the direction of regulation and **B** an undirected graph (Schäfer and Strimmer, 2005), where an edge only means that there is some kind of association between the linked genes.

For the purpose of inference of GN from data, graphical models provide the mathematical theory to determine the gene interactions. Since the pioneer work of Friedman et al. (2000), several models have been proposed (for a review, see Xia et al. (2004), van Someren et al. (2002), and de Jong (2002)). Bayesian networks (BN) are of special interest for two main reasons: 1) the ability to handle noisy data, such as microarray outputs, and 2) prior knowledge can easily be incorporated into the model.

We developed a framework for GN reconstruction, based on a BN model that captures direct interactions between genes through nonparametric regression and that has multiple search schemes for learning the network. The base functions used for regression are B-splines. This approach of function approximation by nonparametric regression leads to flexibility of the model, since it does not assume any fixed regulation function. The model has a number of search

schemes that are based on two search algorithms, known as random hill-climbing and greedy hill-climbing. This framework has many options for initializing the start graph, including empty, random, score-driven, and partial correlation initialization.

METHODS

Before the mathematical enunciation of the model, it is important to discuss some principles behind GN inference. The main objective of an algorithm to infer GNs is to find the correct set of regulators for the genes in the net. Within the BN terminology, this means finding the correct parents for each node. As an example, we assume that expression levels of a regulated gene *a* are dictated by its parents genes. If gene *a* is mainly regulated by genes *b* and *c*, on a universe of *n* genes, the goal is to identify, among all possible combinations of *n*-1 other genes, that the ideal regulators (parents) for *a* are *b* and *c*. We also should find, from expression data, that the mRNA levels of *a* have their best approximation when we take them as a function of the parents' expression, that is, $E(a) \approx F(E(b), E(c))$. Hence, we need to make use of a mathematical tool to perform this approximation of gene expression given its parents.

The model uses nonparametric regression with B-splines to capture the relationships between genes, as introduced by Imoto et al. (2002) and further developed by Bastos and Guimarães (2005). While some BN models for GNs use a fixed regulation law (Nachman et al., 2004), our approach is flexible on this issue, not restricting the interactions to behavior to a fixed function.

Bayesian network model for gene networks

Let $X = (X_1, X_2, \ldots, X_n)^T$ be a random *n*-dimensional vector containing the genes to be analyzed, and assume that *G* is a directed graph. Under BN theory, genes are random variables, and it is possible to decompose their joint probability into a product of conditional probabilities:

$$P(X_1, X_2, \ldots, X_n) = P(X_1 \mid P_1)\, P(X_2 \mid P_2) \times \ldots \times P(X_n \mid P_n) \qquad \text{(Equation 1)}$$

where $P_i = (P_1^{(i)}, P_2^{(i)}, \ldots, P_{qi}^{(i)})^T$ is a *qi*-dimensional vector of parent variables of X_i in *G*.

Suppose that *s* observations $x_1, x_2, \ldots, x_s$ of random vector *X* are performed, and the observations of P_i are denoted $p_{1i}, p_{2i}, \ldots, p_{si}$, where $p_{ji} = (p_{j1}^{(i)}, \ldots, p_{jqi}^{(i)})^T$ is a *qi*-dimensional observation vector of parent genes. For example, assume X_{ns} as an $n \times s$ matrix, given that each of its components x_i (i = 1, . . . , *n*) is a vector of length *s*.

Therefore, $X_{ns} = (x_1, \ldots, x_n)^T = (x_{(1)}, \ldots, x_{(s)}) = (x_{ij})_{i=1,\ldots,n\,;\, j=1,\ldots,s}$, $x_i = (x_{i1}, \ldots, x_{is})$, $x_j = (x_{1j}, \ldots, x_{nj})^T$, and x_i^T is the transposed vector of x_i. If X_1, for instance, has a parent vector $P_1 = (X_4, X_5)^T$, we can obtain $p_{11} = (x_{14}, x_{15})^T, \ldots, p_{s1} = (x_{s4}, x_{s5})^T$.

Equation 1 still holds if we replace the *P* probabilities by density functions:

$$f(x_{1j}, x_{2j}, \ldots, x_{nj}) = f_1(x_{1j} \mid P_j^{(1)})\, f_2(x_{2j} \mid P_j^{(2)}) \times \ldots \times f_n(x_{nj} \mid P_j^{(n)}). \qquad \text{(Equation 2)}$$

where x_{ij} is a particular value taken by X_{ij}. We then need to construct the conditional densities $f_i(x_{ij} \mid p_j^{(i)})$, where i = 1, . . . , *n* and j = 1, . . . , *s*. Following Tamada et al. (2003) and Imoto et al. (2002), we used nonparametric regression models for capturing the relations between the

expression of gene *i* on sample *j*, x_{ij} and the expression levels of parents on *j*, $p_{ji} = (p_{j1}^{(i)}, \ldots, p_{jqi}^{(i)})^T$ through the equation:

$$x_{ij} = m_1(p_{j1}^{(i)}) + m_2(p_{j2}^{(i)}) + \ldots + m_{qi}(p_{jqi}^{(i)}) + \varepsilon_{ij},\ i = 1, \ldots, n; j = 1, \ldots, s, \quad \text{(Equation 3)}$$

where $m_k (k = 1, \ldots, qi)$ are smoothing functions from $\Re$ (the set of real numbers) to $\Re$, and ε_{ij} follows a normal distribution with mean 0 and variance σ_i^2. For each function m_k, we assume that:

$$m_k(p_{jk}^{(i)}) = \sum_{m=1}^{M_{ik}} \gamma_{mk}^{(i)} b_{mk}^{(i)}(p_{jk}^{(i)}),\ j = 1, \ldots, s;\ k = 1, \ldots, q_i, \quad \text{(Equation 4)}$$

where $\{b_{1k}^{(i)}, b_{2k}^{(i)}, \ldots, b_{Mik}^{(i)}\}$ is a preconceived set of "basis functions" (such as Fourier series, polynomial bases, B-splines, and others), the coefficients $\gamma_{1k}^{(i)}, \ldots, \gamma_{Mik}^{(i)}$ are unknown parameters, and M_{ik} is the number of basis functions.

In our case, the basis functions are degree 3 B-splines; the coefficients $\gamma_{1k}^{(i)}, \ldots, \gamma_{Mik}^{(i)}$ are estimated from data, and $M_{ik} = 20$. There is a special algorithm to learn the parameters from data (Eilers and Marx, 1996). We do not detail how B-splines work as function approximation, in nonparametric regression; it is enough to know that they will "construct" a regulation function for each gene with the parent expression values. As we know the parents' values $p_{ji} = (p_{j1}^{(i)}, \ldots, p_{jqi}^{(i)})^T$, the B-splines generate an estimated expression for gene *i*, $\hat{x}_{ij}$.

It remains to evaluate the quality of the B-spline approximation for a given set of parents P_i. This is done with the probability density function customized for a nonparametric regression model, using B-splines (Imoto and Konishi, 2000), when the i*th* gene has *qi* parents:

$$f_i(x_{ij} \mid \mathrm{P}_j^{(i)}; \gamma_i; \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\{x_{ij} - \hat{x}_{ij}\}^2}{2\sigma_i^2}\right]. \quad \text{(Equation 5)}$$

$$\hat{x}_{ij} = \sum_{k=1}^{qi} m_k(p_{jk}^{(i)}).$$

This result allows us to define the BN:

$$f(x_{ij} \mid \theta_G) = \prod_{i=1}^{n} f_i(x_{ij} \mid \mathrm{p}_j^{(i)}; \theta_i),\ j = 1, \ldots, s, \quad \text{(Equation 6)}$$

where $\theta_G = (\theta_1^T, \ldots, \theta n^T)^T$ is a parameter vector in graph *G*, and θ_i is a parameter vector in the model of f_i, that is, $\theta_i = (\gamma_i^T, \sigma_i^2)^T$ or $\theta_i = (\mu_i, \sigma_i^2)^T$.

Multiple search schemes and graph selection criterion

To learn the GN from data, we need a search algorithm that starts with an initial solution and traverses, using a criterion, across the space of possible networks to find the optimal network. To guide the search, we use scoring functions, also called evaluation criteria, which

assign scores to each network according to its adherence to the observed data. Most of the evaluation criteria used to analyze a network use a common idea, the maximization (or minimization) of the posterior probability,

$$p(G|\mathrm{X}) = \frac{p(G)\,p(\mathrm{X}|G)}{p(\mathrm{X})}, \qquad \text{(Equation 7)}$$

where p(G) is the prior probability of the graph G, and p(X | G) is the probability of the observations given the graph. The term p(X) (prior probability of the data) is constant and not related to p(G), and therefore will not be taken into account in the model's evaluation. We use the Bayesian Information Criterion score, which has a term to measure the accuracy of the model in predicting data and another term to penalize the complexity of the model. For that reason, the more complex the graph structure, the worse its evaluation.

The space of solutions determined by all the acyclic directed graphs (a BN) grows exponentially with the number of nodes (Chickering, 1996). Under this scenario, we have implemented two heuristic search algorithms to seek for the optimal net, random hill-climbing and greedy hill-climbing (Ott et al., 2004). Briefly, both random hill-climbing and greedy hill-climbing make local changes in the network, through the following operations: 1) add a parent, 2) remove a parent, 3) reverse link direction, and 4) nothing. The difference between them is that the random approach chooses the modification at random, and only accepts it if it improves the overall score, while the greedy one evaluates the best possible move, applying the operators for the current network. Our implementation uses the greedy hill-climbing with many random restarts in order to better explore the search space.

The framework also has a variety of initialization options that define the starting graph for the learning procedure. We can begin the search with an empty graph, a random graph, a score-driven generated graph, and a graph that starts with interactions found by the partial correlation analysis (de La Fuente et al., 2004).

Next we present the results on applying the model to artificial and real biological expression data.

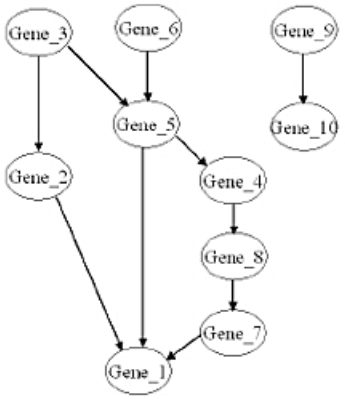
RESULTS AND DISCUSSION

Comparing search schemes

To compare the multiple search schemes available in the framework, we have used an artificial GN with 10 genes (Imoto et al., 2002). The artificial GN was sampled 100 times, each one being perturbed by a Gaussian noise (to mimic microarray experiments). This network, as well as the regulation functions that associate the genes, can be seen in Figure 3.

For random hill-climbing, we tested five different configurations for the initial network:

A. Score-driven: the parents of each gene are defined by the Bayesian information criterion score;
B. Random: the number of parents is sampled from a uniform distribution and they are chosen randomly among the genes;



Regulation functions

$$x_1 = x_2^2 + 2\sin(x_5) - 2x_7 + \varepsilon_1, \qquad x_2 = [1 + \exp(-4x_3)]^{-1} + \varepsilon_2$$
$$x_3 = \varepsilon_3, \qquad x_4 = \frac{x_5^6}{3} + \varepsilon_4$$
$$x_5 = x_3 - x_6^2 + \varepsilon_5, \qquad x_6 = \varepsilon_6$$
$$x_7 = \begin{cases} -1 + \varepsilon_7, & x_8 \leq -0.5 \\ x_8 + \varepsilon_7, & -0.5 < x_8 \leq 0.5 \\ 1 + \varepsilon_7, & 0.5 < x_8 \end{cases} \qquad x_8 = \frac{\exp(-x_4 - 1)}{2} + \varepsilon_8$$
$$x_9 = \varepsilon_9 \qquad x_{10} = \cos x_9 + \varepsilon_{10}$$

$$\varepsilon_1 = \varepsilon_4 = \varepsilon_5 = N(0, 0.4)$$
$$\varepsilon_3 = \varepsilon_6 = \varepsilon_9 = N(0, 0.5)$$
$$\varepsilon_2 = \varepsilon_7 = \varepsilon_8 = \varepsilon_{10} = N(0, 0.1)$$

Figure 3. Artificial gene network used for simulation studies, along with hypothetical regulation functions. ε coefficients represent the Gaussian noise that affects the expression of each gene.

C. Partial correlation: we input the initial graph with the links identified by the partial correlation analysis;
D. Partial correlation + score-driven: the links from the correlation analysis plus a score-driven initialization;
E. Partial correlation + random: the links from the correlation analysis, followed by a random initialization.

We performed 100 executions with each scheme and evaluated the distribution of the score for the learned networks (Figure 4). The lower the score, the better the learned GN. We see from this simulation that configurations A, B, and D produced solutions with scores in the interval $[3.5 \times 10^3, 3.55 \times 10^3]$, which suggests that they had analogous performances. Configuration E generated solutions within two intervals. The best scores were achieved by scheme C, in which we only used the links found by the correlation analysis to start the graph, producing scores ranging from 3.31×10^3 to 3.36×10^3.

The same procedure was used with the greedy hill-climbing. After all simulations, we applied a *t*-test to decide which configuration was the best. The *t*-test points to scheme C, that is random hill-climbing and partial correlation initialization, as the best configuration to begin the search with.

Pheromone response pathway

As an application using a real biological system, we have applied the BN model to the *Saccharomyces cerevisiae* pheromone response pathway. This signaling pathway is responsible for a series of complex physiological changes in preparation for mating in haploid yeast cells, including changes in the expression of about 200 genes, arrest in the G1 phase of the cell cycle and membrane fusion of mating partners (Bardwell, 2005).

The transmission of the signal starts when the pheromone peptide is recognized by cell surface receptors, leading to a signaling cascade that ultimately activates genes needed for mating. STE2 acts as the specific MATα cell pheromone receptor, while STE3 is the receptor

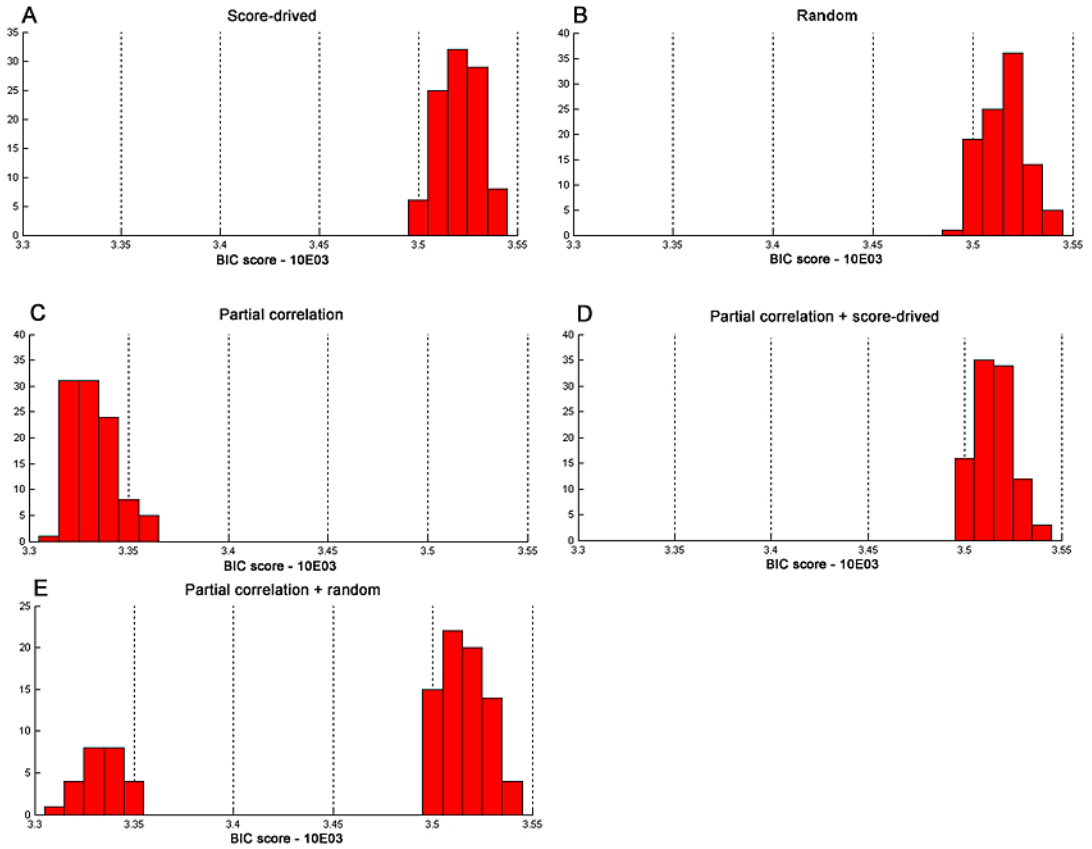


Figure 4. Histograms of the final scores for 100 executions of the framework, under different initialization schemes (see text for details) and using the random hill-climbing search. BIC = Bayesian information criterion.

for the MATa cells (Lewin, 2000). After pheromone recognition, the G-protein coupled to the pheromone receptor (a trimeric complex formed by GPA1, STE4 and STE18) recruits its effectors in order to propagate the signal to an MAPK signaling cascade that finally activates the STE12 transcription factor, the major player that will command the DNA transcription. The G-protein effectors are the STE5/STE11 and FAR1/CDC24 complexes as well as the STE20 protein kinase (Bardwell, 2005). A schematic view of the mating pathway is shown in Figure 5.

We generated GNs for two microarray datasets. As the first dataset, we used a subset of 12 genes from the cell cycle expression data generated by Spellman et al. (1998). These data have 18 samples (time series experiment). Ten of the genes, STE2 or STE3, STE18, STE20, STE5, STE11, STE7, FUS3, GPA1, FAR1, and STE12, act directly on the mating pathway, while AGA1 and FUS1 are required for cell fusion (Bardwell, 2005). Figure 6 shows the GNs (GN-1 and GN-2) inferred from this dataset, using STE2 and STE3, respectively. The nodes were colored depending on the known function of the gene (see Table 1 for the mapping). For this reason, it is desirable that genes with related functions occur near each other in the graph.

For each experiment, we ran 600 executions (i.e., 600 learned networks) and the output network was defined by a "voting criterion". In conformity with this criterion, we chose a threshold, and only the links that appeared above this threshold across all executions composed

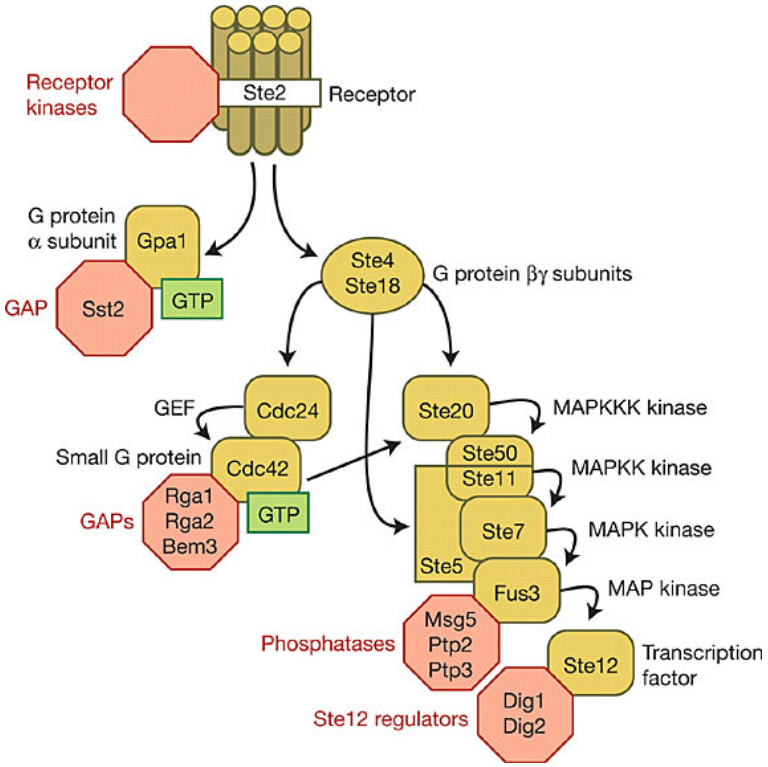


Figure 5. Pheromone response signaling pathway. Steps in the transmission of the initial signal (pheromone peptide recognition) from the membrane to the nucleus (STKE, 2005).

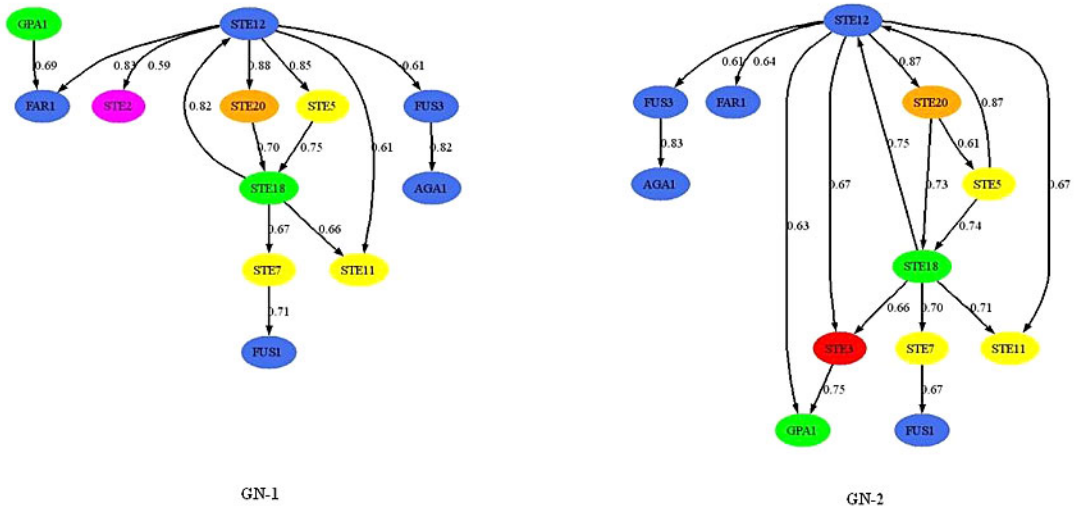


Figure 6. Gene networks (GN) inferred from the Spellman et al. (1998) yeast dataset. Experiment 1 (GN-1): 12 genes involved with the pheromone response pathway, including STE2. Experiment 2 (GN-2): 12 genes, including STE3. In both networks, STE12 is the gene that controls regulation at the transcriptional level in the pheromone response pathway, which is fairly consistent with biological knowledge. See text for description of genes and analysis of the recovered networks. The nodes are colored according to function (Table 1).

Table 1. Color and function of genes set that appear in the networks (adapted from Hartemink et al., 2002).

Gene	Color	Function of corresponding protein
STE2	magenta	transmembrane receptor peptide (present only in MATa strains)
STE3	red	transmembrane receptor peptide (present only in MATα strains)
GPA1	green	component of the heterotrimeric G-protein (Gα)
STE18	green	component of the heterotrimeric G-protein (Gβ)
FUS3	blue	mitogen-activated protein kinase (MAPK)
STE7	yellow	MAPK kinase (MAPKK)
STE11	yellow	MAPKK kinase (MAPKKK)
STE5	yellow	scaffolding peptide holding together Fus3, Ste7, and Ste11 in a large complex
STE12	blue	transcriptional activator
STE20	orange	p21-activated protein kinase (PAK)
FAR1	blue	substrate of Fus3 that leads to G1 arrest; known to bind to STE4 as part received of complex of proteins necessary for establishing cell polarity required for shmoo formation after mating signal has been received
FUS1	blue	required for cell fusion during mating
AGA1	blue	anchor subunit of α-agglutinin complex; mediates attachment of Aga2 to cell surface

the resultant network. In our experiments, we used a threshold value of 55% of the trials. The program run took about 20 h in a Pentium 4 with 2.4 GHz and 1 GB RAM.

From Figure 6, we can observe that in both networks (GN-1 and GN-2), STE12 has an important position, because it has the highest number of children. Thus, the recovered networks suggest that this gene has an important role in transcriptional regulation. Indeed, as indicated by the biological literature, STE12 is the most important transcription factor that is activated by the pheromone response pathway (Zeitlinger et al., 2003), and it influences the transcription of a number of other genes.

For GN-1, the network successfully predicts known regulatory interactions between STE12 and genes STE2, FUS3 and FAR1, which are positively acting components of the pathway (Bardwell, 2005). It did not capture known interactions with AGA1, a gene involved in cell fusion (White and Rose, 2001), nor were interactions found with GPA1, a negative regulator of the pathway (Bardwell, 2005). However, GN-1 reports three interesting relations of STE12 with STE20, STE5 and STE11. STE5 is an adaptor protein that binds to and activates STE11, the upstream kinase on the MAPK cascade. STE20 is another protein in the neighborhood that is another activator for STE11. GN-1 indicates some kind of regulation (activation or inhibition) of these genes by STE12, which we did not find evidenced in a specialized database (SCPD, 2005). We hypothesize that STE12 is in someway upregulating these genes, because MAPK targets will act towards maintaining the transcription of STE12 target genes (e.g., inhibition of Dig1 and Dig2, which are inhibitors of STE12, thus benefiting the maintainability of mating genes transcription). The diagram in Figure 7 provides an illustration of this scenario. The links around STE12, inferred by GN-1, are highlighted (red) and the links in black depict interactions at protein level. This diagram also allows one to note that GNs encode protein-DNA interactions besides only protein-protein relationships, as the case of the signaling pathway in Figure 5.

GN-1 also indicates that FAR1 expression is controlled by STE12 and GPA1. One of the functions of FAR1 is to promote G1 cell-cycle arrest. Given the fact that GPA1 is a negative

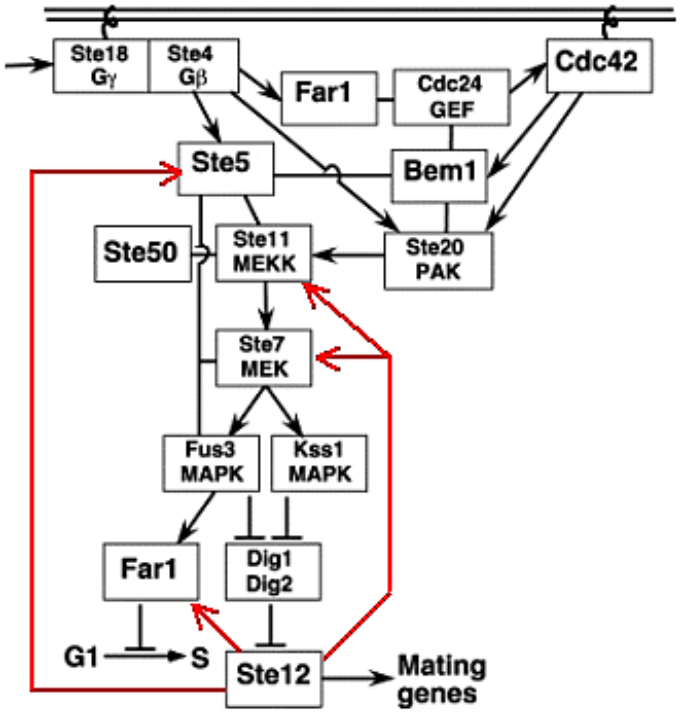


Figure 7. Wiring diagram of the yeast pheromone response pathway (adapted from Bardwell, 2005), along with some regulatory links for STE12, found by gene network-1. A black line indicates a regulatory relationship (activation or inhibition) at the protein level (binding, phosphorylation, and so on). Red-labeled lines identify some of the genes regulated by STE12 (protein-DNA interactions).

regulator of the pheromone response (see above), this link (GPA1, FAR1) is a feasible hypothesis, and it remains to be determined by which mechanism GPA1 down-regulates FAR1.

STE18 is the gene with the second-most regulatory connections in GN-1 (Figure 6). STE18 expresses one of the subunits of the G-protein complex, bound to the pheromone receptor. GN-1 shows that mRNA expression of STE18 affects the regulation of STE7 and STE11 (both protein kinases), as well as STE12, the main transcription factor.

GN-2, which contains STE3 instead of STE2, has also revealed STE12 as the most important gene in the transcriptional regulation of the pathway. GN-2 has found the same interactions around STE12 in the first experiment, with STE3 (in place of STE2), FUS3, FAR1, STE20, STE5 (inverted), and STE11, previous justified. It has also found two interactions evidenced in the literature (Bardwell, 2005) not found by GN-1: a direct interaction (STE12, GPA1) and a second-level interaction (STE12, AGA1).

The second dataset (Hartemink et al., 2002) had 12 genes that are directly or closely involved with the mating pathway, and it was used to infer a third network GN-3. The expression data consisted of 20 observations of the two types of haploid yeast cells (α and a), under several conditions, including exposure to different nutritive media, and exposure to various types of stress, such as heat, oxidative species, excessive acidity, and excessive alkalinity.

GN-3 is presented in Figure 8. STE12 again is the most connected gene, but here some important regulatory connections, such as (STE12, FUS1) and (STE12, STE2) were not discov-

ered, even though several others were identified, including (STE12, FAR1) and (STE12, GPA1). This example shows that even for the same subset of genes, the design of the experiment is a key issue for gene network inference. The first dataset was generated in a time series fashion, whereas in the second one, there were changes in the conditions, but not in time. To properly analyze the dynamics of expression, it is better to use time series microarray data (Bar-Joseph, 2004).

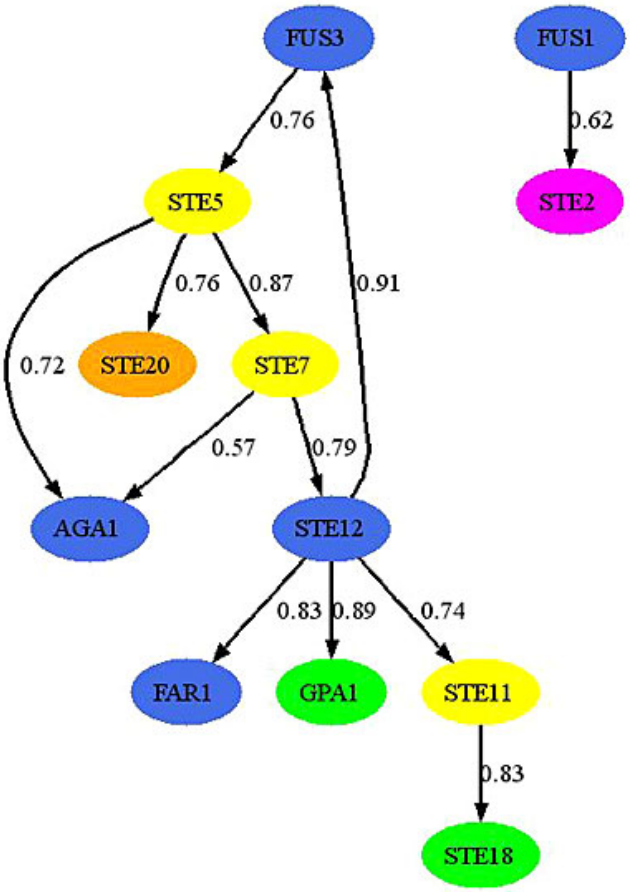


Figure 8. Gene network 3 (GN-3) recovered from the second dataset of observational expression data across a variety of conditions. FUS1 and STE2, both regulated by STE12, appear linked and isolated in the network.

A last important comment can be made about the interaction (STE12, FAR1), which appears in all three networks. According to information from the SCPD yeast promoter database (SCPD, 2005), FAR1 has only one transcription factor, called MCM1, which is found by a motif tool. Nonetheless, our networks hint that STE12 also regulates FAR1 by some mechanism, which could be through the binding of STE12 to an upstream region of that gene, as depicted in Figure 9. Indeed, a chromatin immuno-precipitation assay revealed that STE12 also binds to the upstream region of FAR1 (Ren et al., 2000), although it does not exhibit a motif for STE12. This fact best exemplifies what kind of knowledge about transcriptional control one could extract from GNs, and in that manner rationally design wet experiments.

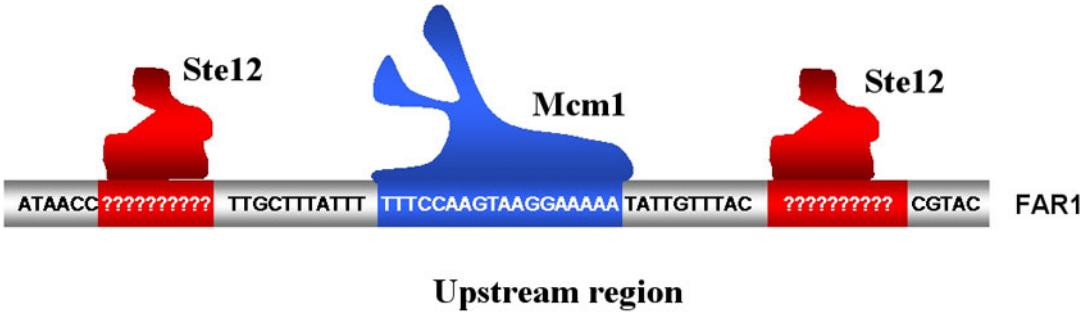


Figure 9. Gene networks inferred from all datasets indicate that STE12 is a transcriptional regulator of FAR1 that is not found by motif tools.

CONCLUDING REMARKS

We have introduced a BN model for GNs, and we have tested it with both artificial and real biological networks. We analyzed the yeast pheromone response pathway, and we demonstrated the usefulness of GNs as a computational approach for the analysis of transcriptional regulation. In summary, a GN can be used, among other things, to i) define transcriptional factors (activators and inhibitors) for a target gene and ii) find co-regulated genes. The intention of the efforts for developing both theories and software for network analysis is that these networks could provide useful clues about biological systems, thus helping with the design and refinement of wet experiments.

The BN model is suitable for small networks. A learning scheme that scales-up to a large number of variables should be investigated, and is a future goal. Nowadays, the finding of an efficient reconstruction method with no constraints in the number of nodes using BN is a cutting-edge problem (Bar-Joseph, 2004).

We are aware of the limitations of gene networks as a way to understand the behavior of a biological system, in terms of phenotype. These occur because 1) there is a low correlation between expression level and protein level (Ideker et al., 2001), and 2) much of the cellular regulation occurs post-translationally, and genomic-scale technologies to measure protein levels are still at beginning stages of development (Rice and Stolovitzky, 2004). For this reason, and to attenuate this weakness, a trend in this field of network analysis is information fusion. According to this concept, the algorithms to learn the networks should incorporate other sources of biological data, such as location data, sequence data and protein-protein interactions. Some researchers have already proposed models in this direction (e.g., Yamanashi et al., 2004; Bernard and Hartemink, 2005) and this will become an active research topic in the next years.

ACKNOWLEDGMENTS

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support. The authors also thank A.J. Hartemink for providing the yeast microarray data.

REFERENCES

Bar-Joseph Z (2004). Analyzing time series gene expression data. *Bioinformatics* 20: 2493-2503.

Bardwell L (2005). A walk-through of the yeast mating pheromone response pathway. *Peptides* 26: 339-350.

Bastos G and Guimarães KS (2005). A simpler Bayesian network model for genetic regulatory network inference. Proceeding of International Joint on Neural Networks, Montreal, Canada.

Bernard A and Hartemink AJ (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. *Pac. Symp. Biocomput.*, Hawaii, USA, pp. 459-470.

Chickering D (1996). Learning Bayesian networks is NP-complete. In: Learning from data: AI and statistics V (Fisher D and Lenz HJ, eds.). Springer-Verlag, New York, NY, USA, pp. 121-130.

de Jong H (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol*. 9: 67-103.

de la Fuente A, Bing N, Hoeschele I and Mendes P (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20: 3565-3574.

Eilers PHC and Marx BD (1996). Flexible smoothing with B-splines and penalties. *Stat. Sci.* 11: 89-121.

Friedman N (2004). Inferring cellular networks using probabilistic graphical models. *Science* 303: 799-805.

Friedman N, Linial M, Nachman I and Pe'er D (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol*. 7: 601-620.

Hartemink AJ, Gifford DK, Jaakkola TS and Young RA (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput*. pp. 437-449.

Ideker T, Thorsson V, Ranish JA, Christmas R, et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.

Imoto S and Konishi S (2000). B-spline nonparametric regression models and information criteria. Proceedings of 2nd Int. Symp. on Frontiers of Time Series Model, pp. 240-241.

Imoto S, Goto T and Miyano S (2002). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.*, pp. 175-186.

Lewin B (2000). Genes VII. Oxford University Press, New York, NY, USA.

Nachman I, Regev A and Friedman N (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* 20 (Suppl 1): I-248-I-256.

Ott S, Imoto S and Miyano S (2004). Finding optimal models for small gene networks. *Pac. Symp. Biocomput.*, Hawaii, USA, pp. 557-567.

Ren B, Robert F, Wyrick JJ, Aparicio O, et al. (2000). Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.

Rice JJ and Stolovitzky G (2004). Making the most of it: pathway reconstruction and integrative simulation using the data at hand. *Biosilico* 2: 70-77.

Roweis S and Ghahramani Z (1999). A unifying review of linear Gaussian models. *Neural Comput*. 11: 305-345.

Schafer J and Strimmer K (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21: 754-764.

SCPD (2005). The promoter database of *Saccharomyces cerevisiae*. Available at: http://rulai.cshl.edu/SCPD/. Accessed August, 2005.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273-3297.

STKE (2005). Science Signal Transduction Knowledge Environment (STKE). Available at: http://stke.sciencemag.org/cgi/content/full/sci;306/5701/1508/FIG1. Accessed August, 2005.

Tamada Y, Kim S, Bannai H, Imoto S, et al. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics* 19 (Suppl 2): II-227-II-236.

van Someren EP, Wessels LF, Backer E and Reinders MJ (2002). Genetic network modeling. *Pharmacogenomics* 3: 507-525.

White JM and Rose MD (2001). Yeast mating: getting close to membrane merger. *Curr. Biol*. 11: R16-R20.

Xia Y, Yu H, Jansen R, Seringhaus M, et al. (2004). Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem*. 73: 1051-1087.

Yamanishi Y, Vert JP and Kanehisa M (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 20 (Suppl 1): I-363-I-370.

Zeitlinger J, Simon I, Harbison CT, Hannett NM, et al. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* 113: 395-404.