Physica A 390 (2011) 2198–2207

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Coarse-grain reconstruction of genetic networks from expression levels

L. Diambra

Laboratorio de Biología de Sistemas – CREG-UNLP, Av. Calchaqui Km 23.5 CP 1888, Florencio Varela, Argentina

ARTICLE INFO

Article history:
Received 3 August 2010
Received in revised form 19 January 2011
Available online 18 February 2011

Keywords:
Gene network
Reverse engineering
Information theory

ABSTRACT

In the postgenome era many efforts have been dedicated to systematically elucidate the complex web of interacting genes and proteins. These efforts include experimental and computational methods. Microarray technology offers an opportunity for monitoring gene expression level at the genome scale. By recourse to information theory, this study proposes a mathematical approach to reconstruct gene regulatory networks at a coarse-grain level from high throughput gene expression data. The method provides the *a posteriori* probability that a given gene regulates positively, negatively or does not regulate each one of the network genes. This approach also allows the introduction of prior knowledge and the quantification of the information gain from experimental data used in the inference procedure. This information gain can be used to choose those genes that will be perturbed in subsequent experiments in order to refine our knowledge about the architecture of an underlying gene regulatory network. The performance of the proposed approach has been studied by *in numero* experiments. Our results suggest that the approach is suitable for focusing on size-limited problems, such as recovering a small subnetwork of interest by performing perturbation over selected genes.

1. Introduction

Gene expression is regulated by proteins that enhance or block polymerase binding at the promoter region. These biochemical reactions constitute the edges of the gene regulatory networks. One of the key issues in modern biology is the elucidation of the structure and function of gene regulatory circuits at the system level [1]. To address this challenge many efforts have been devoted to the task of developing computational methods capable of inferring the interaction between genes from expression levels both on small pathways [2,3] and on the genome-wide scale (see [4] for a review). Several models for gene regulatory networks have been proposed in order to infer network interactions [5–8], such as Bayesian networks [9–11], Boolean networks [12] and linear models [13,14,16]. Once a regulatory network model has been chosen, it is possible, in principle, to recover its parameters with some accuracy. Of course, more detailed models will require more extensive experimental data. In general, these data are not available for the genome-wide scale assuming complex model. However, we can concentrate on a simpler task, such as: Who is regulating whom? and, Is that an up-regulation or a down-regulation? The idea behind restricting our questions to this qualitative information level is to reduce the amount of data needed to infer valuable and robust biological knowledge even when dealing with noisy data. In any case, the detailed information offered by more detailed modeling is not useful without a careful significance analysis of these predictions. In this sense, this study proposes a mathematical approach to infer gene networks at the coarse-grain level. The inference process is to be accomplished according to the information theory (IT) in the framework of the maximum entropy principle [17,18]. IT has proved to be suitable for devising techniques for analyzing gene expression and network reconstruction [19,20,7], where gene expression levels were regarded as random variables. Here, complementing these previous studies, each putative interaction has been considered as a random variable. *In numero* experiments show that, in

E-mail addresses: ldiambra@creg.org.ar, ldiambra@gmail.com.

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.02.021

this case, the IT parlance also provides a powerful framework to discuss questions related to the modeling process such as: (i) how to incorporate *a priori* information about gene interaction; (ii) how to assess the likelihood of the inferred paths; (iii) how to quantify the information provided by the experimental data; and (iv) how to design experiments in order to identify subnetworks.

2. Methods

2.1. Network reverse engineering

In general, a genetic network can be modeled by a set of nonlinear differential equations $\dot{x}_i = f_i\left(x_1(t), \ldots, x_N(t)\right)$, where $x_i(t)$ is the expression level of gene i at time t, and f_i is the regulatory function governing the expression of gene i [21]. Near a steady state, the nonlinear system can be approximated by a set of linear differential equations, $\dot{\mathbf{x}} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is a weighted connectivity matrix [22]. In order to uncover the connectivity matrix, an external perturbation can be applied to the level of transcripts $\mathbf{b} = (b_1(t), \ldots, b_N(t))^T$ By repeating the procedure M times, a measurement matrix $\mathbf{X}$ is obtained, where columns denote the experiments and rows indicate individual genes. Thus, the dynamics can be approximated by

$$\dot{\mathbf{X}} = \mathbf{W}\,\mathbf{X} + \mathbf{B} \tag{1}$$

where $\dot{\mathbf{X}}$ and $\mathbf{B}$ follow the same notation as $\mathbf{X}$.

Usually, inferring a genetic network attempts to retrieve the weight matrix $\mathbf{W}$ using time-series RNA expression data or steady state data. Algorithms that use time-series data have two alternatives: (i) estimate the rates of change of the transcript level ($\dot{\mathbf{X}}$) from the time series; (ii) convert the model to a discrete dynamic system [14,15]. On the other hand, there are algorithms that use steady-state data [3,6], in this case

$$-\mathbf{B} = \mathbf{W}\,\mathbf{X}. \tag{2}$$

In principle, the inference method presented here can deal with time-series RNA expression data or steady-state data, because both equations are equivalent. Here, the more general case will be derived, i.e. Eq. (1), but the numerical simulations presented in the Results section only consider the steady-state case for simplicity. Experimental details about how to obtain steady-state expression-level data are in [3,6].

2.2. Entropy maximization

In the present work, the maximum entropy principle is applied to obtain the probability distribution $P\left(\mathbf{W}|D_M\right)$ from the data $D_M = \left\{\mathbf{X}, \dot{\mathbf{X}}, \mathbf{B}\right\}$. After that, using a maximum *a posteriori* criterion, the gene interaction matrix $\mathbf{I}$ is selected. The elements I_{ij} can take only three values, depending on the type of influence of gene j on gene i, $I_{ij} = 1$ for activation (direct or indirect), $I_{ij} = -1$ for repression and $I_{ij} = 0$ when gene j does not have any influence on gene i. In order to infer weights consistent with D_M, it is assumed that each set of weights $\mathbf{W}$ is realized with probability $P\left(\mathbf{W}|D_M\right)$. In other words, a normalized probability distribution is introduced over the possible sets $\mathbf{W}$, which satisfy

$$\langle \mathbf{W} \rangle = \int P\left(\mathbf{W}|D_M\right) \mathbf{W} \mathrm{d}\mathbf{W}. \tag{3}$$

The relative entropy related to an *a priori* probability distribution P_0, is given by

$$H_r\left(D_M|P_0\right) = -\int P\left(\mathbf{W}|D_M\right) \ln\left[\frac{P\left(\mathbf{W}|D_M\right)}{P_0(\mathbf{W})}\right] \mathrm{d}\mathbf{W}, \tag{4}$$

where $P_0\left(\mathbf{W}\right)$ is an appropriate *a priori* distribution. The negative relative entropy H_r, known as Kullback–Leibler divergence [23], defines the information gained after D_M has been used in the inference procedure. Thus, in this framework, the inference process takes place through a modification of the probability distribution on weight space due to incoming data.

Thus, following the central tenets of the maximum entropy principle, relative entropy is maximized subject to the constraints Eq. (3). Thus, the *a posteriori* probability distribution yields

$$P\left(\mathbf{W}|D_M\right) = \exp\left(-\left(1+\lambda_0\right)\right) \exp\left(-\mathbf{W}\cdot\boldsymbol{\Gamma}\right) P_0\left(\mathbf{W}\right), \tag{5}$$

where λ_0 is the Lagrange multiplier associated with the normalization condition, and $\boldsymbol{\Gamma}$ the Lagrange multipliers associated with the constraints of Eq. (3), which are determined once P_0 is properly selected.

In order to select P_0, it is assumed that the weights are restricted to the values of I_{ij}, i.e. $w_{ij} = 0, \pm 1$, and then a three-peaked *a priori* distribution is used, which is described by

$$P_0\left(\mathbf{W}\right) = (2\pi a)^{-N/2} \prod_{ij}^{N} \left[p_{ij}^0 \mathrm{e}^{-\frac{w_{ij}^2}{2a}} + p_{ij}^+ \mathrm{e}^{-\frac{(w_{ij}-1)^2}{2a}} + p_{ij}^- \mathrm{e}^{-\frac{(w_{ij}+1)^2}{2a}} \right], \tag{6}$$

where $p^{x}_{i,j}$ is the *a priori* probability for gene j to regulate positively ($x = +$), negatively ($x = -$) or to not regulate ($x = 0$) gene i. Of course, $p^{0}_{ij} + p^{+}_{ij} + p^{-}_{ij} = 1$ for each pair i, j. The parameter a can be regarded as a constraint smoothness parameter. By replacing this choice in Eq. (5), the *a posteriori* probability distribution is obtained as a sum of three Gaussians,

$$P\left(\mathbf{W}|D_M\right) = \frac{1}{(2\pi a)^{N/2}} \prod_{ij}^{N} \left[\hat{p}^{0}_{ij} \mathrm{e}^{-\frac{\left(w_{ij}+a\Gamma_{ij}\right)^2}{2a}} + \hat{p}^{+}_{ij} \mathrm{e}^{-\frac{\left(w_{ij}+a\Gamma_{ij}-1\right)^2}{2a}} + \hat{p}^{-}_{ij}\, \mathrm{e}^{-\frac{\left(w_{ij}+a\Gamma_{ij}+1\right)^2}{2a}} \right] \tag{7}$$

where $\hat{p}^{x}_{ij}$ is the *a posteriori* probability for gene j to regulate positively ($x = +$), negatively ($x = -$) or to not regulate ($x = 0$) gene i. These probabilities are defined by $\hat{p}^{+}_{ij} = p^{+}_{ij}\mathrm{e}^{-\Gamma_{ij}}/z_{ij}$, $\hat{p}^{-}_{ij} = p^{-}_{ij}\mathrm{e}^{\Gamma_{ij}}/z_{ij}$ and $\hat{p}^{0}_{ij} = p^{0}_{ij}/z_{ij}$, where $z_{ij} = 1 + p^{+}_{ij}\left(\mathrm{e}^{-\Gamma_{ij}} - 1\right) + p^{-}_{ij}\left(\mathrm{e}^{\Gamma_{ij}} - 1\right)$ guarantees normalization. Furthermore, the relative entropy of the *a posteriori* distribution Eq. (4) is given by

$$H_r\left(D_M, P_0\right) = -\sum_{i}^{N} I_g\left(i|D_M, P_0\right), \tag{8}$$

where $I_g(i)$ is the information gain of gene i with respect to P_0 obtained from using the data D_M that is defined by

$$I_g\left(i|D_M, P_0\right) = \sum_{j}^{N} \left[\frac{a}{2}\Gamma^2_{ij} - \ln\left(z_{ij}\right) - \frac{1}{z_{ij}}\left(p^{+}_{ij}\Gamma_{ij}\mathrm{e}^{-\Gamma_{ij}} - p^{-}_{ij}\Gamma_{ij}\mathrm{e}^{\Gamma_{ij}}\right) \right]. \tag{9}$$

The multipliers Γ_{ij} are obtained after solving the equation

$$\langle w_{ij} \rangle = -a\Gamma_{ij} + z^{-1}_{ij}\left(p^{+}_{ij}\mathrm{e}^{-\Gamma_{ij}} - p^{-}_{ij}\mathrm{e}^{\Gamma_{ij}}\right) \tag{10}$$

where $\langle w_{ij} \rangle$ are subject to the constraints imposed by D_M.

2.3. Network inference and IT

Our *central* idea is that of reinterpreting, following information in D_M in a particular fashion,

$$\dot{\mathbf{X}} - \mathbf{B} = \langle \mathbf{W} \rangle\, \mathbf{X}. \tag{11}$$

Thus, all of the possible networks that are consistent with Eq. (11) can be written as

$$\langle \mathbf{W} \rangle = \left(\dot{\mathbf{X}} - \mathbf{B}\right) \cdot \mathbf{U} \cdot \mathrm{diag}(s^{-1}_j) \cdot \mathbf{V}^T + \mathbf{C} \cdot \mathbf{V}^T \tag{12}$$

$\mathbf{C} = (c_{ij})$ is an $N \times N$ matrix, where c_{ij} is zero if $s_j \neq 0$ and is otherwise an arbitrary scalar coefficient. $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ correspond to the singular value decomposition of matrix $\mathbf{X}^T$, i.e. $\mathbf{X}^T = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$, where $\mathbf{U}$ is a unitary $M \times N$ matrix of left eigenvectors, $\mathbf{S}$ is a diagonal $N \times N$ matrix containing the eigenvalues $\{s_1, \ldots, s_N\}$, and $\mathbf{V}$ is a unitary $N \times N$ matrix of right eigenvectors. Without loss of generality, let all nonzero elements of s_j be listed at the end, and s^{-1}_j in Eq. (12) are taken to be zero if $s_j = 0$. The general solution (12) can be written as

$$\langle \mathbf{W} \rangle = \mathbf{W}_{L_2} + \mathbf{C} \cdot \mathbf{V}^T, \tag{13}$$

where $\mathbf{W}_{L_2}$ is the particular solution with the smallest L_2 norm. If $M < N$, many weights $\mathbf{W}$ are compatible with the available information. The information contained in the data set D_M can be used in different ways. Each of these leads to a different probability distribution that exhibits diverse properties. In this sense, following the prescription $\langle \mathbf{W} \rangle = 0$ in Eq. (13), the knowledge that gene regulatory networks are sparse can be made use of. Thus, we have $\mathbf{C} \cdot \mathbf{V}^T = -\mathbf{W}_{L_2}$, which is an overdetermined problem [22]. This equation is solved by the interior point method for L1 regression. The $c_{i,j}$ values thus obtained are replaced in Eq. (13) and have a particular solution. This solution will be denoted as $\mathbf{W}_{L_1}$. Of course $\boldsymbol{\Gamma}$ is obtained by solving Eq. (10) using $\langle \mathbf{W} \rangle = \mathbf{W}_{L_2}$ or $\langle \mathbf{W} \rangle = \mathbf{W}_{L_1}$. In the following sections these alternatives will be considered independently. Notice that for $M \geq N$, $\mathbf{W}_{L_2} = \mathbf{W}_{L_1}$.

After determining the *a posteriori* distribution, the gene interaction matrix I must be selected. In order to do that, the maximum *a posteriori* criterion is taken into account, i.e. the selection is accomplished by choosing the highest *a posteriori* probability from $\{\hat{p}^{0}_{ij}, \hat{p}^{+}_{ij}, \hat{p}^{-}_{ij}\}$ for each pair i, j. For example, if $\hat{p}^{+}_{ij}$ is greater than $\hat{p}^{0}_{ij}$ and $\hat{p}^{-}_{ij}$, then $I_{ij} = 1$, indicating that gene j activates gene i.

In order to achieve the best model, the idea is to use the information contained in D_M and the knowledge that gene regulatory networks are sparse. The formalism presented here offers an alternative to the prescription that selects $\mathbf{W}_{L_1}$ from all possible solutions (12). This alternative consists in setting $p^{+}_{ij} = p^{-}_{ij} \ll p^{0}_{ij}$. In this way, the knowledge that the gene regulatory network is sparse can be introduced by assigning a much lower value to the *a priori* probabilities of interaction than that of *a priori* probabilities of absence of interaction. Furthermore, as the inference processes occur row by row, any other relevant *a priori* information about the gene in consideration (such as known interactions, type of gene, etc.) could be included in these probabilities. For example, if gene k encodes a helix-turn-helix or a zinc finger protein, high probabilities can be assigned for column k (p^{+}_{ik} and p^{-}_{ik}).

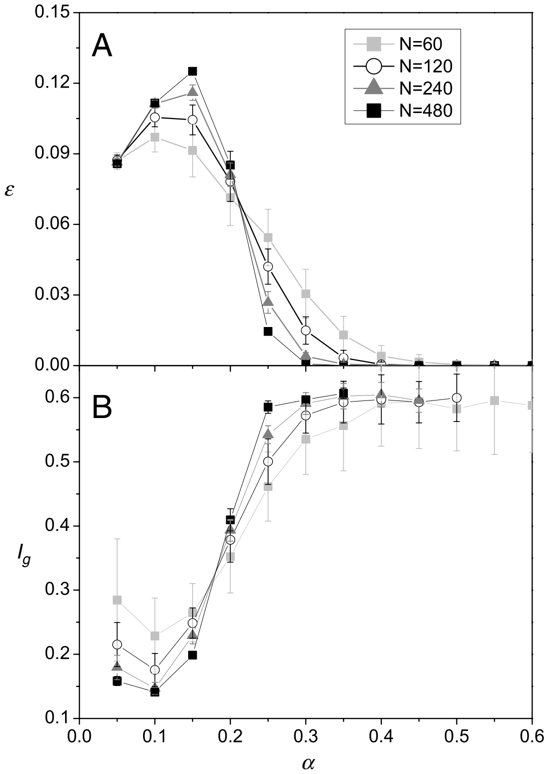


Fig. 1. Performance. (A) Prediction error ε as a function of the ratio $\alpha = M/N$ for gene networks with 60 genes (squares), 120 genes (circles) and 240 genes (triangles), averaged over 50 networks. (B) The information gain I_g curves associated with each performance. In all cases the performances were obtained using W_{L_1} prescription, equal a priori probabilities (i.e. $p_{ij}^{+} = p_{ij}^{-} = p_{ij}^{0} = 1/3$ for all i and j), $k/N = 0.05$ and $a = 0.01$.

3. Results

In order to systematically benchmark the inference performance of this method, the linear data-generating model of Eq. (2) was used. The M random state (the columns of matrix $\mathbf{X}$) were generated in the range $[-1, 1]$ and the perturbation $\mathbf{B}$ was computed as $-\mathbf{W} \cdot \mathbf{X}$, where $\mathbf{W}$ is the matrix to be reconstructed in a coarse-grained sense. Thus the pair $\mathbf{B}, \mathbf{X}$ constitutes the available information D_M. First, we use two different network architectures: random network (RN) and scale-free network (SFN). To build the connectivity matrix $\mathbf{W}$ of sparse RN the following procedure was used: for each matrix element a random number r between [0, 1] was sorted; if $r < k/2N$, a negative random value chosen from a uniform distribution in the range $[-2.0, -0.1]$ was assigned to the matrix element; if $r > 1-k/2N$, the matrix element was a positive random number in the range $[0.1, 2.0]$, and otherwise the matrix element was zero. In the case of SFN the connections follow a preferential attachment scheme, as described in [24], where at every step a new gene with k regulatory entries is added. The probability to choose regulatory genes to act on this new gene depends linearly on the number of regulatory interactions of those genes. After generating the adjacency matrix, each nonzero element of this matrix is replaced by a random number uniformly distributed. Both procedures result in networks where the total number of connections is kN. The condition $k \ll N$ ensures sparseness.

After defining which prescription was used for the mean values $\langle w_{ij} \rangle$, the set of uncoupled nonlinear equation (10) was solved and the *a posteriori* probability for each putative interaction was evaluated. A finite value of parameter a is necessary to find a numerical solution of Eq. (11) when the amount of data is small ($\alpha < 0.2$) and/or noisy. In this paper, $a = 0.01$ has been used, because this value was sufficiently large to guarantee the solution of Eq. (11) in the numerical experiments. No significant difference was detected for $a = 0.05$; however, higher values lead to a worse performance. After this procedure the most likelihood $\mathbf{I}$ can be selected. The performance of the inference procedure, was measured by the prediction error $\varepsilon = N^{-2} \sum_{ij}^{N} e_{ij}$, where e_{ij} is defined by

$$e_{ij} = \begin{cases} 0 & \text{if sign } (w_{ij}) = I_{ij} \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

Fig. 1(A) depicts the prediction error ε as a function of α defined as the ratio of number of experiments and number of genes, i.e. $\alpha = M/N$. These have been tested in three different size networks with $k/N = 0.05$, in which all a priori probabilities are assumed to be equal (i.e. $p_{ij}^{+} = p_{ij}^{-} = p_{ij}^{0} = 1/3$) and $a = 0.01$. For small values of M, the method mistakenly infers a percentage of interactions that depends on the network's size N and on k. The ε function does not have a monotonic decreasing behavior at low α, which is more apparent at higher N. However, a solution obtained with lower α always has

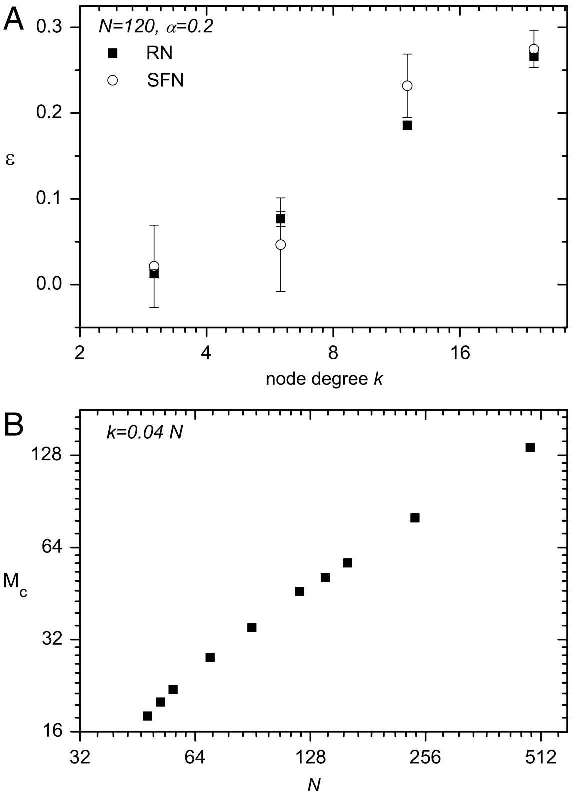


Fig. 2. Size and node degree. (A) Prediction error ε as a function of the k for gene networks with $N = 120$ genes and $M = 24$ measurements averaged over 50 networks. Filled squares correspond to RN and open circles correspond SFN. (B) The critical number of measurements needed to reconstruct the matrix I without errors as a function of the network size N with $k = 0.04N$. In all cases the performances were obtained using W_{L_1} prescription, equal a priori probabilities (i.e. $p_{ij}^{+} = p_{ij}^{-} = p_{ij}^{0} = 1/3$ for all i and j), and $a = 0.01$.

a smaller number of predicted interactions, at a given significance level, than solutions obtained with higher α, even when comparing solutions with a similar prediction error. Fig. 1(B) depicts the associated information gain $I_g = \sum_i I_g(i|D_M)$ as a function of α. These curves almost follow the prediction error in a complementary way, i.e. I_g increases when the error decreases, and saturates when the prediction error reaches zero, but it increases monotonically with α for large networks. This fact suggests that I_g can be used to illustrate the performance of a given data set D_M in the network inference process, since its computation (see Eq. (9) does not require one to know the interaction matrix **W**, in opposition to the ε computation. The prediction error decays rapidly as α increases and the gene interaction matrix is completely recovered with a α value that decreases with the network's size. This performance was obtained using the $\mathbf{W}_{L_1}$ prescription. Similar simulations (data not shown) performed with the $\mathbf{W}_{L_2}$ prescription reveal that, in these cases, the prediction error ε remains close to unit until $\alpha = 1$, where it decays abruptly. In the simulations it was observed that the mean performance depends on the network size and the degree of connectivity k. Fig. 2(A) shows the dependence of the prediction error ε over the node degree k of two kinds of networks: RN (filled squares) and SFN (open circles). Since no dependence on the network type was observed, for the following steps, the RN type was used, which has a smaller variability than the SCN. In Fig. 2(B), we depict the critical number of measurement M_c required to recover the matrix I without error as a function of the network's size N.

Many times, when dealing with an incomplete data set $M \ll N$, only a percentage of the interactions is inferred correctly. If the likelihood of the inferred paths cannot be assessed, this partial reconstruction has a small predictive value in real life. The methodology proposed here can assess the likelihood of the predicted interaction straightforwardly through the a posteriori probability. In this sense, only those predicted interactions with an a posteriori probability that is greater than some significance level can be selected. To illustrate this point, a network with 60 genes with $k/N = 0.05$ was simulated. The related connectivity matrix **W** is represented in Fig. 3(left), row i corresponds to the genes that regulate the activity of gene i, while column j corresponds to the genes regulated by gene j. The weight values w_{ij} are depicted following a linear gray scale, where white(black) corresponds to the maximum(minimum) values of weights, and the gray background represents the absence of interaction. This network is randomly perturbed with $M = 20$ different experiments ($\alpha = 1/3$). With this amount of data, ε is usually around $\sim$0.03 (see Fig. 1). Nevertheless, in a real world problem one does not known which interactions were inferred correctly and which were inferred incorrectly. By means of the information theory approach, the a posteriori probabilities were computed and the inferred interaction matrix **I** and the associated likelihood were derived. Fig. 3(right) represents the inferred connectivity matrix **I**, by assuming that all a priori probabilities are equal (i.e. $p_{ij}^{+} = p_{ij}^{-} = p_{ij}^{0} = 1/3$). Circles indicate the interactions with an *a posteriori* probability greater than 0.99. In this case there are 79 interactions whose associated *a posteriori* probabilities are greater than 0.99, and all these predictions were correct. Fig. 4(A) depicts the number

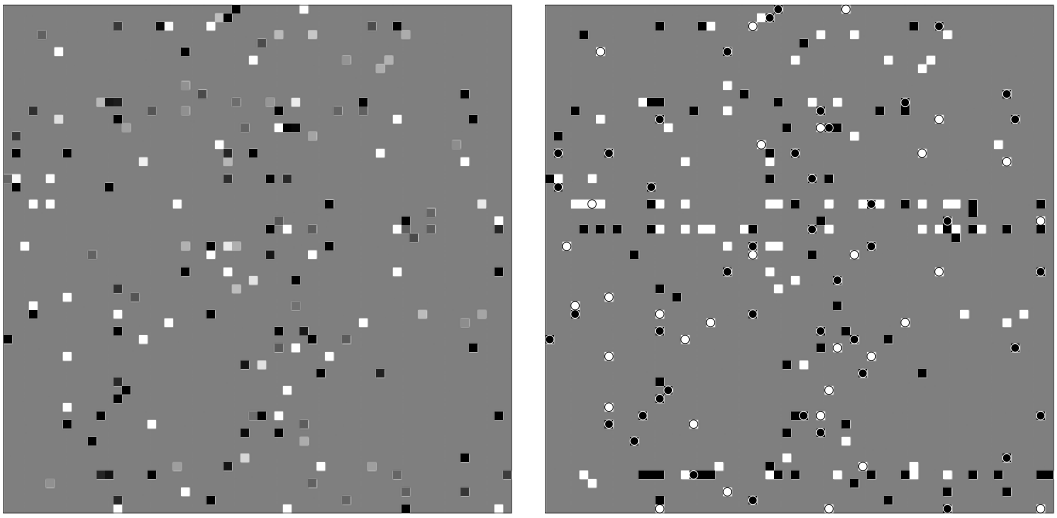


Fig. 3. Likelihood assessment. Left: connectivity matrix **W** representation related to a random network of 60 genes with $k/N = 0.05$. Rows correspond to regulated genes, while columns correspond to the genes acting as regulators. The interaction weights w_{ij} are represented following a linear gray scale, where white corresponds to $w_{ij} = 2$, while black to $w_{ij} = -2$. The gray background represents the absence of interaction, i.e. $w_{ij} = 0$. Right: gene interaction matrix **I** inferred after 20 random perturbation experiments, using W_{L_1} prescription, $a = 0.01$ and $p^+_{ij} = p^-_{ij} = p^0_{ij} = 1/3$. Circles indicate the 76 interactions with an *a posteriori* probability greater than 0.99. Wrong predictions (51, ∼1.5% of the putative interactions), which in this case correspond to the regulatory inputs of three genes.

of interaction calls predicted correctly, with an *a posteriori* probability greater than a given threshold, relative to the total number of interaction calls in the regulatory network as a function of the threshold for two different values of M (squares $M = 20$ and circles $M = 10$). These results suggest that gene networks can be partially recovered even with small amounts of data, mainly for those genes that interact strongly. The a posteriori probabilities associated with the noninteraction were slightly greater than 1/3 in most of the cases. In Fig. 4(B) we can see the number of noninteraction calls predicted wrongly, with an *a posteriori* probability greater than a given threshold, relative to the total number of noninteraction calls in the regulatory network as a function of the threshold for two different values of M. We note that the performance of predictions is different depending on whether it is an effective interaction or not.

Unfortunately, all measurements are subject to observational noise. Consequently, it is important to assess to what extent the performance of the inference procedure is affected by noise. To simulate this condition in the numerical experiment, the available information D_M (both input and output) was corrupted by an additive Gaussian noise with mean zero and standard deviation η. This inference procedure was performed for networks with $N = 60$, under the same condition as for the previous assessment ($p^+_{ij} = p^-_{ij} = p^0_{ij} = 1/3$ and $a = 0.01$). However, in this case the method based on the prescription of sparseness assumed in W_{L_1} could not correctly recover the gene interaction matrix **I** when the noise level was $\eta = 0.3$ (even for smaller η). Left panels of Fig. 5 indicate the prediction error by using both $\mathbf{W}_{L_1}$ (black squares) and $\mathbf{W}_{L_2}$ (gray circles) assuming that the *a priori* probabilities for activation, repression or absence of interaction are equal. We have considered the prediction error relative to interaction calls and to noninteraction calls separately in the top and bottom panels of Fig. 5, respectively. For the nonsparse P_0 (left panels) the prediction error of the interaction calls decreases as more data become available, clearly the $\mathbf{W}_{L_2}$ prescription has better performance than $\mathbf{W}_{L_1}$ in this case. On the other hand, the prediction power relative to the noninteraction calls becomes worse when more data are added in the case of $\mathbf{W}_{L_1}$ prescription and is null for the $\mathbf{W}_{L_2}$ prescription. However, the network can be partially reconstructed by using an alternative constraint of sparseness. This alternative consists in introducing the knowledge of sparseness of the interaction matrix through the a priori probabilities. This is achieved by setting $p^{\pm}_{ij} \approx 0$ in the inference procedure. Right panels of Fig. 5 depict the prediction error as a function of α when the *a priori* probabilities were set to $p^{\pm}_{ij} = 0.025$. The sum of these probability values corresponds to the percentage of genes that are regulated by one gene. With such *a priori* information, it is possible to notably improve the performance prediction of a false interaction at the expense of some performance in the prediction of the interactions. The mean node degree of the network is generally not known in advance. However, the prediction ability is robust for underestimations of the *a priori* probabilities. Simulations using sparser *a priori* probabilities, $p^{\pm}_{ij} = 0.01$ for example, give almost the same results (data not shown) as the last one. This implies that it is possible to partially recover the interaction matrix even with noisy data, by setting low values for the *a priori* probabilities $p^{\pm}_{ij}$. In the right panels of Fig. 5, the overall prediction performance obtained by the $\mathbf{W}_{L_2}$ prescription is comparable with that obtained by $\mathbf{W}_{L_1}$ prescription using $p^{\pm}_{ij} \approx 0$, in contrast to the case that deals with clean data. This is relevant because at low α values, the computational cost of $\mathbf{W}_{L_1}$ solution is very high. Furthermore, when data are corrupted by noise, it was observed that the prediction error has a peak around $\alpha = 1$. This peak arises because, as consequence of additive noise, the pair $\boldsymbol{X} - \boldsymbol{B}$ does not satisfy Eq. (2). Some values of the diagonal matrix **S** become very small or zero as a consequence of the inconsistency in the equation system. Similar effects were reported in

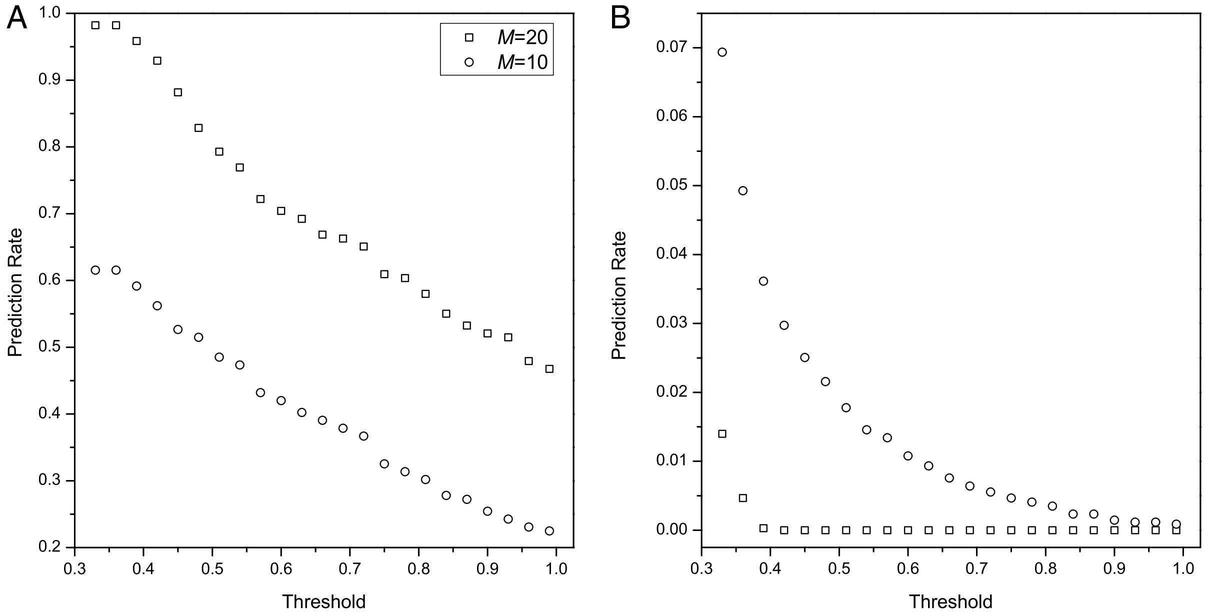


Fig. 4. Likelihood assessment corresponding to the example of Fig. 3. (A) Rate of interactions predicted correctly, with an a posteriori probability greater than a given threshold, as a function of the threshold for $M = 20$ (squares) corresponding to the example of Fig. 3, and $M = 10$ (circles). (B) Rate of noninteractions predicted wrongly, with an *a posteriori* probability greater than a given threshold, as a function of the threshold for $M = 20$ (squares) and $M = 10$ (circles).

neural network learning [25]. One alternative to avoid this low performance could be the use of a technique borrowed from the data compression field, which consists in approximating the matrix $\mathbf{X}^T$ by a lower rank matrix [26,27].

The partial recovery referenced above does not pursue to recover a subnetwork, which mainly infers strong interactions around the whole network. However, in many cases this is crucial to recovering the complete subnetwork associated with a given gene or path of interest. Here the term subnetwork refers to a network formed by nodes connected between them, but not connected, or weakly connected, to the other nodes of the whole network. The inference approach and information gain tool presented in this study could be used to establish new relationships between genes and to propose new experiments. By means of cycles of experiments-datamining, the knowledge about the subnetwork can be refined until its complete recovery, even in the presence of observational noise. For that purpose the following protocol could be used: (i) perform an initial perturbation where the gene of interest is overexpressed, and obtain the genome expression profile; (ii) compute the information gain for each gene with these experimental data; (iii) select the genes for which the information gain is greater than a given threshold; (iv) iterate the first two steps, perturbing each one of the genes that were selected in the third step and have still not been perturbed, until no new gene has an information gain greater than the threshold. Fig. 6(A) illustrates the result of three of these experiments-datamining cycles. First, the gene that belongs to the subnetwork of interest, gene g1, is initially overexpressed (with a level of 10.0, while the other gene levels are randomly selected in the range $[-0.5, 0.5]$). Then the input–output network is measured and this measurement is subject to observational noise with $\eta = 0.30$. The information gain of this experiment is computed for each gene using $p_{ij}^{+} = p_{ij}^{-} = 0.01$ as an *a priori* probability. Subsequently, those genes with I_g greater than 1.0 are selected. I_g suggests that gene g6 is regulated by g1. By repeating the above step with gene g6, the results indicate that genes g2 and g3 are regulated by g6. The above step is repeated with gene g2 and subsequent genes with high information gain values in ensuing experiments, until no new gene with an information gain greater than the threshold appears. Fig. 6(B) illustrates a list of experiments where the first column corresponds to the gene that was perturbed in the experiment, and the second column corresponds to the genes that appear to be regulated by the perturbed gene. In the last two experiments no new regulated genes appeared (which were not indicated in the first column list). The above analysis provides a causal link between two genes, but it does not indicate whether the regulation is positive or negative. In order to extract this information, the inference analysis was performed using the ten overexpression experiments pooled in D_M ($M = 10$). When the inference procedure was applied with these data, 19 out of 24 interactions in the subnetwork were inferred correctly, 10 of them with the *a posteriori* probability greater than 0.99. However, the *a priori* probabilities provided by the information contained in the list of Fig. 5(B) are included. By setting $p_{ij}^{+} = p_{ij}^{-} = 0.5$ (or 1/3) for all the pairs i, j indicated in the list, and $p_{ij}^{+} = p_{ij}^{-} = 0.01$ for the rest, 23 out of 24 interactions in the subnetwork are inferred, 19 of them with an *a posteriori* probability greater than 0.99 (Fig. 6(C)). The performance obtained above does not differ significantly if the inference procedure is implemented using $\mathbf{W}_{L_2}$ or $\mathbf{W}_{L_1}$ prescriptions. Nevertheless, $\mathbf{W}_{L_2}$ is computationally cheaper than $\mathbf{W}_{L_1}$ since the latter requires linear programming optimization. The above example about subnetwork inference suggests that this novel scheme can be reused to infer additional subnetworks until the whole network is recovered with $M \simeq N$ experiments.

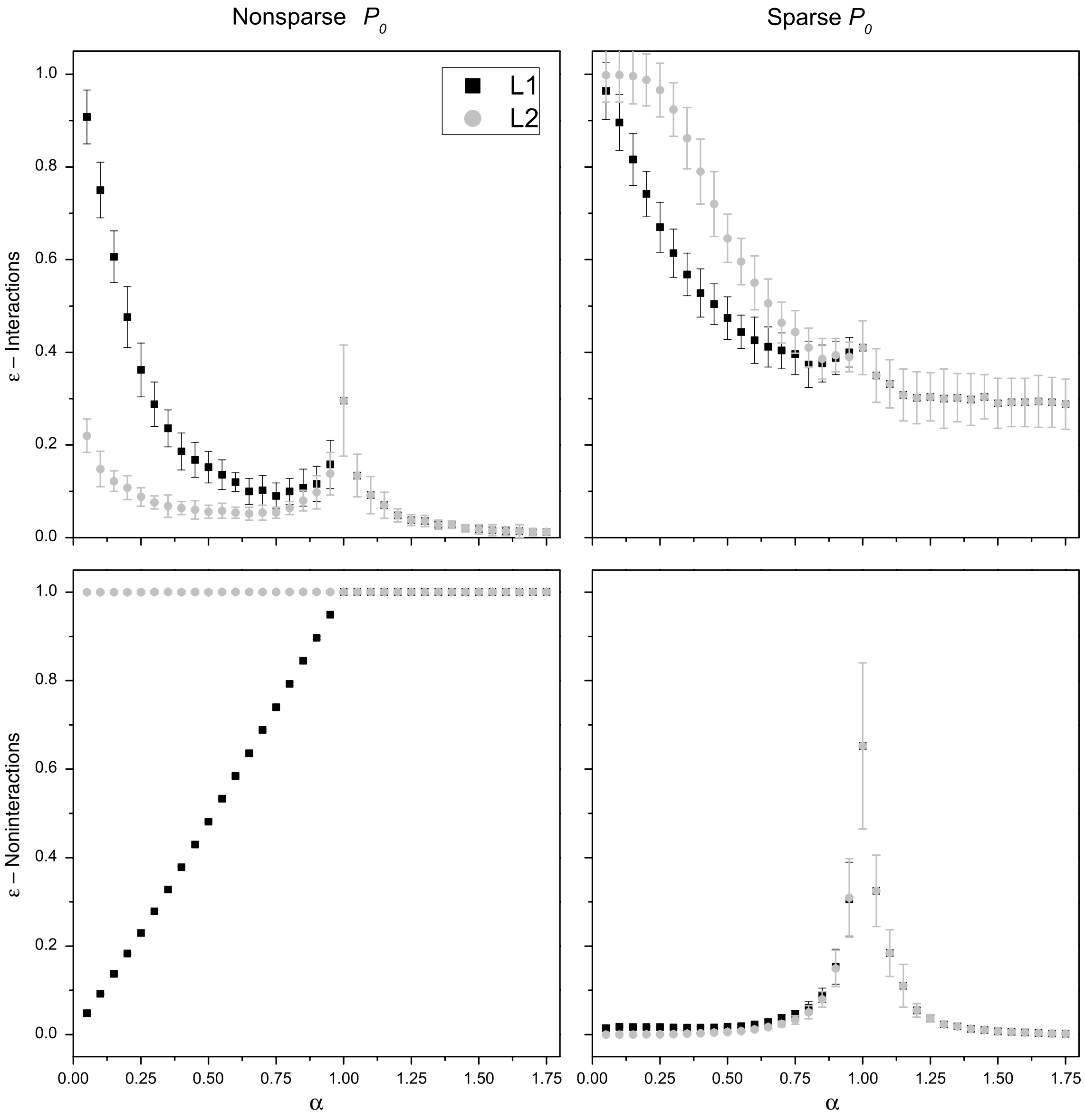
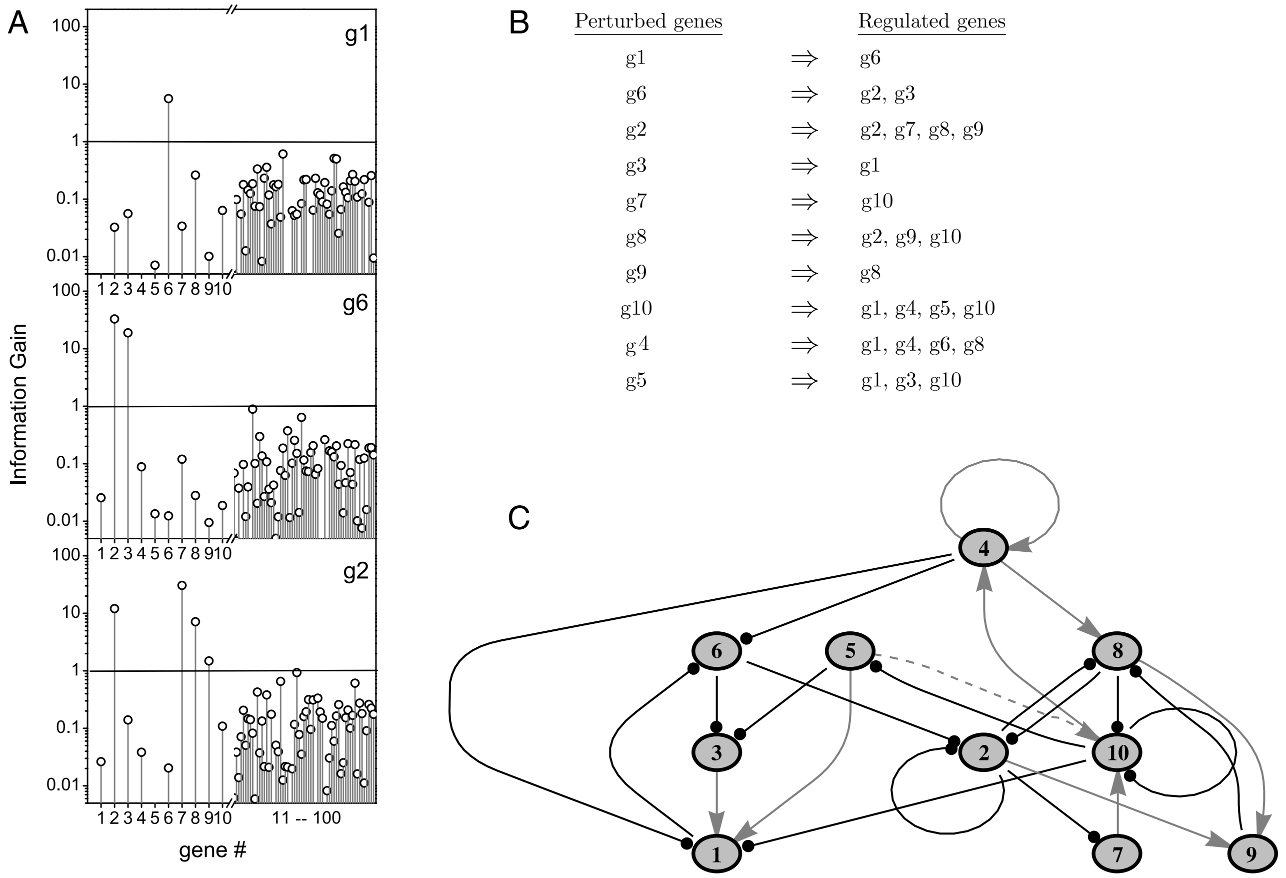


Fig. 5. Inference using noisy data. Prediction error ε as a function of the ratio α for gene networks with 60 genes with $k/N = 0.05$. Both input and output data are subject to observational noise of $\eta = 0.30$. The performances were obtained using both W_{L_1} (black squares) and W_{L_2} (gray circles) prescriptions and $a = 0.01$. Top panels correspond to interaction calls while bottom panels correspond to noninteractions calls. In the left panels the a priori probabilities are equal, i.e. $p_{ij}^{+} = p_{ij}^{-} = p_{ij}^{0} = 1/3$ for all i and j, while the *a priori* probabilities are set to be $p_{ij}^{+} = p_{ij}^{-} = 0.025$ and $p_{ij}^{0} = 0.95$ for all i and j.

4. Discussion and conclusions

Information theoretic principles have been applied to infer connections between the nodes of a gene regulatory network. In particular it was postulated that estimating pairwise gene expression profiles can be useful to identify candidate interactions. In this sense, the expression levels are considered as stochastic variables and used to compute the mutual information between each pair of gene expression profiles. The prediction of this kind of approach (known as association network approach) is limited to predicting an undirected graph [7]. Thus, this approach indicates whether two genes interact or not, but it is not able to determine who the regulated (or regulator) gene is, or whether the regulator is an activator or an inhibitor. Between this kind of approach we can mention ARCANE [28]. The approach represented here is one step further, because it allows predicting who the regulator is and which the nature of this regulation is (activator or repressor). Interestingly the new approach could also integrate results previously obtained with ARACNE, by setting to zero all *a priori* probabilities corresponding to putative interactions that were not predicted by ARACNE.

A novel approach for regulatory network inference is presented in this study. Unlike to other methods, this approach pursues to infer the type of interaction rather than a weight that characterizes the interaction quantitatively. Three main features of the proposed method are pointed out. First, it allows introducing global *a priori* information about the network, such as sparseness and other gene-dependent available information, as illustrated in the last example (Fig. 6(C)). Second, the information theory formalism provides a way to quantify the likelihood of the inferred paths, by using the *a*



Perturbed genes		Regulated genes
g1	⇒	g6
g6	⇒	g2, g3
g2	⇒	g2, g7, g8, g9
g3	⇒	g1
g7	⇒	g10
g8	⇒	g2, g9, g10
g9	⇒	g8
g10	⇒	g1, g4, g5, g10
g4	⇒	g1, g4, g6, g8
g5	⇒	g1, g3, g10

Fig. 6. Subnetwork identification. A: Information gain I_g obtained for three "overexpression experiments". First, the gene that belongs to the subnetwork of interest, gene g1, is initially overexpressed, then the input–output network is measured, and this measurement is subject to observational noise of $\eta = 0.30$. The information gain of this experiment is computed for each gene, and genes with I_g greater than a given threshold are selected. I_g suggests that gene g6 is regulated by g1. Repeating the above step with gene g6, it appears that genes g2 and g3 are regulated by g6. The above step is repeated with gene g2 and subsequent genes with high information gain values in subsequent experiments. B: List of experiments, the first column corresponds to the gene that was overexpressed in each experiment, the second column corresponds to the genes that appear to be regulated by the overexpressed gene. C: The subnetwork inferred 23 out of 24 interactions correctly (solid edges) by this inference procedure using W_{L_2} prescription and the above ten "overexpression experiments" together. The information contained in list B was included as *a priori* probabilities, i.e. they were set to $p^+_{ij} = p^-_{ij} = 0.5$ and $p^0_{ij} = 0.0$ for all i, j pairs indicated in the list, and $p^+_{ij} = p^-_{ij} = 0.01$ otherwise.

posteriori probabilities computed with the method. Last, but not least, the information theory formalism also quantifies the information gained with the set of data to be used in the inference procedure. Furthermore, the present IT approach offers a promising perspective as a network inference protocol, and the methodology presented here introduces an information gain measure as a bonus. This study illustrates the way in which this quantity could be a useful tool to identify the downstream regulated genes in overexpression experiments. This feature allows a datamining-assisted way of unravelling the whole network with a number of experiments equal to the number of genes, even when dealing with a high level of observational noise. This IT approach enables the effective use of all the available information, in which each experiment is used as an individual constraint. Thus, the ensuing observation level becomes much richer than the standard one, where all the data define a fitness function to be optimized. Efficient management leads to more realistic results in inference.

The learning protocol presented here constitutes an additional inference technique of interest not only for basic research but also as an application for an very interesting real world problem without paying an excessive computational cost.

Acknowledgements

The author thanks Christina McCarthy for critical review of the manuscript. L.D. is researcher of CONICET (Argentina).

References

[1] T. Ideker, T. Galitski, L. Hood, A new approach to decoding life: systems biology, Annu. Rev. Genomics Hum. Genet. 2 (2001) 343–372.

[2] I.M. Tienda-Luna, Y. Yin, M.C. Carrion, Y. Huang, H. Cai, M. Sanchez, Y. Wang, Inferring the skeleton cell cycle regulatory network of malaria parasite using comparative genomic and variational Bayesian approaches, Genetica 132 (2008) 131–142.

[3] T.S. Gardner, D. Di Bernardo, D. Lorenz, J.J. Collins, Inferring genetic networks and identifying compound mode of action via expression profiling, Science 301 (2003) 102–105.

[4] J. Tegnér, J. Björkegren, Perturbations to uncover gene networks, TIG 23 (2007) 34–41.

[5] N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, J.R. Banavar, Dynamic modeling of gene expression data, Proc. Natl. Acad. Sci. USA 98 (2001) 1693–1698.

[6] J. Tegnér, M.K. Yeung, J. Hasty, J.J. Collins, Reverse engineering gene networks, integrating genetic perturbations with dynamical modeling, Proc. Natl. Acad. Sci. USA 100 (2003) 5944–5949.

[7] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, T.S. Gardner, Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, PLoS Biol. 5 (2007) e8.

[8] J. Zola, M. Aluru, S. Aluru, Parallel Information Theory Based Construction of Gene Regulatory Networks, in: Lecture Notes in Computer Science, vol. 5374, 2008, pp. 336–349.

[9] D. Pe'er, A. Regev, G. Elidan, N. Friedman, Inferring subnetworks from perturbed expression profiles, Bioinformatics 17 (2001) S215–S224.

[10] D. Husmeier, Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks, Bioinformatics 19 (2003) 2270–2282.

[11] M.J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, D.L. Wild, A Bayesian approach to reconstructing genetic regulatory networks with hidden factors, Bioinformatics 21 (2005) 349–356.

[12] T. Akutsu, S. Miyano, S. Kuhara, Inferring qualitative relations in genetic networks an metabolic pathways, Bioinformatics 16 (2000) 727–734.

[13] P. D'haeseleer, S. Liang, R. Somogyi, Genetic network inference: from co-expression clustering to reverse engineering, Bioinformatics 16 (2000) 707–726.

[14] P. DH́aeseleer, X. Wen, S. Fuhrman, R. Somogyi, Linear modeling of MMA expression levels during CNS development and injury, Pac. Symp. Biocomput. 41 (1999) 52.

[15] E.P. van Someren, L.F. Wessels, M.J. Reinders, Linear modeling of genetic networks from experimental data, Proc. Int. Conf. Intell. Syst. Mol. Biol. 8 (2000) 355.

[16] E.P. van Someren, L.F.A. Wessels, E. Backer, M.J.T. Reinders, Genetic network modeling, Pharmacogenomics 3 (2002) 507–525.

[17] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Chicago, 1949.

[18] E.T. Jaynes, Information theory and statistical mechanics II, Phys. Rev. 108 (1957) 171–190.

[19] T. Lezon, J. Banavar, M. Cieplak, A. Maritan, N.V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, Proc. Natl. Acad. Sci. USA 103 (2006) 19033–19038.

[20] O. Martínez, M.H. Reyes-Valdés, Defining diversity, specialization, and gene specificity in transcriptomes through information theory, Proc. Natl. Acad. Sci. USA 105 (2008) 9709–9714.

[21] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, L. Chen, Inferring gene regulatory networks from multiple microarray datasets, Bioinformatics 22 (2006) 2413–2420.

[22] M.K. Yeung, J. Tegner, J.J. Collins, Reverse engineering gene networks using singular value decomposition and robust regression, Proc. Natl. Acad. Sci. USA 99 (2002) 6163–6168.

[23] R.D. Levine, M. Tribus, The Maximum Entropy Principle, MIT Press, Boston MA, 1978.

[24] A.L. Barabasi, R. Albert, H. Jeong, Mean-feld theory for scale-free random networks, Physica A 272 (1999) 173–187.

[25] T. Watkin, A. Rau, M. Biehl, The statistical mechanics of learning a rule, Rev. Mod. Phys. 65 (1993) 499–556.

[26] C.D. Cantrell, Modern Mathematical Methods for Physicists and Engineers, Cambridge University Press, Cambridge, 2000, page 514.

[27] J.J. Wei, C.J. Chang, N.-K. Chou, G.J. Jan, ECG data compression using truncated singular value decomposition, IEEE Trans. Inform. Technol. Biomed. 5 (2001) 290–299.

[28] J. Watkinson, K.C. Liang, X. Wang, T. Zheng, D. Anastassiou, Inference of regulatory gene interactions from expression data using three-way mutual information, Ann. NY Acad. Sci. 1158 (2009) 302–313.