



## A model-based optimization framework for the inference on gene regulatory networks from DNA array data

Reuben Thomas<sup>1</sup>, Sanjay Mehrotra<sup>1</sup>, Eleftherios T. Papoutsakis<sup>2</sup> and Vassily Hatzimanikatis<sup>2,\*</sup>

<sup>1</sup>Department of Industrial Engineering and Management Science and <sup>2</sup>Department of Chemical and Biological Engineering, 2145 Sheridan Road, E136, Northwestern University, Evanston, IL 60208-3120, USA

Received on March 8, 2004; revised on April 30, 2004; accepted on June 22, 2004

Advance Access publication July 9, 2004

### ABSTRACT

**Motivation:** Identification of the regulatory structures in genetic networks and the formulation of mechanistic models in the form of wiring diagrams is one of the significant objectives of expression profiling using DNA microarray technologies and it requires the development and application of identification frameworks.

**Results:** We have developed a novel optimization framework for identifying regulation in a genetic network using the S-system modeling formalism. We show that balance equations on both mRNA and protein species led to a formulation suitable for analyzing DNA-microarray data whereby protein concentrations have been eliminated and only mRNA relative concentrations are retained. Using this formulation, we examined if it is possible to infer a set of possible genetic regulatory networks consistent with observed mRNA expression patterns. Two origins of changes in mRNA expression patterns were considered. One derives from changes in the biophysical properties of the system that alter the molecular-interaction kinetics and/or message stability. The second is due to gene knock-outs. We reduced the identification problem to an optimization problem (of the so-called mixed-integer non-linear programming class) and we developed an algorithmic procedure for solving this optimization problem. Using simulated data generated by our mathematical model, we show that our method can actually find the regulatory network from which the data were generated. We also show that the number of possible alternate genetic regulatory networks depends on the size of the dataset (i.e. number of experiments), but this dependence is different for each of the two types of problems considered, and that a unique solution requires fewer datasets than previously estimated in the literature. This is the first method that also allows the identification of every possible regulatory network that could explain the data, when the

number of experiments does not allow identification of unique regulatory structure.

**Availability:** The implementation of the algorithm in AMPL is available on request from the authors.

**Contact:** vassily@northwestern.edu

### 1 INTRODUCTION

The cell-wide monitoring of gene (Eisen *et al.*, 1998; Tamayo *et al.*, 1999) and protein expression and interactions (Dezso *et al.*, 2003; Giot *et al.*, 2003) generate a large amount of data that reflect the responses of the genetic/protein networks to environmental and genetic perturbations. DNA-microarray (or proteome analysis) data are used to cluster genes or proteins with respect to their similarity to expression changes compared with a control cellular condition or a 'global' or 'reference' set of transcripts. Clustering of expression profiles of unknown genes with genes of known functions may provide information regarding their function (Eisen *et al.*, 1998; Iyer *et al.*, 1999; Sorlie *et al.*, 2001; Tamayo *et al.*, 1999). Ultimately, however, it is anticipated that DNA-microarray and proteome data will be used to construct genetic or protein interaction networks either between individual genes or groups of genes (clusters) generated by one of the clustering methods. Thus, more sophisticated analysis of large-scale transcriptional and proteome data will require model-based identification methods in order to infer possible regulatory architectures in a genetic or protein network.

A genetic network model aims to capture the interrelated regulatory mechanisms between genes. Several genetic network models have been proposed, which integrate biochemical pathway information and expression data to trace genetic regulatory interactions (Akutsu *et al.*, 2000; Dasika *et al.*, 2004; Di Bernardo *et al.*, 2004; Ideker *et al.*, 2001; Lin *et al.*, 2003; Maki *et al.*, 2001; Moriyama *et al.*, 1999; Noda *et al.*, 1998; Wu *et al.*, 2004). They range from abstract Boolean descriptions to detailed mechanistic models to neural

\*To whom correspondence should be addressed.

network models (Bower and Bolouri, 2001; D’Haeseleer *et al.*, 2000). Every representation has its advantages and limitations. Some of these models have been used for the reverse engineering of the regulatory interactions of the network, i.e. the inference of the regulatory interactions between the expression of the genes in the network and their product proteins from mRNA profiles.

One of the first approaches to reverse engineering a genetic network based on the information of mRNA profiles employed a Boolean modeling framework (Somogyi *et al.*, 1997). Boolean networks model the state of the gene as either ON or OFF and the input–output relationships are postulated as logical functions. Further improvements of the Boolean-based frameworks introduced Mutual Information. (Liang *et al.*, 1998) and an entropy-based approach (Ideker *et al.*, 2000) in order to identify the experiments required for the discrimination between alternate genetic networks predicted by these frameworks. Rigorous analysis has led to the identification of the number of experiments required for identification of the Boolean model that is consistent with the data (Akutsu *et al.*, 2000; Somogyi *et al.*, 1997). However, in real systems, transcript levels vary in a continuous manner implying that the key assumptions underlying Boolean network models may not be appropriate, and thus, more general models may be necessary (de Jong, 2002).

Several continuous modeling frameworks have also been proposed, and they employ linear models (Chen *et al.*, 1999; D’Haeseleer *et al.*, 1999; Di Bernardo *et al.*, 2004; Weaver *et al.*, 1999; Wu *et al.*, 2004), hybrid Boolean and power law models (Akutsu *et al.*, 2000) and Bayesian models (Friedman *et al.*, 2000). Dasika *et al.* (2004) developed a linear model for inferring time delay in genetic networks. However, linear models cannot capture the inherent non-linearities of biological systems. A possible approach then is to use the S-system framework whereby the non-linearity of genetic regulation can be reasonably captured (Hlavacek and Savageau, 1992; Lin *et al.*, 2003; Maki *et al.*, 2001; Savageau, 1983, 1988).

In this paper, we propose an optimization framework for the identification of regulation in a genetic network using the S-system modeling formalism. Specifically, we are interested in whether we can infer a set of possible genetic regulatory networks that are consistent with observed expression patterns. The data that will be used to identify the network are obtained from the simulation of a model of a gene network, which corresponds to data obtained from DNA-microarray experiments. We specifically consider two cases of changes in the mRNA expression patterns. One is due to the changes in biophysical properties of the system that alter the molecular-kinetic constants and/or message stability. The second is due to knock-out experiments, whereby genes are removed from the system. We reduce the identification problems to optimization problems and we describe the algorithmic procedure that has been developed for solving the optimization problems.

We investigated the following two questions associated with the capabilities of the proposed method:

- Does the method actually find the regulatory network from which the data were generated?
- How does the number of possible alternate genetic regulatory networks depend on the size of the dataset, i.e. number of experiments?

An understanding of these issues would enable better design of experiments for data collection in order to infer meaningful genetic networks. This is a systematic study that provides specific evaluation of the use of continuous, S-systems models for the inference of genetic regulatory networks and a broader understanding of the relationship between the number of experiments and the model size or complexity.

## 2 METHODS

### 2.1 S-system modeling of genetic networks

Mathematical modeling of gene and protein expression begins with the formulation of the mass-balance equations for the mRNA and protein corresponding to any gene  $i$ :

$$\begin{aligned}\frac{dM_i}{dt} &= V_{sm,i} - V_{dm,i}, \\ \frac{dP_i}{dt} &= V_{sp,i} - V_{dp,i},\end{aligned}\quad (1)$$

where  $M_i$  and  $P_i$  denote the concentrations of mRNA and protein, respectively;  $V_{sm,i}$  and  $V_{sp,i}$  denote the synthesis rates of mRNA and protein, respectively; and  $V_{dm,i}$  and  $V_{dp,i}$  denote the degradation rates of mRNA and protein, respectively.

In the most common formalism of S-systems, and in most mathematical models of genetic networks, the degradation rates of mRNA and protein are assumed to be first order with respect to the corresponding species:

$$\begin{aligned}V_{dm,i} &= \beta_i \cdot M_i, \\ V_{dp,i} &= \delta_i \cdot P_i,\end{aligned}\quad (2)$$

where the first-order degradation constants,  $\beta$  and  $\delta$ , are inversely proportional to the half-life of the corresponding species.

The kinetics of protein synthesis rate is also considered to be first order with respect to mRNA concentration:

$$V_{sp,i} = \gamma_i \cdot M_i, \quad (3)$$

where  $\gamma_i$  is a rate constant that quantifies the translation efficiency.

The power of the S-systems representation lies in its ability to describe the non-linearities involved in transcriptional regulation by assuming that the mRNA synthesis rates follow power-law kinetics with respect to the concentration of the protein species that regulate the synthesis of the corresponding mRNA species. For example, if the synthesis of mRNA

for species  $i$  is induced by protein species  $j$  and it is repressed by protein species  $k$ , the corresponding transcription rate can be written as follows:

$$V_{sm,i} = \alpha_i \cdot P_j^{\varepsilon_{ij}} \cdot P_k^{\varepsilon_{ik}}, \quad (4)$$

where the exponents,  $\varepsilon_{ij}$  and  $\varepsilon_{ik}$ , capture the non-linearities of the regulatory mechanism. A positive value for  $\varepsilon_{ij}$  captures the inducer activity of  $P_j$ . Similarly, a negative value for  $\varepsilon_{ik}$  captures the repressor activity of  $P_k$ . In addition, the values of the rate constants,  $\alpha_i$ , quantify the strength of the regulatory interactions in the genetic network since they quantify the transcription efficiency of the active complexes of the regulatory proteins with the corresponding regulatory regions. The higher their value is, the stronger the effect of changes in the levels of the regulatory proteins on the transcription efficiency of the target genes is.

The choice of the value of the exponents in the power-law representation depends on how well it approximates the non-linear aspects of the corresponding regulatory mechanisms. Choosing a value for the exponent allows the description of the differential effects of distinct genes on each other as the first derivative of the power-law representation depends on the concentration of the corresponding protein and it is not constant.

Most of the mechanisms in the regulation of gene expression follow hyperbolic, saturation kinetics of the form:

$$V_{sm,i} \propto \frac{P_j}{K_j + P_j} \quad \text{and} \quad V_{sm,i} \propto \frac{K_k}{K_k + P_k} \quad (5)$$

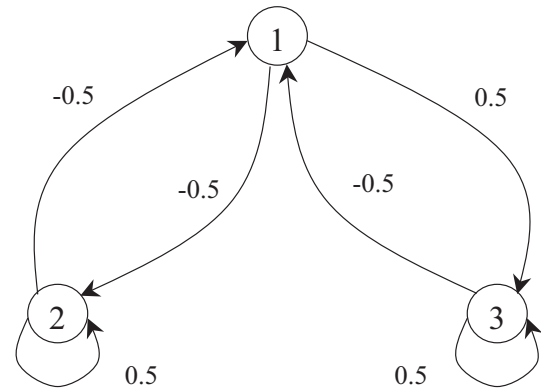
for the inducer and the repressor, respectively. Voit and Savageau (1987) have studied in detail how many common biochemical rate laws, including the ones in Equation (5), could be most accurately approximate by using power-law expression. Following their analysis, we argue in the Appendix 1 that the best choices for approximating the non-linear expression of Equation (5) are the following power-law expressions:

$$V_{sm,i} \propto P_j^{0.5} \quad \text{and} \quad V_{sm,i} \propto P_k^{-0.5} \quad (6)$$

for the inducer and the repressor, respectively. Although more complex mechanisms and mathematical representations are possible, the S-systems formalism provides a substantial improvement with respect to linear models and is a minimal representation of complex, multiparametric rate expressions (Hlavacek and Savageau, 1998; Savageau, 1991).

## 2.2 Prototype system

We will consider here a simple network of three genes (Fig. 1). Using the S-system formalism, the mass balances for the



**Fig. 1.** Graph representation of the three-gene network. The nodes of the graph represent the genes. The number on the arc from gene  $k$  to gene  $j$  represents how gene  $k$  regulates gene  $j$ . The sign of this number indicates whether this is a positive regulation (induction) or a negative one (repression). If there is no arc between gene  $k$  and gene  $j$  then gene  $k$  plays no role in regulating the synthesis of the mRNA of gene  $j$ .

mRNA of each gene and its protein product can be written as:

$$\begin{aligned} \frac{dM_1}{dt} &= \alpha_1 P_2^{-0.5} P_3^{-0.5} - \beta_1 M_1, & \frac{dP_1}{dt} &= \gamma_1 M_1 - \delta_1 P_1, \\ \frac{dM_2}{dt} &= \alpha_2 P_1^{-0.5} P_2^{0.5} - \beta_2 M_2, & \frac{dP_2}{dt} &= \gamma_2 M_2 - \delta_2 P_2, \\ \frac{dM_3}{dt} &= \alpha_3 P_1^{0.5} P_3^{0.5} - \beta_3 M_3, & \frac{dP_3}{dt} &= \gamma_3 M_3 - \delta_3 P_3, \end{aligned} \quad (7)$$

where  $M_i$  and  $P_i$  represent the concentrations of the  $i$ -th mRNA and protein, respectively.  $\alpha_i$ ,  $\beta_i$ ,  $\gamma_i$  and  $\delta_i$  are the kinetic constants of the respective reactions.

The following set of assumptions summarizes what has been already stated and will be used throughout this paper:

- (1) A gene cannot act as a regulator for more than a certain number of genes. Typically, this number would be much smaller than the total number of genes.
- (2) The degradation reactions for both the mRNAs and the proteins follow first-order kinetics.
- (3) The protein synthesis rate is a first-order expression with respect to its corresponding mRNA concentration.
- (4) The non-linear relationship between gene expression and a regulatory protein is approximated by Equation (6).

At steady-state, using the transformation presented in Appendix 2, the mass-balance equations of the system are

reduced to the following three equality equations:

$$\begin{aligned}\phi_1 - 0.5 \ln(M_2) - 0.5 \ln(M_3) &= \ln(M_1), \\ \phi_2 - 0.5 \ln(M_1) + 0.5 \ln(M_2) &= \ln(M_2), \\ \phi_3 + 0.5 \ln(M_1) + 0.5 \ln(M_3) &= \ln(M_3),\end{aligned}\quad (8)$$

where  $\phi_i$ 's are expressions involving the kinetic parameters of mRNA and protein synthesis and degradation (Appendix 2).

This transformation allows expression of the connectivity of the network in terms of mRNA concentrations eliminating protein concentrations. This eliminates the need for information about protein levels without ignoring the fact that it is the proteins that carry out the regulatory actions. The steady-state assumption is not very restrictive, as one can easily argue that in most cases of actual experimental conditions, the pseudo-steady-state assumption can be readily justified to mRNA and protein species so that mRNA profiles from transient batch experiments can be employed to generate the necessary data for the identification problem.

### 2.3 The physiological studies under consideration

Analysis of DNA-microarray data aims to eventually infer the genetic regulatory network based on observations of the relative changes in message abundance between two different cellular states. These changes can be the result of changes at the genome level, such as point mutations or gene knock outs, or changes in the environmental conditions, such as changes in oxygen concentration, pH or nutrient and growth factor concentration. However, from the biophysical point of view, these changes can be classified into two groups with respect to their effect on the system parameters and structure. In one case, some mutations will affect the efficiency of transcription or the half-life of the mRNA and protein. For example, while regulatory complexes are formed, their effect on the initiation of transcription is reduced. Such genetic changes, as well as some changes in the environmental conditions may alter the biophysical properties of the regulatory mechanisms and they will affect the transcriptional efficiency of the corresponding genes. In the other case, some mutations could result in the loss of binding of the regulatory protein, and they will ultimately eliminate the corresponding regulatory connection. The extreme example of this case is the deletion of a gene from the genome of the organism in a knock-out experiment. In the first case, when the transcriptional efficiency is affected due to genetic or environmental changes, the values of the corresponding kinetic constants, transcription, and translation efficiency or degradation rate constants, will effectively change. In the second case, when regulatory connections are lost or severely weakened, the mass-balance equations of the corresponding mRNA and protein will be completely eliminated from the system, or the absolute value of the exponents of the corresponding proteins will be set close to zero.

Here, for simplicity we will consider two cases: (1) changes in the transcription efficiency or half-life of the message or protein and (2) knock-out experiments. In both cases, the exponent of the regulatory proteins will remain unchanged, to an absolute value of 0.5, assuming that the causes of the changes of the transcriptional profile are not due to changes in the ability of the regulatory proteins to bind the DNA regulatory regions. We will refer to the first case as a biophysical perturbation study and to the second case as knock-out study.

### 2.4 Algorithm development

Although the algorithms developed for the two cases share some of the underlying concepts, they are also different due to the basic differences of the mechanistic origins of the changes in the transcriptional profiles. We will illustrate the development of these algorithms using the three-gene system described above.

*2.4.1 Biophysical perturbation study* The perturbations considered here result in changes in the kinetic rate constants associated with transcription efficiency and message half-life, the  $\alpha$ 's and  $\beta$ 's, respectively. Let superscript '0' denote the reference steady-state condition of the three-gene network, and superscripts '1' and '2' denote the steady-state conditions arising from two perturbations in the biophysical properties of the system. Under these considerations, Equation (8) for each of the genes could be rewritten as:

$$\begin{aligned}\ln \frac{(\alpha/\beta)_1^1}{(\alpha/\beta)_1^0} - 0.5 \ln \left( \frac{M_2^1}{M_2^0} \right) - 0.5 \ln \left( \frac{M_3^1}{M_3^0} \right) &= \ln \left( \frac{M_1^1}{M_1^0} \right) \\ \ln \frac{(\alpha/\beta)_1^2}{(\alpha/\beta)_1^0} - 0.5 \ln \left( \frac{M_2^2}{M_2^0} \right) - 0.5 \ln \left( \frac{M_3^2}{M_3^0} \right) &= \ln \left( \frac{M_1^2}{M_1^0} \right).\end{aligned}\quad (9)$$

$$\begin{aligned}\ln \frac{(\alpha/\beta)_2^1}{(\alpha/\beta)_2^0} - 0.5 \ln \left( \frac{M_1^1}{M_1^0} \right) - 0.5 \ln \left( \frac{M_2^1}{M_2^0} \right) &= \ln \left( \frac{M_2^1}{M_2^0} \right) \\ \ln \frac{(\alpha/\beta)_2^2}{(\alpha/\beta)_2^0} - 0.5 \ln \left( \frac{M_1^2}{M_1^0} \right) - 0.5 \ln \left( \frac{M_2^2}{M_2^0} \right) &= \ln \left( \frac{M_2^2}{M_2^0} \right).\end{aligned}\quad (10)$$

$$\begin{aligned}\ln \frac{(\alpha/\beta)_3^1}{(\alpha/\beta)_3^0} + 0.5 \ln \left( \frac{M_1^1}{M_1^0} \right) + 0.5 \ln \left( \frac{M_3^1}{M_3^0} \right) &= \ln \left( \frac{M_3^1}{M_3^0} \right) \\ \ln \frac{(\alpha/\beta)_3^2}{(\alpha/\beta)_3^0} + 0.5 \ln \left( \frac{M_1^2}{M_1^0} \right) + 0.5 \ln \left( \frac{M_3^2}{M_3^0} \right) &= \ln \left( \frac{M_3^2}{M_3^0} \right).\end{aligned}\quad (11)$$

This system of equations could be compactly written in matrix notation as:

$$A + CX = X, \quad (12)$$

where

$$\begin{aligned}
 \mathbf{A} &= \begin{pmatrix} \ln \frac{(\alpha/\beta)_1^1}{(\alpha/\beta)_1^0} & \ln \frac{(\alpha/\beta)_1^2}{(\alpha/\beta)_1^0} \\ \ln \frac{(\alpha/\beta)_2^1}{(\alpha/\beta)_2^0} & \ln \frac{(\alpha/\beta)_2^2}{(\alpha/\beta)_2^0} \\ \ln \frac{(\alpha/\beta)_3^1}{(\alpha/\beta)_3^0} & \ln \frac{(\alpha/\beta)_3^2}{(\alpha/\beta)_3^0} \end{pmatrix}, \\
 \mathbf{X} &= \begin{pmatrix} \ln \left( \frac{M_1^1}{M_1^0} \right) & \ln \left( \frac{M_1^2}{M_1^0} \right) \\ \ln \left( \frac{M_2^1}{M_2^0} \right) & \ln \left( \frac{M_2^2}{M_2^0} \right) \\ \ln \left( \frac{M_3^1}{M_3^0} \right) & \ln \left( \frac{M_3^2}{M_3^0} \right) \end{pmatrix} \quad \text{and} \quad (13) \\
 \mathbf{C} &= \begin{pmatrix} 0 & -0.5 & -0.5 \\ -0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}. \quad (14)
 \end{aligned}$$

In a common experiment, the unknowns are usually the elements of the matrices  $\mathbf{C}$  and  $\mathbf{A}$ . The element values of matrix  $\mathbf{X}$  are obtained from DNA array experiments, or for our studies here, from the simulation model. If there is no knowledge about the exact biophysical parameters that are changing during the perturbation, the problem of inferring the regulatory connections is a computationally hard problem (see Discussion section). In the following studies, we make the assumption that the changes in kinetic constants of the mRNA synthesis and degradation reactions are known. Hence, we will assume that matrix  $\mathbf{A}$  is known. Although this is not the case in most of the known experimental reports, one could design such experiments through genetic and environmental perturbations that will target the expression of specific genes, similar to the experiments by Ideker *et al.* (2001) and Gardner *et al.* (2003). For the scope of this work, analysis of the system under these assumptions will provide an understanding on the performance of the developed methodologies and some general conclusions about the type of information required in order to infer meaningful regulatory networks.

**2.4.2 Knock-out study** The knock-out study involves inactivating or deleting a gene from the system. The corresponding simulation studies involve the removal of the mass balances of the mRNA and protein, while the other kinetic constants remain unchanged between the different knock-out systems. However, eliminating completely a gene from the system causes numerical problems in simulating this experiment. If the gene that is being removed negatively regulates another gene in the system then the synthesis rate of the regulated gene

will tend to be infinity since the concentration of the regulator gene that is removed is raised to a negative power [see Equation (4)]. To overcome this problem and to ensure that the expression level of the ‘knocked out’ gene remains close to zero, we set the kinetic constants corresponding to the synthesis and degradation reactions (of this gene) to zero. A small threshold number is added to each of the protein concentrations in Equation (7). Therefore, the ‘knocked out’ gene would still be present at a non-zero but in very small concentration.

For the three-gene system, let the two experiments performed involve the removal of genes 1 and 2, respectively. The relevant mass-balance equations would then be:

For gene 1,

$$-0.5 \ln \left( \frac{M_2^2}{M_2^0} \right) - 0.5 \ln \left( \frac{M_3^2}{M_3^0} \right) = \ln \left( \frac{M_1^2}{M_1^0} \right). \quad (15)$$

For gene 2,

$$-0.5 \ln \left( \frac{M_1^1}{M_1^0} \right) + 0.5 \ln \left( \frac{M_2^1}{M_2^0} \right) = \ln \left( \frac{M_2^1}{M_2^0} \right). \quad (16)$$

For gene 3,

$$\begin{aligned}
 0.5 \ln \left( \frac{M_1^1}{M_1^0} \right) + 0.5 \ln \left( \frac{M_3^1}{M_3^0} \right) &= \ln \left( \frac{M_3^1}{M_3^0} \right). \\
 0.5 \ln \left( \frac{M_1^2}{M_1^0} \right) + 0.5 \ln \left( \frac{M_3^2}{M_3^0} \right) &= \ln \left( \frac{M_3^2}{M_3^0} \right).
 \end{aligned} \quad (17)$$

There is no convenient way to write these equations as one compact matrix equation. But it should become clear later that whatever technique is used to solve the inference problem of the biophysical perturbation studies should be able to solve the inference problem involving the knock-out studies.

## 2.5 Optimization model

If our model (of the genetic network) was true, then we would expect Equation (12) (in the biophysical perturbation study) to be satisfied exactly. However, in light of the several assumptions that we made including the model itself, there is no reason to expect that our model represents the actual genetic network in every detail. There is also the added feature that there will always be noise in the data obtained from a real system. Under these considerations, Equation (12) can be modified to include these deviations:

$$\mathbf{R} = \mathbf{A} + \mathbf{C}\mathbf{X} - \mathbf{X} \Rightarrow \mathbf{R} = \mathbf{A} + (\mathbf{C} - \mathbf{I})\mathbf{X}, \quad (18)$$

where matrix  $\mathbf{R}$  represents the deviation from exact equality of the Equation (12), and  $\mathbf{I}$  is the identity matrix. If  $\mathbf{R}$  is the zero matrix, then the mass-balance equations are satisfied. The closer  $\mathbf{R}$  is to the zero matrix, the closer Equation (12) represents the mass-balance equations. Here, we will use the

Euclidean norm as a measure of the distance of  $\mathbf{R}$  from the zero matrix, although other matrix norms could also be used to quantify this distance.

The problem then is the identification of regulatory networks that satisfy the mass-balance equations as closely as possible. This problem can then be formulated as an optimization problem as follows:

**PROBLEM 1.** *Given known matrices  $\mathbf{X}$  and  $\mathbf{A}$ , find the matrix  $\mathbf{C}$  that minimizes matrix  $\mathbf{R}$  such that*

- *The elements of  $\mathbf{C}$  are either  $-0.5$ ,  $0$  or  $0.5$ . This corresponds to the assumption that a protein can be either repressor ( $-0.5$ ) or inducer ( $0.5$ ) of the same gene, or there is no regulatory interaction ( $0$ ).*
- *The number of non-zero elements in each row of  $\mathbf{C}$  is less than or equal to a fixed number,  $k$ . This corresponds to the assumption that we do not expect a gene to regulate more than a certain number of genes.*

For a general case, we assume that we are dealing with an  $n$ -gene system and that we have transcriptional data from an  $l$  number of experiments. The problem is then reduced to determining  $n^2$  unknowns, the elements of matrix  $\mathbf{C}$ .

A key observation is that this problem of determining the best possible values of  $n^2$  unknowns is equivalent to  $n$  problems where in each problem we have  $n$  unknowns (variables) to be determined.

**THEOREM 1.** *Solving Problem 1 (in  $n^2$  dimensions) is equivalent to solving  $n$  problems, each of them in  $n$  dimensions.*

**PROOF.** See Appendix 3.

The optimization problem for the three-gene network can be formulated as follows. Let  $\mathbf{Y} = \mathbf{C} - \mathbf{I}$ , and  $y_i^T$  represents the  $i$ -th row of matrix  $\mathbf{Y}$ . For example, for gene 1

$$\mathbf{y}_1^T = [e_1 \cdot d_{1,1,1} + e_2 \cdot d_{1,1,2} - 1 \quad e_1 \cdot d_{1,2,1} + e_2 \cdot d_{1,2,2} \quad e_1 \cdot d_{1,3,1} + e_2 \cdot d_{1,3,2}], \quad (19)$$

where superscript T denotes the transpose of the corresponding vectors and matrices,  $e_i$  are the possible non-zero values of the elements of matrix  $\mathbf{C}$  ( $e_1 = 0.5$  and  $e_2 = -0.5$ ) and  $d_{i,j,k}$  are binary variables that can only take values 0 or 1.

The identification of the value of these binary variables is essentially equivalent to the identification of the regulatory architecture of the genetic network. For example, if  $d_{1,2,1} = 1$ , then protein product of gene 2 induces mRNA synthesis of gene 1, and the binary variable  $d_{1,2,2}$  is set to 0 since we do not allow a gene to be both an inducer and a repressor of the same gene. Similarly, if both  $d_{1,3,1}$  and  $d_{1,3,2}$  are equal to zero, then gene 1 is not subject to regulation by protein product of gene 3.

Let  $\mathbf{a}_1^T$  be the first row of the  $\mathbf{A}$  matrix and  $\|\cdot\|^2$  denotes the norm of a vector as defined before. The objective can then

be stated as

$$\min \|\mathbf{X}^T \mathbf{y}_1 + \mathbf{a}_1\|^2. \quad (20)$$

There are also two constraints that should be satisfied based on the assumptions of our approach. To ensure that the elements of  $\mathbf{y}_1$  are only  $-0.5$ ,  $0$  or  $0.5$ , the binary variables,  $d_{i,j,k}$ , should satisfy the following inequality constraints:

$$\begin{aligned} d_{1,1,1} + d_{1,1,2} &\leq 1. \\ d_{1,2,1} + d_{1,2,2} &\leq 1. \\ d_{1,3,1} + d_{1,3,2} &\leq 1. \end{aligned} \quad (21)$$

The constraint on the maximum number of regulators would be described by the following inequality equation:

$$d_{1,1,1} + d_{1,1,2} + d_{1,2,1} + d_{1,2,2} + d_{1,3,1} + d_{1,3,2} \leq k. \quad (22)$$

Similar optimization problems can be written for the other two rows. However, since the number of experiments that would be performed would typically be less than the number of genes, it is likely that we would get a large number of inferred alternate genetic networks that result in the same distance of matrix  $\mathbf{R}$  from the zero matrix. In order to identify every possible solution that minimizes the norm of matrix  $\mathbf{R}$  we add to the problem an additional constraint on the binary variable for every previously found solution. If  $\mathbf{d}^1$  is the first solution generated by the optimization algorithm, we solve again the problem with the additional following inequality constraint:

$$\begin{aligned} d_{1,1,1}^1 \cdot d_{1,1,1} + d_{1,1,2}^1 \cdot d_{1,1,2} + d_{1,2,1}^1 \cdot d_{1,2,1} + d_{1,2,2}^1 \cdot d_{1,2,2} \\ + d_{1,3,1}^1 \cdot d_{1,3,1} + d_{1,3,2}^1 \cdot d_{1,3,2} \leq D^1 - 1, \end{aligned} \quad (23)$$

where  $D^1$  is the number of non-zero binary variables in the first solution.

$$D^1 = d_{1,1,1}^1 + d_{1,1,2}^1 + d_{1,2,1}^1 + d_{1,2,2}^1 + d_{1,3,1}^1 + d_{1,3,2}^1. \quad (24)$$

In the  $r$ -th solution the problem will involve  $r - 1$  additional constraints of the form:

$$\begin{aligned} d_{1,1,1}^{r-1} \cdot d_{1,1,1} + d_{1,1,2}^{r-1} \cdot d_{1,1,2} + d_{1,2,1}^{r-1} \cdot d_{1,2,1} + d_{1,2,2}^{r-1} \cdot d_{1,2,2} \\ + d_{1,3,1}^{r-1} \cdot d_{1,3,1} + d_{1,3,2}^{r-1} \cdot d_{1,3,2} \leq D^{r-1} - 1 \end{aligned} \quad (25)$$

with

$$D^{r-1} = d_{1,1,1}^{r-1} + d_{1,1,2}^{r-1} + d_{1,2,1}^{r-1} + d_{1,2,2}^{r-1} + d_{1,3,1}^{r-1} + d_{1,3,2}^{r-1} \quad (26)$$

and will provide the  $r$ -th alternative regulatory network consistent with the experimental data.

The generalization of the problem formulation for  $n$  genes and  $l$  experiments is straightforward (Appendix 4). The optimization model for the data from the knock-out studies will be similar except for minor changes (see Appendix 4).

## 2.6 Simulated data

It has been proposed that algorithms developed for the inference on biological regulatory network inference should be evaluated with respect to their ability to successfully recover complex regulatory networks from simulated but biologically reasonable data (Smith *et al.*, 2002). An increasing number of studies is employing the use of simulated data for the development and evaluation of inference algorithms. (Gardner *et al.*, 2003; Hartemink *et al.*, 2002; Husmeier, 2003; Mendes *et al.*, 2003; Michaud *et al.*, 2003; Tegner *et al.*, 2003; Yeung *et al.*, 2002). The optimization model developed here was evaluated using data from a simulated, randomly generated genetic network modeled using the S-system formalism. In generating the data, we solve for the steady-state of the mRNA and protein mass balances for different values of the biophysical parameters (biophysical perturbation study) or after removal of the mRNA and protein mass balances of a gene species (knock-out study). The optimization problem was solved using the Mixed-Integer Nonlinear Programming (MINLP) solver on the NEOS server (Czyzyk *et al.*, 1998; Dolan, 2001; Gropp and More, 1997). The MINLP solver (Fletcher and Leffyer, 1998) employs the standard Branch and Bound procedure (Nemhauser and Wolsey, 1988) (for integer programming) using a depth-first search. The Branch and Bound method in short consists of developing a solution tree where each node represents an integer relaxation of the original problem. As one moves down the tree, more variables are constrained to take integer values. The bounding procedure allows one to implicitly examine; without actually solving any problem nodes. This procedure is not polynomial in the size of the problem and in a worse case could take an exponentially long time.

## 3 RESULTS

We investigated the relationship between the properties of the genetic network, such as the number of genes, and the number and type of experiments required for the identification of the regulatory architecture of the network. Theoretical considerations suggest that two main properties of the system that underlie this relationship are the number of genes in the system and the maximum number of regulatory inputs per gene. Since the number of experiments is usually smaller than the size of the genetic network, there could exist multiple, alternative regulatory structures that could explain the data. Therefore, we also investigated the dependence on the number of alternative regulatory structures on the number of experiments and on the system properties.

We first performed biophysical perturbation studies. In these studies, we chose the number of genes,  $n$ , we then formulated the corresponding mRNA and protein mass balances, and we randomly assigned regulatory inputs in the various transcription rate expressions allowing a maximum number,  $k$ , of inputs per gene. In addition, a regulatory protein is not

allowed to be both repressor and inducer for the same gene. We then used this system to perform simulation experiments where in each experiment a number,  $\eta$ , of biophysical parameters was randomly chosen and their values were perturbed relative to the reference values. The simulated profile on the concentration of the mRNAs was used for the calculation of the expression log-ratios and it constitutes one experimental dataset. Additional experiments were performed in a similar way, with a different, randomly chosen set of perturbed biophysical parameters for each experiment. The results of these studies are summarized in Table 1.

Analysis of the results in Table 1 suggests that there exist a very large number of alternative regulatory networks that can explain the data when the number of experiments is less than the number of the genes in the network. The set of alternative solutions includes the one used to generate the data. For systems with the same number of genes and the same number of allowable regulatory inputs per gene, but with different regulatory structures, we observed different number of alternative solutions for the same number of experiments. These observations suggest that the number of the alternative solutions depends also on the structure of the regulatory network as well as the combination of the biophysical parameters that are responsible for the changes in the expression profiles. The number of alternative network solutions rapidly falls to a unique solution when the ratio of the number of experiments to the numbers of genes is about one-half.

These observations could be explained by considering the structure of the system. The experimental data satisfy Equation (12) which can be rewritten as:

$$(C - I)X = -A. \quad (27)$$

Each element in a column of matrix  $X$  is the ratio of the expression profile of a gene from an experiment relative to the expression profile of the reference system. The number of multiple alternative solutions could be due to the linear dependency of the rows in matrix  $X$  since linearly dependent expression profiles for the same genes across different experiments provide redundant information. Since:

$$X = -(C - I)^{-1}A, \quad (28)$$

the linear dependence on the expression profiles will be due to either linear dependence in the set of perturbed biophysical parameters (i.e. columns of matrix  $A$ ), or due to the specific properties of the matrix  $(C - I)^{-1}$ . The elements of matrix  $A$  were generated randomly, therefore there is a very low probability we would expect to get linearly dependent columns in this matrix that will give  $X$  with linearly dependent columns. An examination of the data generated indicated that this was hardly ever the case. Therefore, the linear dependence is due to the structure of the  $(C - I)^{-1}$  matrix.

Matrix  $C$  has a specific structure in that all its elements were either  $-0.5$ ,  $0$  or  $0.5$ . Among the possibilities are, a specific

**Table 1.** Biophysical perturbation experiments. The results from experiments on three different networks.  $n$ , number of genes in the network;  $k$ , maximum number of regulatory inputs per gene; and  $\eta$ , number of parameters perturbed in each experiment

$n$	$k$	$\eta$	Number of experiments											
			2	3	5	7	10	15	20	25	30	35	40	
5	3	3	1, 1, 1 (0.115)	1, 1, 1 (0.427)	1, 1, 1 (0.8288)	US (0.954)	US (0.994)	US (1)	US (1)	US (1)	US (1)	US (1)	US (1)	
10	3	3	16, 1, 1 (0)	1, 2, 1 (0)	4, 2, 1 (0.046)	1, 1, 2 (0.258)	2, 1, 1 (0.629)	US (0.915)	US (0.982)	US (0.996)	US (0.999)	US (1)	US (1)	
15	5	5	$>10^6$ , 64, $>10^2$ (0)	$>10^2$ , 1, $>10^2$ (0)	2, 1, 4 (0.022)	1, 1, 16 (0.2)	1, 1, 1 (0.599)	1, 1, 1 (0.917)	US (0.985)	US (0.997)	US (1)	US (1)	US (1)	
20	5	5	$>10^7$ , $>10^6$ , 64 (0)	48, 64, 32 (0)	$>10^2$ , 32, 16 ( $10^{-5}$ )	4, 1, 16 (0.0098)	1, 16, 4 (0.1642)	1, 1, 1 (0.6379)	US (0.8865)	US (0.9675)	US (0.9909)	US (0.9975)	US (0.9993)	
25	5	5	$>10^6$ (0)	$>10^3$ , $>10^4$ , $>10^3$ (0)	8, 48, 64 ( $10^{-10}$ )	4, $>10^3$ , 48 ( $10^{-5}$ )	8, 8, 96 (0.0147)	1, $>10^2$ , 2 (0.2729)	US (0.6439)	US (0.8567)	US (0.9463)	US (0.9804)	US (0.9929)	
30	5	5	CNS (0)	$>10^5$ (0)	$>10^6$ , $>10^5$ , $>10^6$ (0)	$>10^3$ , $>10^2$ , $>10^3$ ( $10^{-5}$ )	$>10^2$ , 8, 8 ( $10^{-4}$ )	1, 2, 32 (0.0616)	US (0.3349)	US (0.6377)	US (0.8278)	US (0.9229)	US (0.9664)	
40	5	10	CNS (0)	CNS (0)	$>10^6$ ( $10^{-10}$ )	$>10^3$ , $10^3$ , 2 ( $10^{-5}$ )	10, 16, 1 (0.0231)	4, 1, 1 (0.3873)	1, 1, 1 (0.7728)	US (0.9307)	US (0.9800)	US (0.9943)	US (0.9984)	
50	5	10	CNS (0)	CNS (0)	CNS ( $10^{-21}$ )	CNS ( $10^{-9}$ )	16, $>10^2$ , 2 ( $10^{-4}$ )	4, 2, 8 (0.0676)	1, 8, 1 (0.3893)	1, 1, 1 (0.7215)	US (0.8891)	US (0.9583)	US (0.9846)	
60	5	10	CNS (0)	CNS (0)	CNS (0)	CNS ( $10^{-17}$ )	CNS ( $10^{-8}$ )	$>10^2$ , 64, $>10^3$ (0.0033)	1, 1, 1 (0.1044)	1, 24, 2 (0.3927)	1, 1, 1 (0.6735)	US (0.8446)	US (0.9301)	

Numbers in the parentheses denote the expected probability for uniquely identifying the correct structure [calculated using Equation (31)]. CNS, cannot solve (solver was unable to find a solution) and US, unique solution.

subset of elements in two rows of  $(C - I)^{-1}$  are the same or are related by a factor of 0.5. For example, consider the following five-gene network,

$$C = \begin{pmatrix} -0.5 & 0.5 & -0.5 & 0 & 0 \\ 0 & 0 & 0 & -0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 & -0.5 \\ 0.5 & 0.5 & 0 & 0 & -0.5 \\ 0.5 & 0 & 0 & 0 & 0.5 \end{pmatrix}. \quad (29)$$

This particular structure is characteristic of systems where the transcription of more than one gene is regulated by the same regulatory proteins and with the same type of regulation (such genes would have similar promoter organization and binding sites). In prokaryotic systems, this would apply to all genes in an operon. For example, in the above example, the transcription of both genes 4 and 5 is induced by genes 1 and 2, and it is repressed by gene 5.

The matrix  $(C - I)^{-1}$  for this network will then be:

$$(C - I)^{-1} = \begin{pmatrix} -0.5556 & -0.2222 & 0.5556 & 0.1111 & 0.2222 \\ -0.2222 & -0.8889 & 0.2222 & 0.4444 & -1.1111 \\ -0.5556 & -0.2222 & -1.4444 & 0.1111 & -1.7778 \\ -0.1111 & -0.4444 & 0.1111 & -0.7778 & 0.4444 \\ -0.5556 & -0.2222 & 0.5556 & 0.1111 & -1.7778 \end{pmatrix}. \quad (30)$$

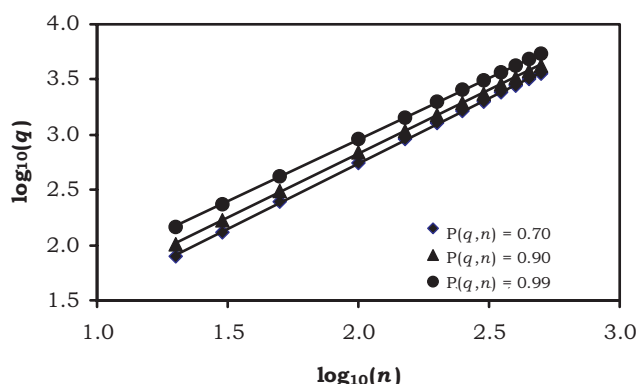
In two experiments where the transcription rate constants corresponding to genes 1 and 2 only are perturbed, the expression profiles of genes 1, 3 and 5 will be exactly the same while changes in genes 2 and 4 will be related by a factor of 0.5. Examination of the data suggested that this was the case when multiple solutions were found when the number of experiments was close to half the number of genes. Therefore, we would be able to identify a unique regulatory mechanism if either every kinetic constant was perturbed at least once or if matrix  $C$  did not have ‘special properties’.

The probability of randomly generating a  $C$  matrix having the special property appears hard to compute. However, we can find an expression for the probability that after  $l$  experiments every kinetic constant has been perturbed at least once, which is equivalent to the probability of identifying uniquely the correct regulatory structure from  $l$  experiments. We derived a recursive expression for a lower bound on the probability that every kinetic constant was perturbed at least once, when  $l$  experiments are performed on an  $n$ -gene network with  $\eta$  kinetic constants being perturbed in each experiment (see Appendix 5 for derivation of the formula):

$$P(q, f) = P(q - 1, f - 1) \frac{n - f + 1}{n} + P(q - 1, f) \frac{f}{n}, \quad 1 \leq f \leq n, \quad q \geq 1 \quad (31)$$

$$P(1, f) = 0, \quad 2 \leq f \leq n$$

$$P(q, 1) = \left(\frac{1}{n}\right)^{(q-1)}, \quad q \geq 1,$$



**Fig. 2.** The number of experiments,  $l$ , that guarantees identification of a unique and correct regulatory structure [within a probability threshold,  $P(q, n)$ ].  $n$ , number of genes in the network.  $q = l \times \eta$ ; where  $l$  represents number of experiments and;  $\eta$  is the number of biophysical parameters perturbed in each experiment.

where  $P(q, n)$  is the probability of interest and  $q$  is the product,  $l \times \eta$ .

This result suggests that, for the same probability of identifying uniquely the correct regulatory structure, the number of experiments can be reduced if we increase the number of biophysical parameters perturbed in each experiment. The simulation results also suggest that as this probability is increasing, the number of alternative structures identified is decreasing rapidly. In some cases even though there was a very low or a zero probability, a unique network was identified since the estimated probability [Equation (31)] provides a lower bound and the structure of the corresponding matrix  $C$  did not require that every kinetic constant to be perturbed at least once.

Figure 2 illustrates the dependence of the number of experiments, for a fixed number of perturbed biophysical parameters in each experiment, required to guarantee within a probability threshold the developed method will identify uniquely the correct regulatory structure. It appears that the number of experiments is exponentially increasing with the number of the genes in the network. Through linear regression of the data in Figure 2, we estimated that:

$$P(l \times \eta, n) = \lambda \cdot n^\kappa \quad (32)$$

with  $\lambda$  ranging between 5.32 [for  $P(l \times \eta, n) = 0.70$ ] and 2.34 [for  $P(l \times \eta, n) = 0.99$ ], and  $\kappa$  ranging between 1.12 [for  $P(l \times \eta, n) = 0.70$ ] and 1.18 [for  $P(l \times \eta, n) = 0.99$ ].

Even with a new mathematical result on the problem formulation we were able to decompose the problem of finding the pathways into  $n$  sub-problems, and reduce the complexity of working with  $n^2$  variables to problems with only  $n$  integer variables, all the present state-of-the-art integer programming solvers used were not able to handle problems for large number genes ( $>100$ ) (Table 2). In addition, even for problems

**Table 2.** Knock-out experiments.  $n$ , number of genes in the network;  $k$ , maximum number of regulatory inputs per gene; and US, unique solution

$n$	$k$	Number of experiments						
		3	5	7	10	15	20	25
5	3	4, 1, 81	2, 1, 1	1, 1, 1	US	US	US	US
10	3	4, 1, 32	$>10^2$ , 12, $>10^2$	9, 81, 3	1, 1, 1	US	US	US
15	5	$>10^4$ , $>10^6$ , $>10^6$	$>10^5$ , 18, $>10^3$	32, $>10^2$ , 24	54, 1, 1	1, 1, 1	US	US
20	5	$>10^7$ , $10^8$ , $>10^9$	$>10^7$ , $>10^6$ , 24	72, $>10^2$ , $>10^3$	32, 2, 54	6, 72, 4	1, 1, 1	US
25	5	$>10^7$ , $>10^8$	$>10^8$ , $>10^5$	$>10^3$ , $>10^6$ , $>10^8$	$>10^3$ , $>10^2$ , $>10^5$	1, 2, 4	1, $>10^3$ , 2	1, 1, 1

of this size the performance of these solvers was inconsistent; in particular, the performance was very sensitive to the optimality tolerance set for the solution. When a loose solution tolerance ( $10^{-2}$ ) was set, the solver reported solutions as optimal which were clearly not optimal, and it became increasingly difficult to find optimal solutions of the problems when a tighter tolerance ( $10^{-6}$ ). This phenomenon of increased difficulty in solving integer programs when accurate solutions are desired is well known. These results suggest the requirement for the development of algorithms, tailored to the structure and the complexity of the gene regulatory networks and to the associated mathematical representations.

We have also performed single-gene knock-out experiments (see Methods section). We observed that in the majority of the simulation experiments the correct regulatory structure could be uniquely identified only when the number of experiments is equal to or greater than the number of genes in the network (Table 2). This is due to the similarity of these studies to the previous case studies in the regime where one biophysical parameter was perturbed in each experiment. In order to guarantee that every biophysical parameter will be perturbed at least once, in a single-gene knock-out experiment, we should perform as many knock-out experiments as the number of genes in the network. Depending on the regulatory structure of the system, i.e. the structure of the matrix  $C$ , and the genes we knock-out, it would be possible to require a smaller number of experiments (see cases for  $n$  equal to 5 and 10 and  $l$  equal to 3 in Table 2).

#### 4 DISCUSSION

Model-based identification methods allow the identification of the regulatory interactions in genetic networks based on mRNA expression data. Since every mathematical model is an approximation of the real biological system, methods for the identification of regulatory interactions should allow the identification of multiple possible regulatory interactions that are consistent with the experimental data. The framework we presented here is based on S-systems modeling of gene expression and on advanced optimization methodologies that allow the identification of multiple regulatory interactions that are consistent with the experimental data. Our studies focused

on the number and type of experiments required for uniquely identifying the genetic regulatory network of the system under study.

Our studies lead to two very important conclusions for the analysis of DNA array data. We found that if there are two or more genes whose expression profiles across the different experiments are linearly dependent, any model-based identification method will fail to identify a unique regulatory structure consistent with the data. This could be the case for co-regulated genes, as discussed above, or for genes that appear to be co-regulated within the accuracy of the experimental methods. This finding suggests that in this case we should cluster the co-regulated genes into a single ‘gene’ and study the interactions of this new ‘gene’ in the network. The identified regulatory connections could then be due to interactions of single or multiple members of the cluster, and closer investigation of the elements in the cluster might help in refining the regulatory network.

The number of experiments required for uniquely identifying the correct regulatory structure was found to depend both on the number of genes and the number of simultaneously perturbed biophysical parameters in each experiment. We found that the number of the experiments times the number of the biophysical parameters perturbed in each experiment should be nearly half of the number of the genes in the system in order to identify a unique regulatory structure. This relationship is an overestimate with respect to required number of experiments, and fewer experiments might be required depending on the structure of the underlying regulatory structure. This result also suggests that we could decrease the required number of experiments if in each experiment we increase the number of the biophysical parameters we perturb.

Our studies involved a number of simplified assumptions and we have also assumed that the origin and the magnitude of the perturbation in each experiment are known. This might be the case in experiments where specific genes are targeted for overexpression or downregulation, or in the case of knock-out experiments. In the later case, identification of the regulatory structure based on single-gene knock-out experiments might require the knock-out of every gene. We are currently investigating the number of experiments required if we consider two-gene and three-gene knock outs.

The use of a mixed-integer optimization method allow us to identify every possible regulatory structure that is consistent with the experimental data. This capability will be extremely useful for the analysis of real experimental data that are subject to experimental noise, and it provides experimentalists with alternative hypotheses and molecular models that could be evaluated based on the additional knowledge about the system and additional experimental studies. In addition, the method allows us to formulate constraints on the structure of the system that can enforce known regulatory interactions or do not allow genes to act as regulators if it is known that they are not regulatory proteins from experiments such as ChIp-on-chip [Chromatin immunoprecipitation on a chip (DNA array)] (Iyer *et al.*, 2001; Molle *et al.*, 2003; Ren *et al.*, 2000).

When the method identified multiple structures, we also observed that certain identified regulatory interactions were common in every alternative structure and they were present in the system used to generate the experimental data. This suggests that the regulatory interactions that are common in the identified alternative structures are highly probable to be present in the original system. We are currently investigating the generality and the origins of this result, as well as if this result holds when the data are corrupted with noise.

## REFERENCES

- Akutsu,T., Miyano,S. and Kuhara,S. (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.
- Bower,J.M. and Bolouri,H. (2001) *Computational Modeling of Genetic and Biochemical Networks*, *Computational Molecular Biology Series*. MIT Press, Cambridge, MA.
- Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, 29–40.
- Czyzyk,J., Mesnier,M.P. and More,J.J. (1998) The NEOS Server. *IEEE Comput. Sci. Eng.*, **5**, 68–75.
- Dasika,M., Gupta,A., Maranas,C.D. and Varner,J.D. (2004) A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. *Pac Symp. Biocomput.*, 474–485.
- de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.
- Dezso,Z., Oltvai,Z.N. and Barabasi,A.L. (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. *Genome Res.*, **13**, 2450–2454.
- D’Haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.
- D’Haeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.*, 41–52.
- Di Bernardo,D., Gardner,T.S. and Collins,J.J. (2004) Robust identification of large genetic networks. *Pac. Symp. Biocomput.*, 486–497.
- Dolan,E. (2001) The NEOS Server 4.0 Administrative Guide. *Technical Memorandum ANL/MCS-TM 250*, Mathematics and Computer Science Division, Argonne National Laboratory.
- Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.
- Fletcher,R. and Lefflyer,S. (1998) Numerical experience with lower bounds for MIQP branch-and-bound. *SIAM J. Comput.*, **8**, 604–616.
- Friedman,N., Linial,M., Nachman,I. and Pe’er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Gardner,T.S., di Bernardo,D., Lorenz,D. and Collins,J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.
- Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
- Gropp,W. and More,J. (1997) Optimization environments and the NEOS server. In Buhmann, M.D. and Iserles, A. (eds), *Approximation Theory and Optimization: Tributes to M.J.D. Powell*. Cambridge University Press, Cambridge, New York, pp. 167–182.
- Hartemink,A.J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2002) Bayesian methods for elucidating genetic regulatory networks. *IEEE Intell Syst.*, **17**, 37–43.
- Hlavacek,W.S. and Savageau,M.A. (1992) 2 rules that predict molecular mechanisms for the regulation of inducible gene-expression. *FASEB J.*, **6**, A72–A72.
- Hlavacek,W.S. and Savageau,M.A. (1998) Method for determining natural design principles of biological control circuits. *J. Intell. Fuzzy Syst.*, **6**, 147–160.
- Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.
- Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
- Ideker,T.E., Thorsson,V. and Karp,R.M. (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.*, 305–316.
- Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,J., Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
- Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
- Liang,S., Fuhrman,S. and Somogyi,R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 18–29.
- Lin,X., Floudas,C.A., Wang,Y. and Broach,J.R. (2003) Theoretical and computational studies of the glucose signaling pathways in yeast using global gene expression data. *Biotechnol. Bioeng.*, **84**, 864–886.
- Maki,Y., Tominaga,D., Okamoto,M., Watanabe,S. and Eguchi,Y. (2001) Development of a system for the inference of large scale genetic networks. *Pac. Symp. Biocomput.*, 446–458.

- Mendes,P., Sha,W. and Ye,K. (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19**(Suppl. 2), II122–II129.
- Michaud,D.J., Marsh,A.G. and Dhurjati,P.S. (2003). eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics*, **19**, 1140–1146.
- Molle,V., Fujita,M., Jensen,S.T., Eichenberger,P., Gonzalez-Pastor,J.E., Liu,J.S. and Losick,R. (2003) The Spo0A regulon of *Bacillus subtilis*. *Mol Microbiol.*, **50**, 1683–1701.
- Moriyama,T., Shinohara,A., Takeda,M., Maruyama,O., Goto,T., Miyano,S. and Kuhara,S. (1999) A system to find genetic networks using weighted network model. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 186–195.
- Nemhauser,G.L. and Wolsey,L. (1988) *Integer and Combinatorial Optimization*. John Wiley and Sons.
- Noda,K., Shinohara,A., Takeda,M., Matsumoto,S., Miyano,S. and Kuhara,S. (1998) Finding genetic network from experiments by weighted network model. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 141–150.
- Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E., et al. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
- Savageau,M.A. (1969a) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.*, **25**, 365–369.
- Savageau,M.A. (1969b) Biochemical systems analysis. II. The steady-state solutions for an  $n$ -pool system using a power-law approximation. *J. Theor. Biol.*, **25**, 370–379.
- Savageau,M.A. (1976) *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Addison-Wesley, Reading, MA.
- Savageau,M.A. (1983) Models of gene-function—general-methods of kinetic-analysis and specific ecological correlates. *ACS Symp. Ser.*, **207**, 3–25.
- Savageau,M.A. (1988) Introduction to S-systems and the underlying power-law formalism. *Math. Comput. Modell.*, **11**, 546–551.
- Savageau,M.A. (1991) Biochemical systems theory: operational differences among variant representations and their significance. *J. Theor. Biol.*, **151**, 509–530.
- Savageau,M.A. and Voit,E.O. (1982) Power-law approach to modeling biological-systems. 1. Theory. *J. Ferment. Technol.*, **60**, 221–228.
- Smith,V.A., Jarvis,E.D. and Hartemink,A.J. (2002) Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, **18** (Suppl. 1), S216–S224.
- Somogyi,R., Fuhrman,S., Askenazi,M. and Wuensche,A. (1997) The gene expression matrix: towards the extraction of genetic network architectures. *Nonlinear Anal.-Theory Meth. Appl.*, **30**, 1815–1824.
- Sorlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci., USA*, **98**, 10869–10874.
- Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci., USA*, **96**, 2907–2912.
- Tegner,J., Yeung,M.K., Hasty,J. and Collins,J.J. (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl Acad. Sci. USA*, **100**, 5944–5949.
- Voit,E.O. and Savageau,M.A. (1987) Accuracy of alternative representations for integrated biochemical systems. *Biochemistry*, **26**, 6869–6880.
- Weaver,D.C., Workman,C.T. and Stormo,G.D. (1999) Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput.*, 112–123.
- Wu,F.X., Zhang,F.X. and Kusalik,A.J. (2004) Modeling gene expression from microarray expression data with state-space equations. *Pac. Symp. Biocomput.*, 581–592.
- Yeung,M.K., Tegner,J. and Collins,J.J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl Acad. Sci. USA*, **99**, 6163–6168.

## APPENDIX 1

The analysis presented in this appendix is based on the S-system methodology for the power-law representation proposed by Savageau and co-workers (Savageau, 1969a,b, 1976; Savageau and Voit, 1982; Voit and Savageau, 1987). We will consider the case that the transcription rate of a gene is subject to induction regulation by a protein, P. Under the common assumption of fast reversible binding of the protein on the regulatory region of the DNA the transcription rate is described by the following hyperbolic kinetic expression:

$$V_{tr} = k_{tr} \cdot B \cdot \frac{P}{K + P}, \quad (A1.1)$$

where  $k_{tr}$  is the transcription rate constant,  $B$  is the concentration of the protein-binding sites and  $K$  is the dissociation constant of the protein–DNA complex.

Taylor expansion of the kinetic expression in Equation (A1.1) with respect to protein concentration around a reference state can be written as follows:

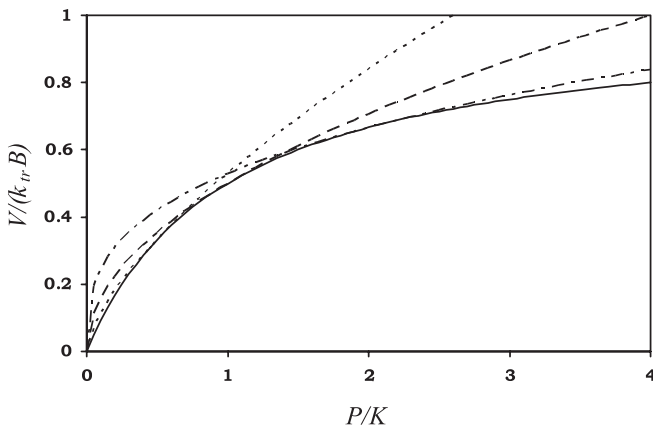
$$V_{tr} = V_{tr,o} + \left. \frac{\partial V_{tr}}{\partial P} \right|_o \cdot (P - P_o) + \text{higher order terms}, \quad (A1.2)$$

where subscript ‘o’ denotes the corresponding quantity evaluated at the reference state conditions.

Ignoring higher order terms and scaling around the reference state leads to the following expression:

$$\begin{aligned} \frac{V_{tr} - V_{tr,o}}{V_{tr,o}} &= \frac{P_o}{V_{tr,o}} \cdot \left. \frac{\partial V_{tr}}{\partial P} \right|_o \cdot \frac{(P - P_o)}{P_o} \Rightarrow \frac{V_{tr} - V_{tr,o}}{V_{tr,o}} \\ &= \left. \frac{\partial \ln V_{tr}}{\partial \ln P} \right|_o \cdot \frac{(P - P_o)}{P_o}. \end{aligned} \quad (A1.3)$$

For any logarithmic function of variable,  $y$ , we can write for up to first-order approximation the Taylor series around



**Fig. A1.** Comparison of hyperbolic rate expression (Equation A1.1) with power-law representation derived around different reference states:  $P_o = 0.5 \cdot K$  ( $\alpha = 0.529$ ,  $\varepsilon = 0.667$ ) (dotted line);  $P_o = K$  [ $\alpha = 0.5$ ,  $\varepsilon = 0.5$ ] (dashed line); and  $P_o = 2 \cdot K$  ( $\alpha = 0.529$ ,  $\varepsilon = 0.333$ ) (dash-dotted line).

a reference value,  $y_o$ , as follows:

$$\ln(y) = \ln(y_o) + \frac{y - y_o}{y_o} \Rightarrow \ln\left(\frac{y}{y_o}\right) = \frac{y - y_o}{y_o}, \quad (\text{A1.4})$$

which suggests that Equation (A1.3) can be rewritten as:

$$\ln\left(\frac{V_{tr}}{V_{tr,o}}\right) = \frac{\partial \ln V_{tr}}{\partial \ln P}\bigg|_o \cdot \ln\left(\frac{P}{P_o}\right) \Rightarrow V_{tr}^{PL} = \alpha \cdot P^\varepsilon, \quad (\text{A1.5})$$

where the superscript ‘*PL*’ indicates the power-law kinetics of the transcription rate with:

$$\alpha = \frac{V_{tr,o}}{P_o^\varepsilon} \quad \text{and} \quad \varepsilon = \frac{\partial \ln V_{tr}}{\partial \ln P}\bigg|_o. \quad (\text{A1.6})$$

For the hyperbolic rate expression in Equation (A1.1), the exponent  $\varepsilon$  can be calculated as a function of the dissociation constant,  $K$ , and the reference protein concentration:

$$\varepsilon = \frac{\partial \ln V_{tr}}{\partial \ln P}\bigg|_o = \frac{K}{K + P_o} \quad (\text{A1.7})$$

The power-law expression in Equation (A1.5) can be used to approximate the kinetic expression in Equation (A1.1) after choosing a reference state and calculating the corresponding values for the constant  $\alpha$  and the exponent  $\varepsilon$ . Figure A1 demonstrates the accuracy of the power-law expression derived for three different reference states:  $P_o = 0.5 \cdot K$  which corresponds to  $\alpha = 0.529$  and  $\varepsilon = 0.667$ ;  $P_o = K$  which corresponds to  $\alpha = 0.5$  and  $\varepsilon = 0.5$ ; and  $P_o = 2 \cdot K$  which corresponds to  $\alpha = 0.529$  and  $\varepsilon = 0.333$ . Over a wide range of protein concentration ( $0-4 \cdot K$ ) the power-law representation with  $\alpha = 0.5$  and  $\varepsilon = 0.5$  is the best non-linear approximation with an average relative error  $((V_{tr} - V_{tr}^{PL})/V_{tr})$  equal

to  $-13.34$  and SD  $18.02$ . For the other two cases, the relative errors (SD) are higher:  $-29.28$  ( $21.13$ ) for  $\alpha = 0.529$  and  $\varepsilon = 0.667$ , and  $-14.65$  ( $42.47$ ) for  $\alpha = 0.529$  and  $\varepsilon = 0.333$ .

The above considerations suggest that kinetic expressions of the form:

$$V \propto \frac{P}{K + P} \quad (\text{A1.8})$$

can be best approximated by a power-law representation of the form:

$$V \propto P^{0.5}. \quad (\text{A1.9})$$

Similar analysis could also show that kinetic expression of the form:

$$V \propto \frac{K}{K + P} \quad (\text{A1.10})$$

can be best approximated by a power-law representation of the form:

$$V \propto P^{-0.5} \quad (\text{A1.11})$$

In general, a relatively accurate power-law approximation can be developed for any type of complex non-linear biochemical kinetics (Savageau, 1976; Savageau and Voit, 1982; Voit and Savageau, 1987).

## APPENDIX 2

According to this representation, the gene expression synthesis rate can be written as:

$$V_i^s = \alpha_i \prod_{j=1}^n P_j^{\varepsilon_{ij}}, \quad (\text{A2.1})$$

where  $V_i^s$  is the synthesis rate of gene  $i$ ,  $\alpha_i$  is the synthesis rate constant,  $P_j$  is the product protein of gene  $j$  which regulates the synthesis of gene  $i$  and  $\varepsilon_{ij}$  is a power that quantifies the strength of regulatory interaction of gene  $i$  by gene  $j$ . Similar power-law kinetic expressions can be formulated for the degradation rates of mRNA, and for the synthesis and degradation rates of proteins.

Based on the power-law rate expressions Equation (A2.1) the mass balances for the mRNA of gene  $i$  and its protein product can be written as follows:

$$\frac{dM_i}{dt} = \alpha_i \prod_{j=1}^n P_j^{\varepsilon_{ij}} - \beta_i M_i^{\theta_{ii}}, \quad (\text{A2.2})$$

$$\frac{dP_i}{dt} = \gamma_i M_i^{\sigma_i} - \delta_i P_i^{\xi_i},$$

where  $M_i$  and  $P_i$  are the concentrations of mRNA and protein products, respectively, of gene  $i$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  are rate constants, and  $\varepsilon_{ij}$ ,  $\theta_{ii}$ ,  $\sigma_i$  and  $\xi_i$  are constants that quantify the strength of interactions. The information about the regulatory

architecture around a genetic network lies in the parameter values of the exponents  $\varepsilon_{ij}$ . Similar to the parameters of the linear model, a positive (negative) value for the exponent  $\varepsilon_{ij}$  implies that gene  $i$  is induced (repressed) by the product of gene  $j$ .

This modeling approach captures some of the essential non-linear features of gene expression and it has yielded invaluable insights about the responses and the design principles of genetic networks.

At steady-state Equation (A2.2) can be solved exactly:

$$\begin{aligned} 0 = \alpha_i \prod_{j=1}^n P_j^{\varepsilon_{ij}} - \beta_i M_i^{\theta_{ii}} &\Rightarrow \ln(\alpha_i/\beta_i) + \sum_{j=1}^n \varepsilon_{ij} \ln(P_j) = \theta_{ii} \ln(M_i). \\ 0 = \gamma_i M_i^{\sigma_i} - \delta_i P_i^{\xi_i} &\Rightarrow \ln(\gamma_i/\delta_i) + \sigma_i \ln(M_i) = \xi_i \ln(P_i). \end{aligned} \quad (\text{A2.3})$$

We can further express the protein products as a function of the mRNA and we can rewrite the steady-state mass-balance equation for the mRNA:

$$\phi_i + \sum_{j=1}^n c_{ij} \ln(M_j) = \ln(M_i), \quad (\text{A2.4})$$

where  $\phi$  and  $c_{ij}$  are the following combinations of the original parameters of the system [Equation (A2.3)]:

$$\phi_i = \ln(\alpha_i/\beta_i)/\theta_{ii} + \sum_{j=1}^n \varepsilon_{ij} \cdot \ln(\gamma_i/\delta_i)/(\theta_{ii} \cdot \xi_i). \quad (\text{A2.5})$$

$$c_{ij} = (\varepsilon_{ij} \cdot \sigma_i)/(\theta_{ii} \cdot \xi_i).$$

We will further assume that the exponents  $\sigma_i$  and  $\xi_i$  in the protein mass-balance and the exponent  $\theta_{ii}$  in the mRNA mass-balance are positive. These assumptions stem from the fact that protein synthesis and protein degradation are monotonically increasing functions of the mRNA ( $M_i$ ) and protein ( $P_i$ ) concentrations, respectively, and the mRNA degradation rate is also a monotonically increasing function of the mRNA ( $M_i$ ) concentration. Under this assumption, the new parameters  $c_{ij}$  have the same sign with the parameters  $\varepsilon_{ij}$ , and determine the regulatory architecture in our genetic network. If protein synthesis, and protein and mRNA degradation and are considered to be first-order ( $\sigma_i, \xi_i$  and  $\theta_{ii}$  equal to 1), as is the case in the main studies here,  $c_{ij}$  is exactly equal to  $\varepsilon_{ij}$ . Therefore, a positive (negative) value for the parameter  $c_{ij}$  implies that gene  $i$  is induced (repressed) by the product of gene  $j$ . The challenge is then to estimate the values of  $c_{ij}$  from measurement of mRNA expression.

In Equation (A2.4),  $M_j$  denotes the (absolute) concentration of mRNA. However, from DNA-array experiments the available information is about relative concentrations of mRNA between two states, i.e. ‘ratios’. Manipulation of Equation (A2.4) allows us to use information about relative

concentrations instead of absolute concentration values:

$$\begin{aligned} \left. \begin{aligned} \phi_i + \sum_{j=1}^n c_{ij} \ln(M_{j,I}) &= \ln(M_{i,I}) \\ \phi_i + \sum_{j=1}^n c_{ij} \ln(M_{j,II}) &= \ln(M_{i,II}) \end{aligned} \right\} \\ \Rightarrow \sum_{j=1}^n c_{ij} \ln(M_{j,I}) - \sum_{j=1}^n c_{ij} \ln(M_{j,II}) &= \ln(M_{i,I}) - \ln(M_{i,II}) \\ \Rightarrow \sum_{j=1}^n c_{ij} \ln(M_{j,I}/M_{j,II}) &= \ln(M_{j,I}/M_{j,II}), \quad (\text{A2.6}) \end{aligned}$$

where  $I$  and  $II$  denote the two different states.

### APPENDIX 3

*Proof of the theorem.* Solving Problem 1 (in  $n^2$  dimensions) is equivalent to solving  $n$  problems, each of them in  $n$  dimensions.

The proof is based on two observations. First the objective to minimize the sum of squares of all elements of the  $\mathbf{R}$  matrix [Equation (18)] is the same as minimizing the sum of squares (taken over all rows of  $\mathbf{R}$ ) of elements of each row. Now the  $r$ -th row of  $\mathbf{R}$  would only contain terms from the  $r$ -th row of the  $\mathbf{C}$  matrix. The elements of the  $r$ -th column of the  $\mathbf{C}$  matrix represent how the  $r$ -th gene regulates the other in the system. The constraint that the  $r$ -th gene would not regulate more than  $k$  genes would involve exactly the  $n$  elements of the  $r$ -th row. Hence, we can solve the original problem one gene at a time.

### APPENDIX 4

The following is the general optimization program to determine the  $r$ -th best solution for the  $r$ -th problem (corresponding to the  $i$ -th gene,  $i = 1, 2, \dots, n$ )

$$\min \|W^T \mathbf{y}_i^r + \mathbf{a}_i\|^2 \quad (\text{A4.1})$$

Subject to:

$$y_{i,j}^r = 0.5d_{i,j,1}^r - 0.5d_{i,j,2}^r - 1 \quad \text{for } j = i, \quad (\text{A4.2})$$

$$y_{i,j}^r = 0.5d_{i,j,1}^r - 0.5d_{i,j,2}^r \quad \text{for } j \neq i, \quad (\text{A4.3})$$

$$d_{i,j,1}^r + d_{i,j,2}^r \leq 1 \quad \text{for } j = 1, 2, \dots, n, \quad (\text{A4.4})$$

$$\sum_{j=1}^n (d_{i,j,1}^r + d_{i,j,2}^r) \leq k, \quad (\text{A4.5})$$

$$\begin{aligned} \sum_{j=1}^n (d_{i,j,1}^r d_{i,j,1}^b + d_{i,j,2}^r d_{i,j,2}^b) &\leq (D^b - 1) \\ \text{for } b = 1, 2, \dots, (r-1), &\quad (\text{A4.6}) \end{aligned}$$

$$D^b = \sum_{j=1}^n (d_{i,j,1}^b + d_{i,j,2}^b) \quad \text{for } b = 1, 2, \dots, (r-1), \quad (\text{A4.7})$$

$$d_{i,j,1}^r, d_{i,j,2}^r \in \{0, 1\} \quad \text{for } j = 1, 2, \dots, n, \quad (\text{A4.8})$$

where  $\mathbf{c}_i^t$  is the transpose of the  $i$ -th row of matrix  $\mathbf{C}$ , and  $k$  is the maximum number of regulatory inputs per gene.

For the ‘Biophysical perturbation studies’,  $\mathbf{W} = \mathbf{X}$  is the  $n \times l$  matrix of log-ratios of the changes in the mRNA concentration and  $\mathbf{a}_k$  is the log-ratio of the change in synthesis rate of the  $k$ -th gene in all the  $l$  experiments. For the ‘Knock-out studies’,  $\mathbf{W} = \mathbf{X}$  if the  $i$ -th gene was not removed in any of the  $l$  experiments, else it is the  $\mathbf{X}$  matrix with the  $i$ -th row removed and the  $\mathbf{a}_i$  vector is the zero vector.

## APPENDIX 5

Consider the following situation. A person begins throwing balls at a group of  $n$  buckets. Each ball (from each throw) has an equal probability of falling in any of the  $n$  buckets. The problem is to determine the probability that after  $h$  throws,  $f$  buckets are filled with one or more balls. Define

$E_{h,f}$  : Event of having  $f$  buckets full after  $h$  throws.

$D_{q,i}$  : Event of the  $q$ -th throw landing in any of  $i$  buckets.

Then by the law of total probability,

$$\begin{aligned}
 P(E_{h,f}) &= P(E_{h-1,f-1})P(D_{h,n-f+1}) \\
 &\quad + P(E_{h-1,f})P(D_{h,f}), \quad \text{or} \\
 P(E_{h,f}) &= P(E_{h-1,f-1}) \binom{n-f+1}{n} + P(E_{h-1,f}) \binom{f}{n}.
 \end{aligned}
 \tag{A5.1}$$

Clearly, the probability of having more than one bucket filled with one throw should be zero, i.e.

$$P(E_{1,f}) = 0, \quad f \geq 2. \tag{A5.2}$$

It should also be clear that the probability of having just one bucket filled after  $k$  throws should be given by,

$$P(E_{h,1}) = \left(\frac{1}{n}\right)^{(h-1)}, \quad h \geq 1. \tag{A5.3}$$

There is an analogy of this situation with that of perturbing every one of the  $n$  kinetic constants in the genetic network, at least once. Each experiment in which  $\eta$  kinetic constants are randomly chosen can thought of being similar to making  $\eta$  throws at the group of  $n$  buckets. The only difference is now that all these  $\eta$  throws fall in different buckets. However between experiments, there can be overlap of the kinetic constants chosen. Hence  $l$  experiments can thought of as similar to making  $l \times \eta$  throws. Because of non-overlap within an experiment, the probabilities computed should be a lower bound on the desired values.

## ACKNOWLEDGEMENTS

This work was supported by grant R01 GM065476 from the NIH National Institute of General Medical Sciences. V.H. acknowledges support from DuPont Young Professor Grant.