

A quantitative method for reverse engineering gene networks from microarray experiments using regulatory strengths.

Alberto de la Fuente
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and
State University
1750 Kraft Drive, Suite 1100
Blacksburg, VA 24061-0477
alf@vbi.vt.edu

Paul Brazhnik
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and
State University
1750 Kraft Drive, Suite 1100
Blacksburg, VA 24061-0477
brazhnik@vbi.vt.edu

Pedro Mendes
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and
State University
1750 Kraft Drive, Suite 1100
Blacksburg, VA 24061-0477
mendes@vbi.vt.edu

ABSTRACT

We propose a method for reverse engineering gene regulatory networks from microarray gene expression data. The method is based on metabolic control analysis and requires data from microarray gene expression experiments where the rate of change of a single gene has been perturbed. Gene regulatory interactions are quantified through “regulatory strengths” which are determined from co-responses of messenger RNA to a common perturbation. The application of the method is illustrated by analyzing data produced with computer simulations of gene regulatory networks. Evidence is shown indicating that the method performs well in some cases even when perturbations are as large as two-fold. A global gene regulatory network can be uncovered by this method if applied systematically to all genes in a genome. Alternatively, it can be applied to a subset of genes in which case it identifies a phenomenological gene network that contains, beside direct interactions, contributions from indirect interactions (via the genes that were not considered or complex formation of the gene products). Supplementary materials are available at <http://www.vbi.vt.edu/~mendes/icsb01-supp.html>

1. INTRODUCTION

Functioning of living systems is partially orchestrated, at the molecular level, by selective expression of genes of a genome. Expression of genes is modulated by specific proteins (activators and/or inhibitors) and metabolite effectors, which are gene products themselves. One can thus consider networks of genes in which some genes modulate the expression of others – gene regulatory networks. Uncovering such networks is essential to the understanding how the biological machinery of cells works and consequently to manipulate it to our advantage.

Microarray [41] and DNA chip [33] technologies have enabled high-throughput monitoring of gene expression on a genome-wide scale [e.g. 8, 9, 12, 23, 30, 40, 46, 47] and are being used with increasing frequency [4]. Analysis of microarray gene expression data is a rapidly developing field with a number of different

methods in current use. The majority of these focuses on identifying the function of open reading frames (ORF) by following a logic of “guilt by association”, and are based on grouping genes by their expression pattern [5, 15, 38, 43, 48]. A second class of analyses attempts to model the dynamics of gene expression as observed in time-course experiments with microarrays [7, 14, 24, 45] and is based on fitting a model to (relative) mRNA time courses. A third class of algorithms attempts to reverse engineer networks of relationships between genes. Most have done so in a qualitative way, as for example using Boolean networks where transcription is represented as all-or-none [1, 32]. Methods of analysis for genome-wide gene expression were reviewed recently by D’Haeseleer *et al.* [13].

Gene regulatory networks are high-level conceptual representations of interactions between genes in an organism. They are usually represented as directed graphs in which each gene is connected to a number of other genes whose expression it modulates (see Fig. 1). Gene networks hide a great deal of biochemical detail, such as all proteins and metabolites. Thus gene networks represent behaviors that result from thousands of hidden variables and therefore should not be expected to be predictive for conditions far from those used to build them. Nevertheless, gene networks can be an excellent means of summarizing experimental gene expression results and of helping us reason about complex cellular phenomena.

Metabolic Control Analysis (MCA), derived from the work of Kacser and Burns [27] and Heinrich and Rapoport [18], is a quantitative formalism that allows one to establish the extent to which individual biochemical reactions control cellular variables such as pathway fluxes and concentrations of biochemical compounds. MCA is a sensitivity analysis of steady states providing measures of how parameter perturbations affect the variables. This is, in its most basic form, done via the flux- and concentration-control coefficients, which are ratios of the relative change in flux or concentration to the relative change in the parameter (the perturbation). Control coefficients are systemic properties and depend on the properties of *all* the components of the system, not only on the one they are related to. MCA provides a means of relating these systemic variables to the kinetic properties of each single reaction, which are expressed as

elasticity coefficients [6]. Of particular relevance here is the fact that the matrix of elasticity coefficients can be calculated from a matrix of control coefficients essentially by inversion [39]. The reader is encouraged to consult the extensive literature on MCA for further details [10, 11, 16, 17, 19, 20].

Here we propose a method for reverse engineering gene regulatory networks from steady state relative gene expression data. Our approach falls under the framework of MCA and requires a specific experimental design based on perturbations of transcription rates. We illustrate the method by analyzing “synthetic” data generated with a mathematical model. The applicability of the method is discussed in the context of current microarray technology and a limited knowledge of protein-protein interactions.

2. METHOD

Microarray experiments quantify gene expression levels essentially as a ratio of the abundance of mRNA in response to a stimulus to its abundance in a reference state (determined by the ratio of fluorescence intensity of two fluorofors):

$$FR_i = \frac{F'}{F} = \frac{[mRNA_i]'}{[mRNA_i]}, \quad (1)$$

where F' and F are respectively the fluorescence intensities of the stimulated and reference state, $[mRNA_i]$ is the reference concentration of the message of gene i ($i = 1, \dots, n$; n being the total number of genes analyzed) and $[mRNA_i]'$ is the concentration of the same message in the new steady state reached after the stimulus has been applied. One should bear in mind that artifacts might arise due to details of the technologies (indeed, some have pointed out, e.g. [49]), although these fall outside the scope of our analysis. As with any other, the quality of the results of the proposed method is only as good as the raw data it processes.

Two relatively new concepts in MCA are the *co-control coefficient* [22] and the *regulatory strength* [28]. A co-control coefficient expresses how two variables change when a single parameter is perturbed, while a regulatory strength expresses how much of a perturbation in a variable is propagated to another through a particular pathway. Both these concepts take a central rôle in our method and are applied specifically to mRNA concentrations. The co-control coefficient, defined for the mRNA concentrations of two genes i and j when a rate of transcription v_m is perturbed, is defined as:

$${}^{v_m}O_j^i = \frac{\partial[mRNA_i]/[mRNA_i]}{\partial[mRNA_j]/[mRNA_j]}. \quad (2)$$

We propose to use a special type of regulatory strengths, those that quantify direct effects, as a quantitative description for gene regulatory networks. The regulatory strength of $mRNA_i$ on $mRNA_j$ through direct action over the transcription of $mRNA_j$ is written as:

$${}^{v_j}R_i^j = \mathcal{E}_{mRNA_i}^{v_j} \cdot C_{v_j}^{mRNA_j} = \frac{\partial v_j / v_j}{\partial [mRNA_i] / [mRNA_i]} \cdot \frac{\partial [mRNA_j] / [mRNA_j]}{\partial v_j / v_j}, \quad (3)$$

where $\mathcal{E}_{mRNA_i}^{v_j}$ is the elasticity of v_j by $mRNA_i$ and $C_{v_j}^{mRNA_j}$ is the concentration-control coefficient of the transcription reaction j on $mRNA_j$. These coefficients quantify how much change a perturbation in a selected variable (mRNA concentration) induces on another through a particular path of regulation. A negative regulatory strength indicates an inhibitory influence, a positive one activation, and zero represents no direct influence. Fig. 1 depicts an example gene network as a directed graph and also as a matrix of direct-effect regulatory strengths.

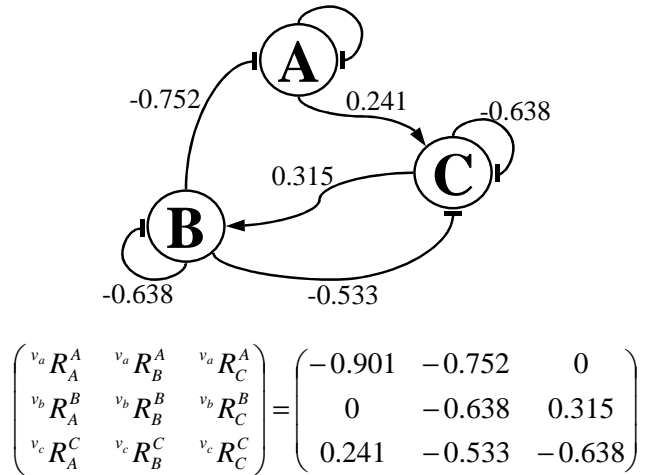


Figure 1: A model gene regulatory network of three genes. Arrows represent activation interactions while lines with blunt ends inhibitory interactions. The numbers next to the lines are the values determined for the corresponding regulatory strengths.

While Eq. 3 defines regulatory strengths, it is not a convenient method for their calculation, as it is hard to determine both the elasticity and control coefficients from experiments. Fortunately, an alternative method exists in which regulatory strengths can be calculated from the co-control coefficients. The latter are easier to determine [21, 22] because the magnitude of the perturbation does not need to be known in this case. A matrix of regulatory coefficients (\mathbf{R}) is equal to the inverse matrix of co-control coefficients (\mathbf{O}) as shown by Hofmeyr and Cornish-Bowden [21]. The task is then to measure co-control coefficients from microarray experiments and to calculate the regulatory strengths from them.

All coefficients in MCA (including the co-control coefficients) are by definition the result of infinitesimal calculations, but in real experiments one can only make finite changes. It is thus useful to reformulate the approximation explicitly using finite changes because this is how experimental data will be processed. All calculations are based on scaled differentials but the

measurements are of finite concentration changes. We will make use of a central finite differences approximation, $\Delta C/C$, to the scaled derivative $\partial C/C$:

$$\frac{\Delta C}{C} = \frac{(C - C^0)}{(C + C^0)/2}, \quad (4)$$

where C^0 is the reference concentration and C is the concentration after perturbation. In microarray experiments, however, one does not determine absolute values but rather a ratio of fluorescence intensities that is equivalent to the ratio of concentrations ($FR=C/C^0$, C^0 being the reference concentration). The use of central finite differences is important, as it is free from the bias that left or right finite differences would introduce. This is especially important when the perturbations are large (as we will illustrate later on). Eq. 4 can then be expressed in terms of the fluorescence ratio FR by dividing denominator and numerator by the same factor C^0 :

$$\frac{\Delta C}{C} = \frac{2\left(\frac{C}{C^0} - 1\right)}{\frac{C}{C^0} + 1} = \frac{2(FR - 1)}{FR + 1}. \quad (5)$$

Using this result to replace infinitesimal changes by finite changes in Eq. 2, an approximation of the co-control coefficients in terms of fluorescence ratios can be written as:

$$v_m \tilde{O}_j^i = \frac{\Delta mRNA_i / mRNA_i}{\Delta mRNA_j / mRNA_j} = \frac{(FR_i - 1)(FR_j + 1)}{(FR_j - 1)(FR_i + 1)}, \quad (6)$$

$v_m \tilde{O}_j^i$ is thus a measurable quantity obtainable directly from microarray data. For a single perturbation experiment on the rate of transcription of gene m (v_m) one can calculate n^2 co-responses $v_m \tilde{O}_j^i$. These are organized in groups of n , each of which makes up a column of a co-control matrix. Since there are n co-control matrices (one for each gene), n experiments are needed (with a single transcription rate perturbed in each), to complete the n co-control matrices:

$$\tilde{O}_i = \begin{bmatrix} v_1 \tilde{O}_i^1 & \dots & v_n \tilde{O}_i^1 \\ \dots & \dots & \dots \\ v_1 \tilde{O}_i^n & \dots & v_n \tilde{O}_i^n \end{bmatrix} \quad (7)$$

Here \tilde{O}_i is the co-control matrix of gene i . Inverting each of these matrices results in a regulatory strength matrix:

$$\tilde{\mathbf{R}}_i = \tilde{O}_i^{-1} \quad (8)$$

To reconstruct the gene regulatory network we then use row i from each matrix $\tilde{\mathbf{R}}_i$ (with $i = 1, \dots, n$) that corresponds to the direct-effect regulatory strengths.

To summarize, the proposed method consists of the following steps:

1. Select the set of n genes for which the regulatory network will be investigated (all genes of a genome or a subset thereof).
2. Perturb the rate of transcription of one single gene.
3. Measure gene expression ratios between the new steady state reached after the perturbation and the reference state, using microarray or DNA chip technology.
4. Use fluorescence ratios FR and Eq. 6 to calculate n^2 co-control coefficients, completing one row of each of the n co-control matrices.
5. Carry out steps 2-4 until all transcription rates in the initially selected set have been perturbed and gradually fill in the co-control matrices.
6. Invert the n co-control matrices to obtain n regulatory strength matrices $\tilde{\mathbf{R}}_i$.
7. From the n matrices $\tilde{\mathbf{R}}_i$ use row i of each to reconstruct the gene regulatory network.

The next section illustrates the application of the method using synthetic data from model gene regulatory networks.

3. RESULTS

To illustrate the proposed method, we will apply it to data produced by artificial gene regulatory networks (computer models). These models were defined and run with the biochemical kinetics simulator Gepasi [35, 36] (available at <http://www.gepasi.org>). We intentionally used nonlinear kinetics for transcription to show that the inherent nonlinearity of the system does not invalidate this linear method. In each case we changed the rate of transcription of each transcription step (as specified above) and calculated the resulting new steady state. Implying that fluorescence signals in microarray experiments are proportional to mRNA concentrations, we further used the ratio of mRNA concentration in perturbed state over the mRNA concentration in the reference state as FR . Proceeding with the remaining steps of the method, we obtained regulatory strengths that allowed us to re-construct the corresponding network graphs. One advantage in using a computer model is the ability to judge how well the method performed. The models used represent several scenarios, including one in which there are hidden variables and another where there are non-additive effects (complex formation between different gene products).

3.1 All players are known

We consider a small regulatory network of three genes and simulate on the computer the experiments described above. The

model is described by the network depicted in Figure 1 and the following system of ordinary differential equations:

$$\begin{aligned}
 \frac{d[A]}{dt} &= \frac{V_a}{1 + \frac{[B]}{K_{iB}}} - k_a[A] \\
 \frac{d[B]}{dt} &= \frac{V_b}{1 + \frac{K_{aC}}{[C]}} - k_b[B] \\
 \frac{d[C]}{dt} &= \frac{V_c}{\left(1 + \frac{[B]}{K_{iB}'}\right)\left(1 + \frac{K_{aA}}{[A]}\right)} - k_c[C]
 \end{aligned} \tag{9}$$

Here [A], [B] and [C] are concentrations of the mRNA species; V_a , V_b and V_c are basal rates of transcription; K_{iB} and K_{iB}' are inhibition constants; K_{aC} and K_{aA} are activation constants; k_a , k_b , and k_c are first-order degradation constants.

We first calculate [A], [B] and [C] for a reference steady state setting all parameter values arbitrarily to unity, except for K_{iB} , K_{iB}' , K_{aA} , and K_{aC} which were set to 0.1.

Then each rate of transcription was modified (one at a time) by changing a value of the basal rate and new steady state concentrations were calculated. Since the analysis was originally based on infinitesimal changes, we started by making “small” changes of 10% on transcription rates. We followed the method described in the previous section and obtained nine direct-effect regulatory strengths which we combined in a new matrix (Fig.1). From this matrix we draw the corresponding wiring diagram and assign quantitative measures to each interaction (the regulatory strengths) as depicted in Fig. 1.

Keeping in mind that small changes in transcription rates are difficult to achieve in practice and/or their effects are even harder to measure, we explored the performance of the method with larger perturbations (under-expression by 50% and over-expression by 200%). The values of the regulatory strengths obtained with these larger perturbations are compared to the theoretical values (calculated with Eq. 3 using elasticity and control coefficients obtained with the simulation software) in Table 1. Table 1 clearly shows that the error due to the finite differences approximation in our method is relatively small for a small perturbation (1.1x) but grows with larger perturbations (0.5x and 2x). Nevertheless, even with the larger perturbations the absolute error is less than 0.075 (17%), which we believe is well below the measurement noise, and thus perfectly acceptable.

Table 1 – Effect of the size of perturbations on the estimation of regulatory strengths. Theoretical values of regulatory strengths for the model of Fig. 1A were calculated using Eq. 3 and the simulation’s values of elasticity and control coefficients. “Experimental” values were calculated by applying different perturbations on rate of transcription (last three columns) and following the method described in the text.

	Theoretical value	1.1x perturbation	0.5x perturbation	2x perturbation
$v_a R_A^A$	-0.901	-0.903	-0.879	-0.916
$v_a R_B^A$	-0.752	-0.757	-0.711	-0.787
$v_a R_C^A$	0	0.001	-0.005	0.004
$v_b R_A^B$	0	0.001	-0.006	0.005
$v_b R_B^B$	-0.638	-0.639	-0.635	-0.646
$v_b R_C^B$	0.315	0.311	0.348	0.288
$v_c R_A^C$	0.241	0.236	0.279	0.213
$v_c R_B^C$	-0.533	-0.544	-0.444	-0.607
$v_c R_C^C$	-0.638	-0.639	-0.628	-0.651

3.2 Hidden variables

In cases where one cannot measure the expression of all the genes, when it is not easy to perturb the transcription rate of some genes, or when it is not convenient to analyze the full network, the

method must be applied only to a subset of the genes. In those cases only the rates of transcription of “available” genes and their mRNA concentrations are measured. Nevertheless, a larger network is responsible for the observations and so one can no longer be sure if the interactions detected by the method are really direct or are a result of the hidden variables. To explore such a

scenario we constructed the five-gene network shown in Fig. 2A (the kinetics were similar to the previous example and are available as supplementary information at <http://www.vbi.vt.edu/~mendes/icsb01-supp.html>).

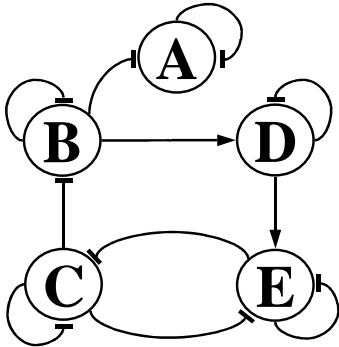


Figure 2A: A model gene regulatory network of five genes. The full mathematical model is supplied as supplementary material at <http://www.vbi.vt.edu/~mendes/icsb01-supp.html>.

Then the transcription rates of genes C, D and E were perturbed by 10% over-expression and their corresponding relative mRNA responses were “measured”. We calculated all the direct-effect regulatory strengths for those three genes and used them to reconstruct the network shown in Fig. 2B.

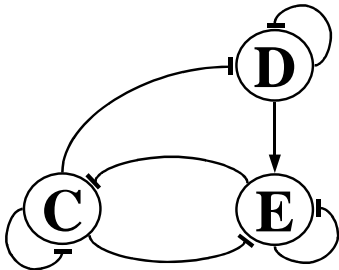


Figure 2B: the network reverse engineered with the method proposed when only genes C, D and E were perturbed and observed.

Comparing the networks of Fig. 2 (the original and the one reconstructed by our method), it is clear that all the direct interactions on the original network between genes C, D and E were recovered from the regulatory strengths. In addition, there is a new arrow from C to D in Fig. 2B that does not exist in the original network (Fig. 2A).

In the original system, C influences B and B influences D, but because B was not included in the analysis these two interactions collapse into a single one which is, of course, only apparent. What this reveals is that if only a subset of genes is considered in the analysis, then the interactions identified with this method are not necessarily direct but can also include indirect effects (through hidden variables). Incidentally this is the same as if one considers that in most cases it is not the mRNAs that interact with transcription but rather their protein products. Because proteins are not represented explicitly here, they can be thought of being

hidden variables and their action is included in the arrows of the gene regulatory network.

3.3 Non-additive effects

Our model of gene networks is a purely additive one, which stems from our use of metabolic control analysis, a formalism based on a first-order Taylor approximation [18]. However, it is not hard to conceive (or find examples) where the regulation of the expression of a gene depends on a combination of several other genes. For example, the products of gene A and gene B may have to bind in order to be capable of activating gene C. Because these interactions may be frequent, it is relevant to see how the method performs with such networks. For that we have created the gene network depicted in Fig. 3A.

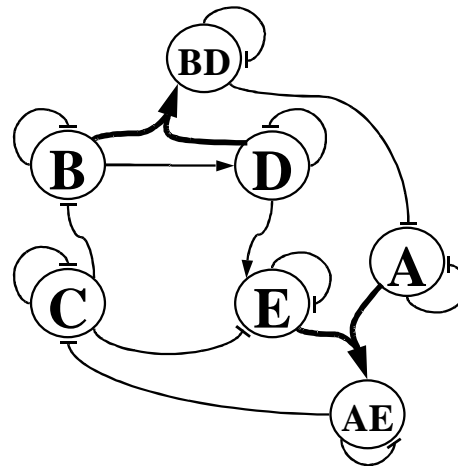


Figure 3A: A gene network where some of the regulation occurs through binary complexes. The full mathematical model is supplied as supplementary material at <http://www.vbi.vt.edu/~mendes/icsb01-supp.html>.

The main feature of this network is the formation of two complexes, AE and BD that are regulators of other genes. We performed 10% over-expression perturbations on all the transcription rates and observed the response on the expression levels of A, B, C, D and E (levels of AE and BD were not monitored). The calculated regulatory strengths describe a gene regulatory network presented in Fig. 3B. The network shows interactions from B to A, from D to A, from A on C and from E on C – these correspond to the interactions that complexes AE and BD have on target genes in the original model, but are recovered by our approach as an additive effect of individual genes. This situation is similar to the previous case of hidden variables: as we did not consider the complexes as separate individual entities, we recovered their effect as additive effects of the constituents of the complexes. This results in extra arrows leading directly from the complex components to the target genes. There are also additional interactions of A on E and E on A, and B on D and D on B which are all negative. We interpret this in the following way: in the original network, increasing A, for example, will tend to increase AE and therefore decrease E. In summary,

although the method does not account directly for complex formation, the effect of such interactions does not get lost in our treatment, but is reflected in the resulting networks as coming from the individual constituents of the complexes.

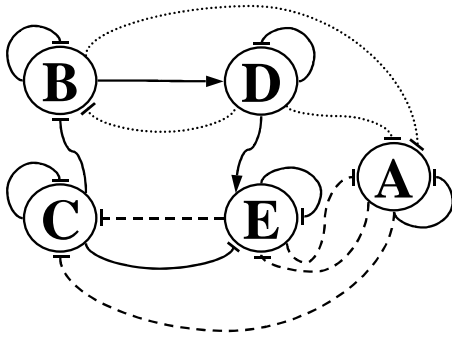


Figure 3B: The gene network reverse engineered with the proposed method. The non-additive regulatory interactions that occur through the complexes in the original system are lost but an equivalent set of additive interactions appears in their place.

4. DISCUSSION

We have proposed a method for inferring the gene regulatory networks from gene expression data. This method requires making perturbations on single rates of transcription, one at a time, and measuring the levels of mRNA of all genes after each perturbation. The method is thus very suitable to be applied together with the increasingly popular microarray or DNA chip technologies. Considering that these technologies can only reliably measure ratios of mRNA concentrations, we derived the method based on those ratios, not on absolute mRNA levels. This does not preclude using the method with quantitative RT-PCR or Northern blot data.

An important issue in applying this method is the need to carry out specific single-gene perturbation experiments. Often gene expression is monitored in response to perturbations that affect several genes or to mutants where one or several genes have been deleted from the genome. Here we argue for perturbations of each single rate of transcription, which implies that the total number of experiments required is equal to the number of genes under consideration.

Many existing methods to reverse engineer gene networks [2, 25, 32, 42] rely on a Boolean representation of gene interactions [29, 44]. In such representations, genes are considered to be either completely “on” or completely “off”. In our models we make no such assumptions and consider expression levels as continuous variables. Other methods approach gene networks by representing them as linearly additive models [14, 24]. This is a better representation as it allows for transcription rates to take on continuous values. Our method falls into this category. Nevertheless, it is of notice that gene regulatory networks obtained by our method (and other linear additive ones) are snapshots of a state of the system and therefore may have limited extrapolation power to states other than the one studied. Some other representations have been proposed [3, 31, 34, 37, 45] that

use non-linear rate functions to represent the dynamics of mRNA concentrations, but they have the disadvantage of including large numbers of parameters to be fit and so require much more data [37, 45] than this method. We believe that first obtaining a representation based on regulatory strengths should precede the application of such non-linear regression models.

One advantage of our method is that it is quantitative. Not only does it find which interactions are present but also how strong they are, as reflected in the regulatory strengths. But to be truly quantitative, extremely small perturbations (infinitesimal) should be applied on the rates of transcription. This is impossible to achieve and with current state-of-the-art of laboratory technique we are limited to considerably large perturbations. Measurement of small changes in gene expression levels is even more problematic. It is thus important to consider that the method will be usually applied with large perturbations and is useful to know how it would perform in such circumstances. We have tested for a simple model the limits of applying large perturbations on transcription rates. We have found out that even with rates 2-fold larger (as could be obtained by duplicating the number of gene copies in the genome), we obtain rather good estimates of the regulatory strengths. Our conclusion is that even for large changes we may be able to reconstruct gene regulatory networks with fairly good resolution. In this case the accuracy is somewhat more limited but the measurement noise is still larger than the finite approximation itself.

We argue for the perturbation of transcription rates rather than perturbations on mRNA levels. But how can one perturb rates of transcription? One way is to transform cells using constructs that include a different promoter sensitive to some chemical compound foreign to that cell (such as constitutive artificial promoters [26] inducible by IPTG). Fortunately, for our purposes, this transformation does not have to be performed on the chromosomes: it suffices to add the gene of interest with the appropriate promoter in a plasmid. On the other hand, gene knockouts are not appropriate for this analysis: when a gene is completely removed, the resulting network is no longer the same as the original one and it becomes impossible to quantify the regulatory strength.

The theory behind our analysis is based on a linearization of the mRNA dynamics around the reference state. Unless the kinetics of mRNA concentrations is linear (which we do not believe to be frequent and indeed it is not the case in our model, see Eqs. 9), the description of the regulatory network with regulatory strengths is only good for that specific cell state. For example, if one were to compare a bacterial culture growing in a rich medium with the same culture growing in a limited medium, one would proceed first by perturbing the culture in rich medium and estimating its corresponding gene network, and then by perturbing the culture in the limited medium and estimating its gene network. The comparison would be at the level of the two gene networks and the focus would be on how the regulatory strengths have changed from one state to the other – a quantitative comparison.

Considering that a network is only good for one state of the organism and that the network recovered with this method may be different from the original one (e.g. when there are hidden variables, Fig. 2, or non-additive effects, Fig. 3), we conclude that gene regulatory networks are phenomenological. To find the “true” mechanistic gene regulatory dynamics one would have to

find the phenomenological networks of many different states of the organism and then infer the non-linear kinetic functions that could accurately describe the behavior of the network in a wide range of conditions.

It is possible that some interactions between genes, though physically present, are not detectable in a particular state by our method. This may happen, for instance, due to saturation in the transcription rates. For example, if A is an activator of B but the transcription rate of B is saturated in A, an increase in the level of A has no measurable effect on B, and consequently the regulatory strength is zero. Hence, in this particular state there is no effective regulation of A on B. That A is able to affect B would become evident only when analyzing another state of the system where the transcription of B was no longer saturated in A, e.g. in different environmental conditions. Each cell type of a multi-cellular organism is a specific state of the genetic network [29] therefore each cell type will have a different *effective* genetic regulatory network with a characteristic set of regulatory strengths between genes.

We showed that the method proposed is capable of providing a lower-resolution gene network when only a subset of genes of a genome is being considered. The ability of the method to deal with an incomplete set of components in a network is of great importance. Not all organisms have yet been fully sequenced and even for those that have, it is very likely that not all genes have yet been identified. In addition, we will probably not be able to apply perturbations on all rates of transcription and some mRNAs may be below accurate detection. Even though our method is not able to indicate that some genes are missing in the analysis, it provides *causal* information about all genes present, reflecting effects of genes that are not considered in the analysis.

One point worth stressing is that the analysis we propose does not rely on the size of the perturbation applied, only that it should be small. This is because the method is based on the ratio of changes in pairs of mRNA species rather than on the effect of the perturbation on each single one. This fact is most convenient because one cannot accurately predict the magnitude of changes in transcription rates when a new gene copy is added (either to a chromosome or in a plasmid). If one were to use the traditional metabolic control analysis rather than co-control analysis, the approach would require accurate measurements of the size of the perturbation. In such a case one would measure concentration-control coefficients and then use these to calculate elasticity coefficients, which describe the effect of one mRNA on a transcription rate (rather than one mRNA on another, as regulatory strengths). We have also performed this alternative analysis on the model of Fig. 1 and were indeed able to describe the correct gene network because we assumed accurate knowledge of the size of perturbation (see supplementary information at <http://www.vbi.vt.edu/~mendes/icsb01-supp.html>). From the experimentalist point of view, however, it is simpler to measure co-control coefficients. Note that both analyses require exactly the same perturbation experiments to be performed.

We argue for carrying out specific perturbation experiments in order to infer the regulatory structure of gene networks of an organism or cell type. Knowledge of such networks is a crucial step towards the understanding the functions of the genes and indeed the cell as a whole. The experimental effort needed for these experiments is perhaps comparable to genome sequencing

projects. To be feasible to apply this method to whole-genomes in a systematic way will require high-throughput technologies. The leap in understanding of cell function that comes from knowing gene regulatory networks is comparable to the leap in knowledge that occurred from single-gene to whole-genome sequences and will help in deciphering the function of many genes that are currently of unknown function. In the meantime it will still be very informative to apply the proposed method to smaller groups of genes.

5. ACKNOWLEDGEMENTS

We thank Karen Schlauch and Bruno Sobral for helpful discussions and comments on the manuscript. Financial support from the Commonwealth of Virginia is gratefully acknowledged.

6. REFERENCES

1. Akutsu, T., Kuhara, S., Maruyama, O. and Miyano, S. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. *Genome Informatics*, 9. 151-160.
2. Akutsu, T., Miyano, S. and Kuhara, S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computational Biology*, 7 (3-4). 331-343.
3. Ando, S. and Iba, H. Quantitative modeling of gene regulatory network: identifying the network by means of genetic algorithm. *Genome Informatics*, 11. 278-280.
4. Blohm, D.H. and Guiseppi-Elie, A. New developments in microarray technology. *Current Opinion in Biotechnology*, 12 (1). 41-47.
5. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences USA*, 97 (1). 262-267.
6. Burns, J.A., Cornish-Bowden, A., Groen, A.K., Heinrich, R., Kacser, H., Porteous, J.W., Rapoport, S.M., Rapoport, T.A., Stucki, J.W., Tager, J.M., Wanders, R.J.A. and Westerhoff, H.V. Control analysis of metabolic systems. *Trends in Biochemical Sciences*, 10 (1). 16.
7. Chen, T., He, H.L. and Church, G.M. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, 4. 29-40.
8. Cho, R.J., Campbell, M.J., Winzler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2 (1). 65-73.
9. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. The transcriptional

- program of sporulation in budding yeast. *Science*, 282 (5389). 699-705.
10. Cornish-Bowden, A. Metabolic control analysis in biotechnology and medicine. *Nature Biotechnology*, 17. 641-643.
 11. Cornish-Bowden, A. and Cárdenas, M.L. *Control of metabolic processes*. Plenum Press, New York and London, 1990.
 12. DeRisi, J.L., Iyer, V.R. and Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278 (5338). 680-686.
 13. D'Haeseleer, P., Liang, S. and Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16 (8). 707-726.
 14. D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*, 4. 41-52.
 15. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences U.S.A.*, 95 (25). 14863-14868.
 16. Fell, D.A. Metabolic control analysis - a survey of its theoretical and experimental development. *Biochem. J.*, 286. 313-330.
 17. Fell, D.A. *Understanding the Control of Metabolism*. Portland Press, London, 1996.
 18. Heinrich, R. and Rapoport, T.A. A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector. *Eur. J. Biochem.*, 42. 97-105.
 19. Heinrich, R. and Schuster, S. *The regulation of cellular systems*. Chapman & Hall, New York, 1996.
 20. Hofmeyr, J.H. Metabolic regulation: a control analytic perspective. *Journal of Bioenergetics and Biomembranes*, 27 (5). 479-490.
 21. Hofmeyr, J.H. and Cornish-Bowden, A. Co-response analysis: a new experimental strategy for metabolic control analysis. *Journal of Theoretical Biology*, 182 (3). 371-380.
 22. Hofmeyr, J.H.S., Cornish-Bowden, A. and Rohwer, J.M. Taking enzyme kinetics out of control - putting control into regulation. *European Journal of Biochemistry*, 212 (3). 833-837.
 23. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95 (5). 717-728.
 24. Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V. and Banavar, J.R. Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences USA*, 98 (4). 1693-1698.
 25. Ideker, T.E., Thorsson, V. and Karp, R.M. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, 5. 305-316.
 26. Jensen, P.R. and Hammer, K. Artificial promoters for metabolic optimization. *Biotechnology and Bioengineering*, 58 (2-3). 191-195.
 27. Kacser, H. and Burns, J.A. The control of flux. *Symp. Soc. Exp. Biol.*, 27. 65-104.
 28. Kahn, D. and Westerhoff, H.V. The Regulatory strength: how to be precise about regulation and homeostasis. *Acta Biotheoretica*, 41. 85-96.
 29. Kauffman, S. Homeostasis and differentiation in random genetic control networks. *Nature*, 224 (215). 177-178.
 30. Kehoe, D.M., Villand, P. and Somerville, S. DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends Plant Sci*, 4 (1). 38-41.
 31. Kyoda, K.M., Morohashi, M., Onami, S. and Kitano, H. A gene network inference method from continuous-value gene expression data of wild-type and mutants. *Genome Informatics*, 11. 196-204.
 32. Liang, S., Fuhrman, S. and Somogyi, R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 3. 18-29.
 33. Lockhart, D.J., Dong, H.L., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C.W., Kobayashi, M., Horton, H. and Brown, E.L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14 (13). 1675-1680.
 34. Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S. and Eguchi, Y. Development of a system for the inference of large scale genetic networks. *Pacific Symposium on Biocomputing*, 6. 446-458.
 35. Mendes, P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends in Biochemical Sciences*, 22. 361-363.
 36. Mendes, P. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Computer Applications in the Biosciences*, 9 (5). 563-571.
 37. Mendes, P. Modeling large scale biological systems from functional genomic data: parameter estimation. in Kitano, H. ed. *Foundations of Systems Biology*, MIT Press, Cambridge, MA, 2001, in press.
 38. Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R., Cluster analysis and data visualization of large-scale gene expression data. in *Pacific Symposium on Biocomputing*, (1998), 42-53.
 39. Reder, C. Metabolic control theory. A structural approach. *Journal of Theoretical Biology*, 135 (2). 175-201.
 40. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn,

- M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24 (3). 227-235.
41. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 (5235). 467-470.
42. Somogyi, R., Fuhrman, S., Askenazi, M. and Wuensche, A. The gene expression matrix: Towards the extraction of genetic network architectures. *Nonlinear Analysis - Theory Methods & Applications*, 30 (3). 1815-1824.
43. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA*, 96 (6). 2907-2912.
44. Thomas, R. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42 (3). 563-585.
45. Wahde, M. and Hertz, J. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, 55 (1-3). 129-136.
46. Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A. and LaRossa, R.A. High-density microarray-mediated gene expression profiling of *Escherichia coli*. *Journal of Bacteriology*, 183 (2). 545-556.
47. Wodicka, L., Dong, H.L., Mittmann, M., Ho, M.H. and Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnology*, 15 (13). 1359-1367.
48. Zhao, L.P., Prentice, R. and Breeden, L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Sciences USA*, 98 (10). 5631-5636.
49. Zhou, Y.-X., Kalocsai, P., Chen, J.-Y. and Shams, S. Information processing issues and solutions associated with microarray technology. in Schena, M. ed. *Microarray biochip technology*, Eaton Publishing, Natick, MA, 2000, 167-200.