Linking the genes: inferring quantitative gene networks from microarray data

Alberto de la Fuente, Paul Brazhnik and Pedro Mendes

Modern microarray technology is capable of providing data about the expression of thousands of genes, and even of whole genomes. An important question is how this technology can be used most effectively to unravel the workings of cellular machinery. Here, we propose a method to infer genetic networks on the basis of data from appropriately designed microarray experiments. In addition to identifying the genes that affect a specific other gene directly, this method also estimates the strength of such effects. We will discuss both the experimental setup and the theoretical background.

The functioning of biological systems is orchestrated by selective expression of genes. Levels of gene products change during development and differentiation, and in response to external perturbations. Activity of genes is regulated by proteins and metabolites, which are produced by proteins. But proteins are also gene products, thus genes can influence each other (induce or repress) through a chain of proteins and metabolites. At the genetic level, it is thus legitimate, and indeed common, to consider gene–gene interactions, and these lead to the concept of the 'gene network'. A gene network is a graphical (Fig. 1) or numerical (Figs 2,3) representation of causal relationships between activities of genes. Gene networks provide a large-scale, coarse-grained view of an organism's physiological state at the genetic level. Uncovering the structure of such networks would help us to understand how cells work and how we can manipulate them to our advantage.

Several approaches have been proposed for inferring gene networks from experimental data [1]. A popular one assumes that genes with similar expression are functionally related to each other, and it is usually put into practice using clustering algorithms. However, this fails to reveal causality, does not quantify effects, and is not itself part of the experimental design. By contrast, we propose a method based on a well-defined experimental design intended to quantify how much the expression of one gene influences the expression of itself and others.

Experimental design

The method proposed here consists of a set of experiments in each of which the rate of expression of a single gene is perturbed by a finite amount (e.g. through the use of antisense RNA, or by engineering promoter sequences [2]). This differs from published microarray experiments where very drastic changes are made [3–9], such as knocking out genes, or perturbing a group of genes simultaneously. The experiments start with the collection of mRNA from a reference steady state. Then, a finite perturbation is applied to a single gene and, once the system is in a new steady state, the gene expression levels are measured against the reference level. These measurements of gene expression are best carried out using microarrays to assess as many genes as possible, but quantitative reverse-transcriptase polymerase chain reaction (qRT–PCR) would also be appropriate. Further perturbations are applied in a systematic way to all other genes, and their effects measured. The underlying gene network is then recovered directly from relative fluorescence levels using the algorithm described below.

Although perturbing the rate of transcription of one gene at a time could be costly and difficult, we believe that this strategy will lead to quantitative information about the gene network, which will have far-reaching consequences and is thus a worthwhile goal. When it is not possible to perturb all the genes, one should at least include all those that are already suspected of taking part in the phenomenon of interest, and perhaps others that have been associated with these (e.g. through clustering of data from other gene expression experiments).

Theoretical background

Our approach adopts the framework of Metabolic Control Analysis (MCA) [10–12]. MCA quantifies, by control coefficients, how sensitive certain cellular state variables (e.g. fluxes and metabolite concentrations) are to parameters affecting biochemical reactions (e.g. rate constants). Control coefficients are systemic properties and so depend on all cellular parameters, not only on those that affect the variable in question directly. Furthermore, MCA relates the control coefficients to the kinetic properties of biochemical reactions, which are properties of the single components (and usually referred to as 'local'). It is important to the present work that MCA also provides a recipe to deduce local kinetic properties from the global behavior of the system. This is exactly what we propose to exploit: co-control coefficients for pairs of mRNAs are first determined from microarray experiments (reflecting global system properties) and then regulatory strengths, quantifying the direct interactions between genes, are calculated. Co-control coefficients are ratios between control coefficients and show how two system variables (different mRNA concentrations, in this case) respond to a common perturbation [13,14] (a change in transcription rate of one gene). Note that the two references cited [13,14] use the term 'co-response coefficient', but we follow a more recent document by the same authors that uses the term 'co-control coefficient' when the parameter perturbed is a rate (http://www.sun.ac.za/biochem/mcanom.html). Regulatory strengths quantify the

Alberto de la Fuente
Paul Brazhnik
Pedro Mendes*
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, 1880 Pratt Drive, Blacksburg, VA 24061, USA.
*e-mail: mendes@vt.edu

 0168-9525/02/$ – see front matter PII: S0168-9525(02)02692-6

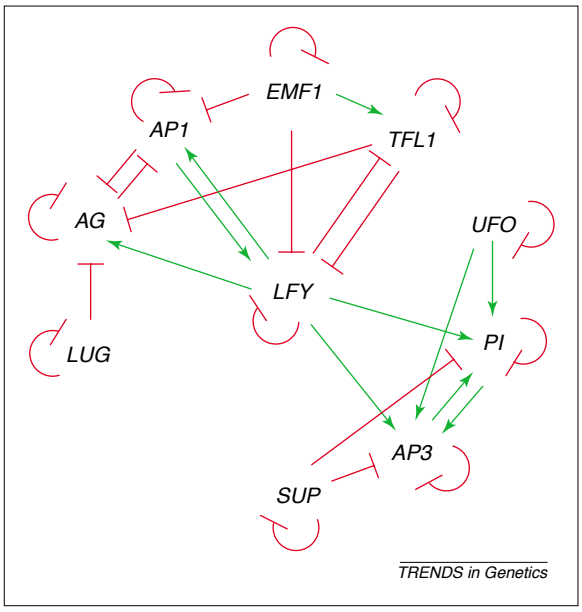

Fig. 1. Graphical representation of the gene network controlling flower morphogenesis in *Arabidopsis thaliana* proposed by Mendoza *et al.* [23]. Direct effects are shown with green for activation, and red for inhibition. The *Arabidopsis* genes are: *LUG*, *Leunig*; *AG*, *Agamous*; *AP1*, *Apetala 1*; *EMF1*, *Embryonic flower 1*; *TFL1*, *Terminal flower 1*; *LFY*, *Leafy*; *SUP*, *Superman*; *AP3*, *Apetala 3*; *PI*, *Pistillata*; *UFO*, *Unusual floral organs*.

Rd =

LUG	AG	AP1	E1MF	T1FL	LFY	SUP	A3P	PI	UFO	
–1	0	0	0	0	0	0	0	0	0	LUG
–0.579	–1.14	–0.184	–0.002	–0.112	0.114	0	0	0	0	AG
–0.009	–0.894	–1.14	–0.109	–0.002	0.124	0	0	0	0	AP1
0	0	0	–1	0	0	0	0	0	0	EMF1
0	0	0	0.094	–1.09	–0.973	0	0	0	0	TFL1
–0.002	–0.005	0.053	–0.103	–0.107	–1.09	0	0	0	0	LFY
0	0	0	0	0	0	–1	0	0	0	SUP
0	0	0	–0.001	–0.001	0.135	–0.119	–1.04	0.192	0.109	AP3
0	0	0	–0.001	–0.001	0.135	–0.119	0.192	–1.04	0.109	PI
0	0	0	0	0	0	0	0	0	–1	UFO

TRENDS in Genetics

Fig. 2. Result of the method applied to the in silico experiments described in the main text. Assuming a base line of 0.01, this matrix represents the gene network of Fig. 1.

fractional changes that a system variable suffers as a consequence of the change in another system variable [15]. Readers are encouraged to consult the extensive literature on MCA [16–21] for further details, as the description below is limited to the essential facts.

Connection of co-control coefficients with regulatory strengths

Co-control coefficients express the concomitant changes in steady-state values of two independent biochemical variables when a single rate is perturbed. In particular, the co-control coefficient ${}^{v_m}O_j^i$ characterizes how the concentrations of $mRNA_i$ and $mRNA_j$ (i.e. $[mRNA_i]$ and $[mRNA_j]$, respectively) change following a perturbation of the transcription rate v_m of a third $mRNA_m$; the coefficients are defined for all of their values including $i = j = m$. MCA is based on infinitesimal changes (perturbations), but as this is impossible to effect in practice, we will hereafter always refer to their finite counterparts:

$$ {}^{v_m}O_j^i = \frac{\Delta[mRNA_i]/[mRNA_i]}{\Delta[mRNA_j]/[mRNA_j]} \qquad \text{[Eqn 1]} $$

The rates considered here are those of mRNA synthesis and degradation, and are aggregated per gene, which is justified by the specific stoichiometry of genetic networks (as the reactions affecting one mRNA do not affect other mRNAs). In each perturbation experiment, j, one can calculate n co-control coefficients ${}^jO_j^i$, with i iterating over all n genes measured. After carrying out experiments for all genes of interest, a subset of co-control coefficients, ${}^jO_j^i$ (calculated by dividing a change in the concentration of $mRNA_i$ by the change in the concentration of $mRNA_j$ that had its rate of change perturbed, $m = j$ in Eqn 1), is organized in a matrix $\mathbf{O} = [{}^jO_j^i]$. This matrix is then inverted to yield a matrix of regulatory strengths, $\mathbf{Rd} = [{}^jR_i^j]$ [13,14]. **Rd** specifies the gene network by direct gene–gene effects [22]; its elements quantify the influence of gene i on gene j through the synthesis or degradation rate of gene j. For further details, see supplementary information at http://www.vbi.vt.edu/~mendes/tig02.html.

Determining co-control coefficients from microarray data

Microarray experiments usually result in ratios of mRNA concentrations in a perturbed state, $[mRNA]$, to their concentrations in a reference state, $[mRNA]^0$, or more precisely, a ratio of fluorescence intensities, FR, that is equivalent to the ratio of concentrations ($FR = [mRNA]/[mRNA]^0$). Such relative measures, as opposed to absolute concentrations, are often seen as an inconvenience. But the proposed method takes advantage of this, because the relative change of the concentration $\Delta[mRNA]/[mRNA]^0$, needed to calculate co-control coefficients, can be directly expressed by fluorescence ratios:

$$ \frac{\Delta mRNA}{mRNA^0} = \frac{mRNA - mRNA^0}{mRNA^0} = FR - 1 \qquad \text{[Eqn 2]} $$

This enables co-control coefficients also to be expressed directly from fluorescence ratios, using a central finite difference approximation to derivatives:

$$ {}^{v_m}O_j^i \approx \frac{(FR_i - 1)(FR_j + 1)}{(FR_j - 1)(FR_i + 1)} \qquad \text{[Eqn 3]} $$

Experimental data can thus be used directly without requiring the laborious calibrations that would be needed to obtain absolute concentrations. Box 1 summarizes how the method should be applied in practice.

An example using *in silico* experiments

Because no data of the kind needed are currently available, we demonstrate the method on simulated experiments with a model gene network. We use the gene network proposed by Mendoza *et al.* [23] to control flower morphogenesis in *Arabidopsis thaliana* (Fig. 1). It is irrelevant for our purposes whether this network is indeed correct or what the molecular details behind it might be. For this illustration, one should assume the model network is the real system. We encapsulate the gene network in a model that follows principles of biochemical kinetics, and is formulated in terms of ordinary differential equations describing the rate of change of the mRNA concentrations. Actual parameter values are not important here, provided that they guarantee the relationships shown in Fig. 1.

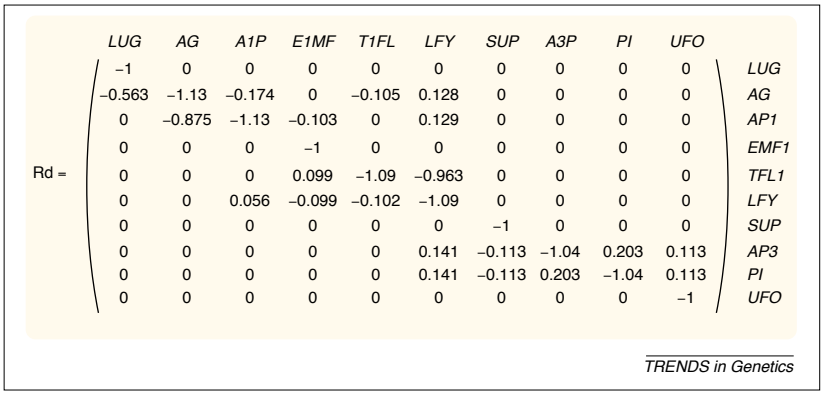

Fig. 3. The matrix of regulatory strengths, calculated by the definition of these coefficients [15]. This matrix represents the gene network of Fig. 1 exactly.

The simulations were carried out with the software GEPASI [24–26] and result in concentration ratios, similar to the fluorescence ratios of real measurements. Additional information about this model can be found at http://www.vbi.vt.edu/~mendes/tig02.html.

The *in silico* experiments performed followed the method described in the previous sections by applying 10% perturbations on the rates of transcription of each gene. Figure 2 depicts the matrix of direct regulatory strengths that is the result of this exercise. By assuming that absolute values of regulatory strengths below 0.01 are below the noise level, and are probably zero, one can reproduce exactly the diagram in Fig. 1 from the matrix of Fig. 2. This demonstrates that the method can recover a gene network from observations of relative mRNA concentrations. The finite approximation used in the method could be a source of error. To assess how well this approximation (e.g. Equation 1) compares with reality, Fig. 3 depicts the solution calculated through the definition of regulatory strength [15]. The estimate obtained by the *in silico* experiment is indeed rather good when compared with the more precise solution of Fig. 3, indicating that 10% perturbations could well be appropriate.

Discussion

It is our thesis that regulatory strengths are good quantitative measures of gene–gene interactions, and they can be determined directly from relative gene expression levels, as obtained in microarray experiments. We present theoretical guidelines (based on MCA) and design for microarray experiments that will enable investigators to infer genetic networks. Because there are as yet no published experiments that conform to such design, the method was illustrated using *in silico* experiments.

Several methods have been proposed to reverse engineer gene networks [27–29], relying on Boolean representations and consider genes completely 'on' or completely 'off' [30,31]. Other methods, including the one presented here, make no such assumptions and represent gene networks as linearly additive models [1,22,32] that consider expression levels as continuous variables. A few other approaches [33–37] use non-linear rate functions to represent the dynamics of mRNA concentrations, but these need to fit larger numbers of parameters and so require a much larger number of experiments [34,36] than the present method. We believe obtaining a representation based on regulatory strengths should precede the application of such non-linear regression models. Our method not only finds which interactions are present, but also their sign and how strong they are. However, because the method is based on a linear approximation, gene networks inferred by it are relevant only to a particular physiological state.

The theory behind this method is formulated on the basis of infinitesimal calculus, and thus is only exact for infinitesimal perturbations. However, real perturbations are always finite, and this introduces errors in our estimations. Therefore, the smaller the perturbations can be made in the experiment, the more reliable will be the final results. In practice, small perturbations are hard to perform and corresponding small responses can easily be missed in the noise. In the *in silico* example, we applied 10% perturbations, but were still able to infer the network correctly. Together with additional *in silico* experiments performed with different networks (data not shown), this indicates that the method can be effective in realistic experimental conditions.

Presently, the signal-to-noise ratio of microarrays is too low to measure the effect of small responses in gene expression. However, the technology is constantly improving, and it is only a matter of time before this approach becomes more feasible (or perhaps a superior technology will appear in the meantime). This contribution will increase the incentives for further technological improvements. Our method requires genome-scale experimental effort, in the form of gene expression rate manipulations, which are currently laborious in practice. This is comparable to the situation with whole-genome sequencing some 15 years ago, when the automated sequencing was not in place to carry out the Human Genome Project, however that was quickly developed thereafter. We expect that, considering the increasing trend of laboratory robotics, and the demonstrated usefulness of this method, high-throughput means of carrying out gene manipulations will become feasible in the not-too-distant future.

This approach requires perturbations to be applied to each gene independently. This is similar to the approach of Wagner [22], except that our method is concerned with the magnitude of changes, whereas Wagner's relies only on identifying which genes changed. Furthermore, that

Box 1. The method summarized

(1) Allow a system of *n* genes to reach a reference steady state and use it in all iterations of step 3 as the reference state.
(2) Perturb the rate of transcription of a single gene and allow the system to reach a new steady state.
(3) Measure the mRNA concentrations of this new state relative to the mRNA concentrations from step 1 using microarrays.
(4) Repeat steps 2–3, until all transcription rates have been perturbed.
(5) Use the fluorescence ratios, *FR*, determined in the experiments above, and calculate the co-control coefficients (Eqn 3), filling the co-control matrix.
(6) Invert the co-control matrix to obtain the regulatory strength matrix, which quantitatively represents the gene network.

approach is incapable of dealing with cyclic relations and cannot distinguish networks that belong to a certain class (it identifies the simplest of that class), whereas our approach has no such limitations.

Another application of MCA to functional genomics is the FANCY method (standing for 'functional analysis of co-responses in yeast') [38]. That method uses the co-response of metabolites to gene knockouts to uncover the functions of genes, and it has recently been demonstrated with great success [39,40]. Although this approach has the use of co-responses in common with ours, the objectives of each are quite different and the similarity is indeed only superficial.

We use the term 'direct' for interactions that run between two genes without being mediated by a third one. In reality, the interactions between any two genes run necessarily through the action of proteins and/or metabolites and are thus never really direct (although the effect of a gene on itself, corresponding to mRNA degradation, is indeed a direct effect). The proteome and metabolome are not considered in the present study, because we focused solely on the genetic level. In general, direct interactions between the genes indicate interactions that run through other molecules that are not considered (metabolites, proteins and unobserved mRNAs). The global gene network of an organism would be uncovered if perturbations were applied to all genes in a genome. However, if only a subset of the genes is manipulated, the method is still capable of identifying a gene network, even though indirect interactions running through the genes that were omitted will appear as direct. We argue that networks obtained by examining subsets of a genome are useful representations of the underlying gene regulatory structure. This is particularly relevant in the sense that not all genes might be known. If gene A influences the expression of gene B through the action of gene C, but gene C is not known, the method will find a regulatory effect of A on B. This is similar to the common situation in MCA where usually only certain parts of metabolism are analyzed, yet it is well known [10] that all other sections of metabolism also contribute to control.

In summary, we propose that a combination of a particular experimental design and the use of microarrays can be used to infer quantitatively the direct interactions between the genes manipulated in the experiment. This contrasts with qualitative analyses that ignore how the experiments were carried out, such as many unsupervised learning methods, of which clustering is perhaps the most popular.

Acknowledgements
We thank the National Science Foundation (grant BES-0120306) and the Commonwealth of Virginia for financial support. We are grateful to Karen Schlauch and Bruno Sobral for stimulating discussions, Tiffany Trent for critical reading of the manuscript, and an anonymous reviewer for suggesting a simplification in the matrix algebra involved.

References

1 D'Haeseleer, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726
2 Snoep, J.L. *et al.* (2002) DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. *Eur. J. Biochem.* 269, 1662–1669
3 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
4 Wodicka, L. *et al.* (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15, 1359–1367
5 Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705
6 Holstege, F.C. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728
7 Kehoe, D.M. *et al.* (1999) DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends Plant Sci.* 4, 38–41
8 Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24, 227–235
9 Wei, Y. *et al.* (2001) High-density microarray-mediated gene expression profiling of *Escherichia coli*. *J. Bacteriol.* 183, 545–556
10 Kacser, H. and Burns, J.A. (1973) The control of flux. *Symp. Soc. Exp. Biol.* 27, 65–104
11 Heinrich, R. and Rapoport, T.A. (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* 42, 89–95
12 Fell, D.A. (1996) *Understanding the Control of Metabolism*, Portland Press
13 Hofmeyr, J.H.S. *et al.* (1993) Taking enzyme kinetics out of control – putting control into regulation. *Eur. J. Biochem.* 212, 833–837
14 Hofmeyr, J.H. and Cornish-Bowden, A. (1996) Co-response analysis: a new experimental strategy for metabolic control analysis. *J. Theor. Biol.* 182, 371–380
15 Kahn, D. and Westerhoff, H.V. (1993) The regulatory strength: how to be precise about regulation and homeostasis. *Acta Biotheor.* 41, 85–96
16 Cornish-Bowden, A. and Cárdenas, M.L. (1990) *Control of Metabolic Processes*, Plenum Press
17 Fell, D.A. (1992) Metabolic control analysis – a survey of its theoretical and experimental development. *Biochem. J.* 286, 313–330
18 Hofmeyr, J.H. (1995) Metabolic regulation: a control analytic perspective. *J. Bioenerg. Biomembr.* 27, 479–490
19 Fell, D.A. (1996) *Understanding the Control of Metabolism*, Portland Press
20 Heinrich, R. and Schuster, S. (1996) *The Regulation of Cellular Systems*, Chapman & Hall
21 Cornish-Bowden, A. (1999) Metabolic control analysis in biotechnology and medicine. *Nat. Biotechnol.* 17, 641–643
22 Wagner, A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics* 17, 1183–1197
23 Mendoza, L. *et al.* (1999) Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics* 15, 593–606
24 Mendes, P. (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.* 9, 563–571
25 Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with GEPASI 3. *Trends Biochem. Sci.* 22, 361–363
26 Mendes, P. and Kell, D. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869–883
27 Liang, S. *et al.* (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 3, 18–29
28 Akutsu, T. *et al.* (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.* 7, 331–343
29 Ideker, T.E. *et al.* (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.* 5, 305–316
30 Kauffman, S. (1969) Homeostasis and differentiation in random genetic control networks. *Nature* 224, 177–178
31 Thomas, R. (1973) Boolean formalization of genetic control circuits. *J. Theor. Biol.* 42, 563–585
32 Holter, N.S. *et al.* (2001) Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* 98, 1693–1698
33 Ando, S. and Iba, H. (2000) Quantitative modeling of gene regulatory network: identifying the network by means of genetic algorithm. *Genome Informatics* 11, 278–280
34 Wahde, M. and Hertz, J. (2000) Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems* 55, 129–136
35 Maki, Y. *et al.* (2001) Development of a system for the inference of large scale genetic networks. *Pac. Symp. Biocomput.* 6, 446–458
36 Mendes, P. (2001) Modeling large scale biological systems from functional genomic data: parameter estimation. In *Foundations of Systems Biology* (Kitano, H., ed.), pp. 163–186, MIT Press
37 Onami, S. *et al.* (2001) The DBRF method for inferring a gene network from large-scale steady-state gene expression data. In *Foundations of Systems Biology* (Kitano, H., ed.), pp. 59–75, MIT Press
38 Teusink, B. *et al.* (1998) Metabolic control analysis as a tool in the elucidation of the function of novel genes. *Methods Microbiol.* 26, 297–336
39 Raamsdonk, L.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50
40 Cornish-Bowden, A. and Cardenas, M.L. (2001) Functional genomics. Silent genes given voice. *Nature* 409, 571–572