Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling

Timothy S. Gardner,[1†] Diego di Bernardo,[1,2†] David Lorenz,[1]
James J. Collins[1*]

[1]Center for BioDynamics and Department of Biomedical Engineering, Boston University,
44 Cummington St., Boston, MA 02215, USA
[2]Telethon Institue for Genetics and Medicine (TIGEM),
Via P. Castellino 111, 80131, Naples, Italy

[†] These authors contributed equally to this work.
[*]To whom correspondence should be addressed; E-mail: jcollins@bu.edu.

The complexity of cellular gene, protein and metabolite networks can hinder attempts to elucidate their structure and function. To address this problem, we used systematic transcriptional perturbations to construct a first-order model of regulatory interactions in a nine-gene subnetwork of the SOS pathway in *Escherichia coli***. The model correctly identified the major regulatory genes and the transcriptional targets of Mitomycin C activity in the subnetwork. This approach, which is experimentally and computationally scalable, provides a framework for elucidating the functional properties of genetic networks and identifying molecular targets of pharmacological compounds.**

Efforts to systematically define the organization and function of gene, protein and metabolite networks include experimental and computational methods for identifying molecular interactions (*1–3*), global structural properties (*4, 5*), metabolic limits (*6*), and regulatory modules and characteristics (*7–9*). These methods have provided valuable insights in many applications,

but they often provide only structural information or require extensive quantitative information which is not generally available, particularly for larger regulatory networks. In previous computational studies (*10–12*), alternative methods have been proposed that would enable rapid deduction of network connectivity and functional properties solely from temporal gene-expression data. However, the acquisition of adequate temporal expression data remains difficult, and the practical utility of such approaches has not been determined.

Here we present a rapid and scalable method that enables construction of a first-order predictive model of a gene and protein regulatory network using only steady-state expression measurements and no prior information on the network structure or function. We use multiple linear regression to determine the model from RNA expression changes following a set of steady-state transcriptional perturbations. The model can be used to identify the regulatory role of individual genes in the network, useful control points in the network, and genes that directly mediate a pharmaceutical compound's bioactivity in the cell. The method, called Network Identification by multiple Regression (NIR), is derived from a branch of engineering called system identification (*13*), in which a model of the connections and functional relationships between elements in a network is inferred from measurements of system dynamics (e.g., the response of genes and proteins to external perturbations).

To apply a system identification method, we assume that the behavior of a gene, protein and metabolite regulatory network can be modeled by a system of non-linear differential equations (*14, 15*). Near a steady-state point (e.g., when gene expression does not change significantly over time), such a non-linear system may be approximated, to the first-order, by a linear system of equations describing the rate of accumulation of each network species following a transcriptional perturbation:

$$\dot{\underline{x}} = \mathbf{A}\underline{x} + \underline{u}, \tag{1}$$

where $\underline{x}$ is a vector representing the concentrations of N RNAs, proteins, and metabolites in the network, $\dot{\underline{x}}$ represents the rate of accumulation of the species in $\underline{x}$, $\underline{u}$ is a vector representing

an external perturbation to the rate of accumulation of the species in $\underline{x}$, and $\mathbf{A}$, the network model, is an $N \times N$ matrix of coefficients describing the regulatory interactions between the species in $\underline{x}$. Next, we must identify the coefficients of $\mathbf{A}$ using only RNA expression changes that result from steady-state transcriptional perturbations. Because we measure RNA but not protein or metabolite species in this study, variables representing proteins and metabolites are not explicitly represented in the network model. Thus regulatory connections in the model are not, in general, physical connections; rather, they represent effective functional relationships between transcripts.

Under the steady-state assumption ($\dot{\underline{x}} = 0$), Eq. 1 reduces to $\mathbf{A}\underline{x} = -\underline{u}$. To identify the network model, we could, in principle, make N distinct perturbations, $\underline{u}$, to the RNAs in a particular network, recover N sets of RNA concentrations, $\underline{x}$, and solve directly for $\mathbf{A}$ (*16*). However, in larger networks it may be impractical to perform a full set of N perturbation experiments, and thus our problem would remain underdetermined. Even with a full set of perturbation experiments, RNA expression data are prone to high levels of measurement noise making the direct solution unreliable. To overcome this problem, we assume that most biochemical networks are not fully connected (*17, 18*), that is, some of the coefficients of $\mathbf{A}$ are zero. Thus, by assuming a maximum of k non-zero regulatory inputs to each gene (where $k < N$), we can transform our underdetermined problem into an overdetermined problem, making it robust both to measurement noise and incomplete data sets.

We next apply multiple linear regression (*19*) to calculate the model coefficients for each possible combination of k regulatory inputs (k coefficients) per gene. The k coefficients for each gene that fit the expression data with the smallest error are chosen as the best approximation of $\mathbf{A}$. Using the standard errors on the RNA measurement data, the algorithm also computes the statistical significance of each recovered coefficient of $\mathbf{A}$ and the overall fit of $\mathbf{A}$. A complete description of the algorithm is provided in the supporting online text.

We applied the NIR method to a nine-transcript subnetwork of the SOS pathway in *E. coli*

(the "test network"). The SOS pathway, which regulates cell survival and repair following DNA damage, involves the *lexA* and *recA* genes, more than 30 genes directly regulated by *lexA* and *recA*, and tens or possibly hundreds of indirectly regulated genes (*20–24*). We chose the nine transcripts in our test network (Fig. 1) to include the principal mediators of the SOS response (*lexA* and *recA*), four other regulatory genes with known involvement in the SOS response (*ssb*, *recF*, *dinI*, *umuDC*), and three sigma factor genes (*rpoD*, *rpoH*, *rpoS*) whose regulatory role in the SOS response is not fully understood. Because much of the regulatory structure of our test-network has been previously mapped, it serves as an excellent subject for the validation of our method. In addition, it serves as an entry point for further study of the SOS pathway which regulates genes associated with important protective pathways relevant to antibiotic resistance (*22, 25*).

We applied a set of nine transcriptional perturbations to the test network in *E. coli* cells (*26*). In each perturbation, we overexpressed a different one of the nine genes in the test network using an arabinose-controlled episomal expression plasmid (Fig. S1). We grew the cells in batch cultures under constant physiological conditions to their steady state (approximately 5.5 hours following addition of arabinose). Cells were maintained in the exponential growth phase throughout all experiments. For all nine transcripts, we used quantitative real-time PCR (qPCR) to measure the change in expression relative to that in unperturbed cells. For each transcript, two qPCR reactions from each of eight replicate cultures were obtained and qPCR data were filtered to eliminate aberrant or inefficient reactions (*26*). The mean expression changes for each transcript in each experiment ($\underline{x}$ in Eq. 1 above) was calculated (*26*) and only those changes that were greater than their standard error were accepted as significant and used for further analysis (that is, $x_i = 0$ if $|x_i| < S_{x_i}$, where x_i is the mean expression change and S_{x_i} is the standard error for transcript i).

Using the nine-perturbation expression data set (the training set—Tables S6-S8) and the NIR algorithm described above, we solved Eq. 1 for $\mathbf{A}$, the model of the regulatory interactions

in the test network (Table S1). The number of input connections per gene (k) was chosen such that the solved model provided a statistically significant fit (as determined by an F-test), was dynamically stable, and provided the best balance between coverage and false positives (*26*). To evaluate the performance of the algorithm, we determined the number of connections in the test network that were correctly resolved in the model, $\mathbf{A}$. A resolved connection was considered correct if there exists a known RNA, protein or metabolite pathway between the two transcripts and if the sign of the net effect of regulatory interaction (that is, activating or inhibiting) is correct, as determined by the currently known network in Fig. 1.

The algorithm correctly identified the key regulatory connections in the network. For example, the model correctly shows that *recA* positively regulates *lexA* and its own transcription, whereas *lexA* negatively regulates *recA* and its own transcription. In addition, the model correctly identified *recA* and *lexA* as having the greatest regulatory influence on the other genes in the test network (Table S5). Overall, the performance (coverage and false positives) of the NIR algorithm was equivalent to that expected based on simulations of 50 random nine-gene networks (Fig. 2). Moreover, for the subnetwork of 6 genes typically considered part of the SOS network (*recA*, *lexA*, *ssb*, *recF*, *dinI*, and *umuDC*), the performance of the algorithm improved significantly. This suggests that some of the false positives identified for the three sigma factors in our model (*rpoD*, *rpoH*, *rpoS*), may be true connections mediated by genes not included in our test network. Furthermore, our simulation results suggest that even small reduction in the measurement noise observed in our experiments (mean noise level $= \mathrm{mean}(S_{x_i})/\mathrm{mean}(x_i) = 68\%$) could lead to substantial improvements in coverage and errors in the network model (Fig. 2). Reductions in experimental noise could be achieved using improved RNA measurement technologies such as competitive PCR coupled with MALDI-TOF mass spectrometry (*27*).

We also tested the performance of the NIR algorithm with an incomplete training set consisting of perturbations to only 7 of the 9 genes. We solved for network models using all 36

combinations of 7 perturbations and found that the algorithm also performed comparably to simulations, albeit with slightly reduced performance than the full nine-perturbation training set (Fig. 2).

Much of the value of the network model lies in its predictive power, that is, its ability to predict expression changes and network behaviors that fall outside of the training data set used to solve the model. Here, we demonstrate its predictive power by using it to distinguish the transcripts that are directly targeted by a pharmacological compound (the compound's mode of action), and transcripts that exhibit secondary responses to the expression changes of the direct targets. Thus, the direct targets represent the minimal subset of transcripts in the model that will produce the observed expression pattern if externally perturbed. Because proteins and metabolites are not measured in this study, the compound may not physically interact with transcripts identified as direct targets, but instead may interact with protein or metabolite intermediates that are not explicitly represented in the network model.

To identify direct transcriptional targets of a compound, we first measure RNA expression changes ($\underline{x}_p$) resulting from treatment with the compound. The activity of the compound is treated as a set of unknown transcriptional perturbations ($\underline{u}_p$) that produce the measured expression changes. From Eq. 1, we calculate the unknown perturbations as $\underline{u}_p = -\mathbf{A}\underline{x}_p$ (*26*). The direct transcriptional targets of a compound are those that exhibit statistically significant values in $\underline{u}_p$. Calculation of the statistical significance of $\underline{u}_p$ is described in the supporting online text.

We first applied our scheme to RNA expression changes that result from the simultaneous controlled perturbation of the *lexA* and *recA* genes. This perturbation might represent the effects of a hypothetical compound, and serves as a well-defined input for validating the predictive power of our model. Although five of the nine test-network genes responded with statistically significant transcriptional changes (Fig. 3A), application of our network model correctly identified only *lexA* and *recA* as the perturbed genes ($2/2 = 100\%$ coverage, $7/7 = 100\%$ specificity—Fig. 3B).

We next applied a Mitomycin C (MMC) perturbation to determine if our scheme could identify the transcriptional targets of MMC bioactivity in the SOS network. Perturbed cells were grown in 0.75 μg/ml MMC and transcriptional changes were measured relative to those in control cells grown in the normal baseline condition (0.5 μg/ml MMC). All genes in the test network showed statistically significant transcriptional increases (Fig 3C). When we applied the network model to the expression data, we correctly identified *recA* as the transcriptional target of MMC bioactivity, with only one false positive, *umuDC* ($1/1 = 100\%$ coverage, $7/8 = 88\%$ specificity—Fig. 3D). Moreover, *recA* was identified at a higher significance level ($P \leq 0.09$) than was *umuDC* ($P \leq 0.22$), suggesting it is the more likely, if not the only, true target. It is also possible, however, that *umuDC* interacts with gene, protein, or metabolite targets of the compound that are not represented in our model. Therefore, *umuDC* may have been correctly identified as a target in our model. We also found that a model recovered using a seven-perturbation training set that excludes the *lexA* and *recA* training perturbations performs nearly as well as the model recovered using a full training set (see supporting online text and Fig. S3).

The NIR method, a form of system identification based on multiple linear regression analysis of steady-state transcription profiles, provides a framework for rapidly elucidating the structure and function of genetic networks using no prior information. The method is robust to high levels of measurement noise, is scalable for larger biochemical networks (*26*), and is equally applicable to transcript, protein and metabolite activity data. With advances in high-throughput measurement methods, it may soon be feasible to include protein and metabolite measurements on a large scale. The model recovered using this method enables the identification of key properties of the network, such as the major regulatory genes, and it provides a mechanism for efficiently identifying the mode of action of uncharacterized pharmacological compounds. These capabilities may facilitate optimization of cellular processes for biotechnology applications and the development of novel classes of therapeutic drugs that account for and utilize the complex

regulatory properties of genetic networks.

References and Notes

1. T. I. Lee, *et al., Science* **298**, 799 (2002).

2. T. Ideker, *et al., Science* **292**, 929 (2001).

3. A. Arkin, P. D. Shen, J. Ross, *Science* **277**, 1275 (1997).

4. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).

5. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabási, *Science* **297**, 1551 (2002).

6. J. S. Edwards, B. O. Palsson, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528 (2000).

7. E. H. Davidson, *et al., Science* **295**, 1669 (2002).

8. S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nature Genet.* **31**, 64 (2002).

9. U. S. Bhalla, R. Iyengar, *Science* **283**, 381 (1999).

10. S. Liang, S. Fuhrman, R. Somogyi, *Proc. Pacific Symp. Biocomp.* **3**, 18 (1998).

11. P. D'Haeseleer, X. Wen, S. Fuhrman, R. Somogi, *Proc. Pacific Symp. Biocomp.* **4**, 41 (1999).

12. E. P. van Someren, L. F. A. Wessels, M. J. T. Reinders, E. Backer, *Proc. 2*nd *Int. Conf. Systems Biol.* pp. 222–230 (2001).

13. L. Ljung, *System Identification: Theory for the User* (Prentice Hall, Upper Saddle River, NJ, 1999).

14. H. H. McAdams, A. Arkin, *Annu. Rev. Biophys. Biomol. Struct.* **27**, 199 (1998).

15. H. de Jong, *J. Comp. Biol.* **9**, 67 (2002).

16. A. de la Fuente, P. Brazhnik, P. Mendes, *TRENDS Genet.* **18**, 395 (2002).

17. D. Thieffry, A. M. Huerta, E. Pérez-Rueda, J. Collado-Vides, *BioEssays* **20**, 433 (1998).

18. H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai, *Nature* **411**, 41 (2001).

19. D. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis* (John Wiley & Sons, Inc., New York, 2001).

20. J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, P. C. Hanawalt, *Genetics* **158**, 41 (2001).

21. G. C. Walker, in *Escherichia Coli and Salmonella: Typhimurium Cellular and Molecular Biology* (American Society for Microbiology, Washington DC, 1996), pp. 1400–1416, second edn.

22. W. H. Koch, R. Woodgate, in *DNA Damage and Repair, Vol. 1: DNA Repair in Prokaryotes and Lower Eukaryotes* (Humana Press, Inc., Totowa, NJ, 1998), vol. 1, pp. 107–134.

23. A. R. Fernández de Henestrosa, *et al.*, *Mol. Microbiol.* **35**, 1560 (2000).

24. P. D. Karp, *et al.*, *Nucleic Acids Res.* **30**, 56 (2002).

25. K. Lewis, *Microbiol. Mol. Biol. Rev.* **64**, 503 (2000).

26. Materials, methods and supporting data are available as supporting material on Science Online.

27. C. Ding, C. R. Cantor, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3059 (2003).

28. We thank Jesper Tegner for his insights on this work. This work was supported by DARPA, NSF, ONR, the Human Frontiers Science Program, and the Telethon Institute of Genetics and Medicine.

Supporting Online Material

www.sciencemag.org

Materials and Methods

SOM Text

Figs. S1 to S5

Tables S1 to S8

References

Figure Captions

Figure 1: Diagram of interactions in the SOS network. DNA lesions caused by Mitomycin C (blue hexagon) are converted to single-stranded DNA during chromosomal replication. Upon binding to ssDNA, the RecA protein is activated (RecA*) and serves as a co-protease for the LexA protein. The LexA protein is cleaved, thereby diminishing the repression of genes that mediate multiple protective responses. Boxes denote genes, ellipses denote proteins, hexagons indicate metabolites, arrows denote positive regulation, filled circles denote negative regulation. Red emphasis denotes the primary pathway by which the network is activated following DNA damage.

Figure 2: NIR algorithm performance. Coverage (correctly identified connections / total true connections) and false positives (incorrectly identified connections / total identified connections) were calculated for SOS models solved using a 9-perturbation training set (main figures) and a 7-perturbation training set (insets). Error bars are not included in the inset for clarity. Experiment (open triangles): Coverage and false positives were calculated by comparing the solved model (Table S1) to connections described in the literature (Table S4 and Fig. 1). Because a non-significant fit was obtained for *recF*, the weights for inputs to *recF* were set to zero in the model. The mean noise observed on the mRNA measurements in our experiments was 68% ($\mathrm{noise} = S_x/\mu_x$, where S_x is the standard deviation of the mean of x, μ_x). Simulations (filled squares): Simulated perturbations were applied to 50 randomly connected networks of 9 genes with an average of 5 regulatory inputs per gene. For each perturbation to each random network, the mRNA expression changes at steady state were calculated. The noise on the perturbations was set to 20%, equivalent to that observed on perturbations in our experiments. The noise on the mRNA concentrations was varied from 10% to 70%.

Figure 3: Cells were perturbed either with a *lexA-recA* double perturbation or MMC. The mean relative expression changes (x), normalized by their standard deviations (S_x), are illustrated for the *lexA-recA* double perturbation (**A**) and the MMC perturbation (**C**). Arrows indicate the genes known to be targeted by the perturbation. Predicted perturbations in the *lexA-recA* experiment (**B**) and the MMC experiment (**D**) were calculated from the expression data in **A** and **C** using the SOS model solved with the nine-perturbation training set (*26*). The predicted perturbations to each gene (u) were normalized by their standard deviations (S_u) to determine statistical significance. In all panels, black bars indicate statistically significant, and grey bars indicate statistically non-significant. Horizontal lines denote significance levels: $P = 0.3$ (dashed), $P = 0.1$ (solid).