Inferring Microbial Genetic Networks

Statistical learning applied to transcript responses help to gauge how genes influence one another and to identify complex networks

Timothy S. Gardner, Skip Shimer, and James J. Collins

Cellular processes are governed by extensive, interconnected networks of genes, proteins, and metabolites. An important challenge in systems biology is to determine the structures and mechanisms by which these complex networks control cell processes.

Recent studies of cellular networks include methods for identifying gene and protein interactions, regulatory modules, and global structural properties. These methods are yielding valuable information and insights, but they often fail to identify the regulatory role of individual elements or the system-wide functional properties of a given network. Computational modeling and simulation also help toward understand network functions, but their utility depends on amassing extensive information, including regulatory structures, network connections, rate constants, and biochemical concentrations. Such data are generally not available, particularly for larger regulatory networks, due to the high cost and slow speed of experimentation.

In recent work, we developed an efficient method for inferring the basic structure and function of microbial genetic networks using limited experimental information. The method, called network identification by multiple regression (NIR), quantitates the influence of genes on one another using measurements of a microbe's transcriptional response to genetic perturbations. NIR assembles this information into a network model that can be used to interpret additional experimental data, make predictions about network behavior, and identify useful control points in the regulatory network.

NIR-inferred network models will be of great value in a variety of applications, including bioremediation, bioproduction of chemicals, and development of new antibiotics. For instance, efforts to develop antibiotics based on natural products or on more recent genomic-based methods that focus on lethal or "essential" genes as drug targets have delivered few new antibiotics to the clinic. Using network maps of bacterial pathways could help to overcome the shortcomings of these and other traditional methods—in part, by identifying genetic pathways and key regulators that serve as targets for new antimicrobial agents that circumvent mechanisms of resistance.

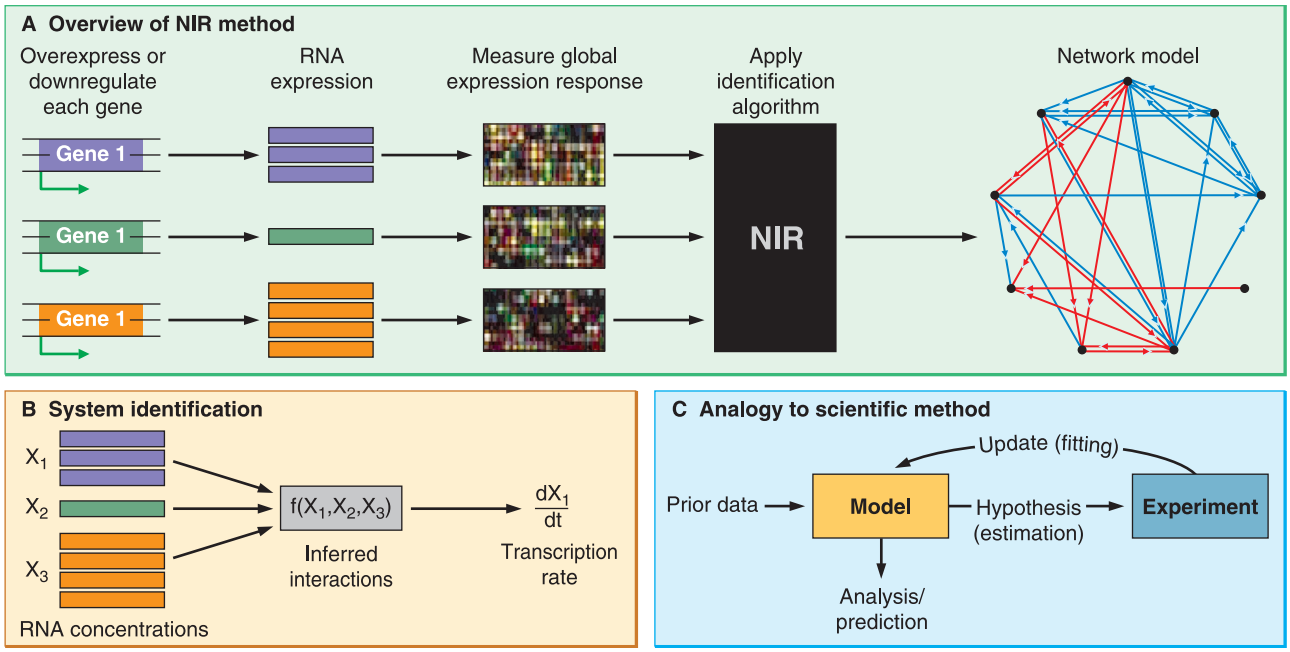
Using network maps of bacterial pathways could help to overcome the shortcomings of traditional methods

Inferring Genetic Networks with the NIR Method

The NIR method departs from structural approaches such as two-hybrid and DNA-binding assays. Structural approaches define networks through measurements of probable physical interactions, but they typically do not provide information on functional relationships between genes. On the other hand, the NIR method, which is a form of system identification, identifies quantitative regulatory relationships between genes under observed cellular conditions. Thus, with the NIR method, we can analyze and predict dynamic responses of gene networks, thereby gathering information that complements the findings of structural approaches.

Timothy S. Gardner is Assistant Professor of Biomedical Engineering and James J. Collins is Professor of Biomedical Engineering at the Center for BioDynamics and Department of Biomedical Engineering, Boston University, Boston, Mass., and Skip Shimer is the Chief Executive Officer at Cellicon Biotechnologies, Inc., Boston, Mass.

FIGURE 1



Overview of the NIR method. (A) A structured set of perturbations is delivered to cells, such as the overexpression or downregulation of one or more genes in each experiment. RNA expression (or, if possible, protein and metabolite activity) is measured for all species in the network. The data set is used by the NIR algorithm to infer a model of the perturbed network. The resulting model may then be used for analysis and prediction of network function. (B) In system identification, the inferred model relates a set of input variables to output responses using a mathematical function. In the NIR method, the input variables (X_1, X_2, X_3), are RNA concentrations of each gene, the output responses are transcription rates of each gene (dx_1/dt), and the function *f* is a weighted sum of the input variables. (C) System identification is analogous to the scientific method.

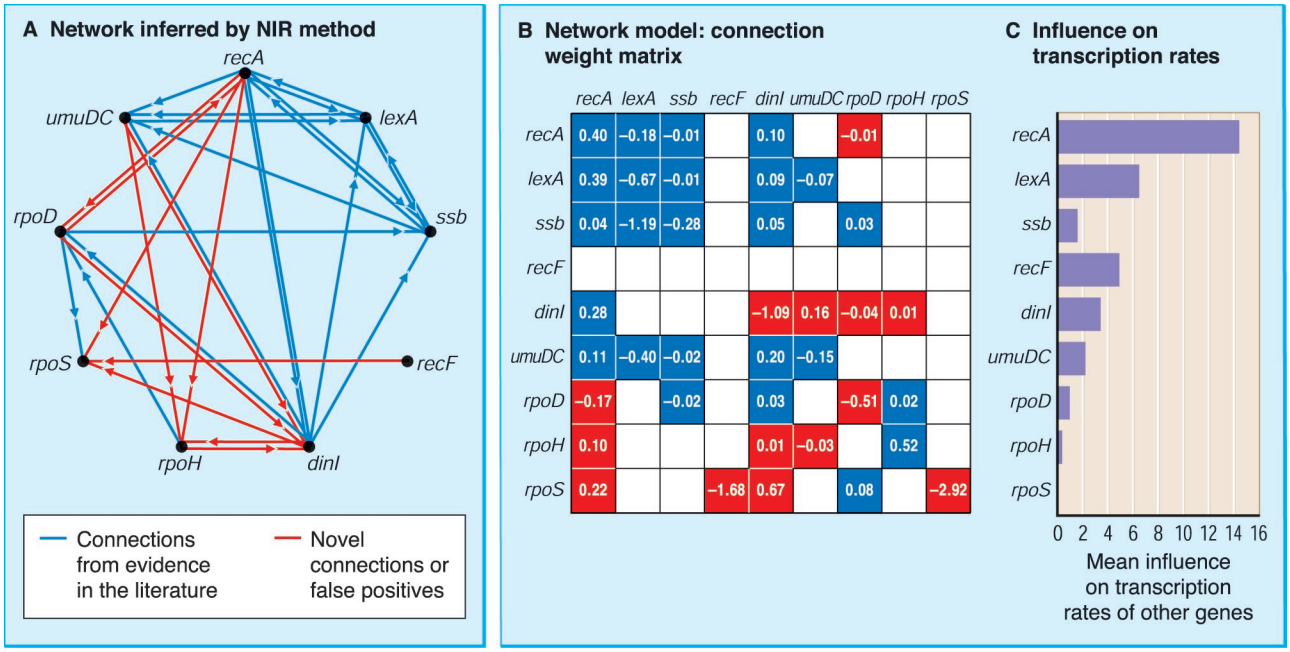
To apply the system identification approach to gene networks, controlled perturbations, such as a set of gene overexpressions, are delivered to a cell, global measurements of that cell's response are obtained, such as through transcript profiling, and an algorithm is applied to the data that identifies or learns a model of the genetic network (Fig. 1A). This model helps to define the relationship between input variables such as RNA concentrations and output variables such as gene transcription rates (Fig. 1B).

The challenge in applying a system identification method is to select a model, a learning algorithm, and an experimental design that accurately and efficiently define the network of interest. In the NIR method, we chose to represent gene interactions with a linear model. That is, the rate of transcription of each gene is represented as a weighted sum of the RNA concentrations of other genes in the network. Thus, the influence of each gene on each other gene, if any, is explained by the calculated weights. The learning algorithm, which uses experimental training data to determine each weight in the model, is based on multiple regression analysis. Finally, training data are collected by overexpressing individual genes and then measuring the steady-state RNA levels of all genes in the network.

The concept of system identification is analogous to the discovery process that a typical scientist intuitively applies (Fig. 1C). Available data are used to generate a preliminary model of the system (a formal computational model in the system identification framework), a hypothesis is generated based on that model (estimation of system outputs based on inputs), an experiment is conducted to test the hypothesis (collection of training data), and the model is updated based on the results of the experiment (fitting/learning).

For a scientist, a model is typically intuitive, or developed with the aid of various text-based, logical, graphical, or mathematical tools. In sys-

FIGURE 2



Inference of *E. coli* subnetwork using the NIR method. (A) The connections identified by the NIR method in a nine-gene subnetwork of the *E. coli* DNA damage response pathway. For visual clarity, strengths and directions of the identified connections are not labeled. Blue lines indicate connections for which there exists evidence in the scientific literature and online databases. Novel connections (or false-positives) are indicated in red. (B) The weight matrix representation of the network identified by the NIR method. The weights determine the influence of the gene in each column on the rate of transcription of the gene in each row. (C) The model is used to calculate the mean influence of each gene on transcription changes in the other genes. The model identifies *recA* and *lexA* as the primary regulatory nodes in the network, which is consistent with existing knowledge.

tem identification, the model construction, hypothesis generation, experimentation, and learning steps are assembled into a formal mathematical or computational framework. The rigor of that framework enables subsequent processing, interpretation, and analysis of complex, multivariate data, such as those generated by microbial genetic networks, that are ordinarily beyond the reach of human intuition.

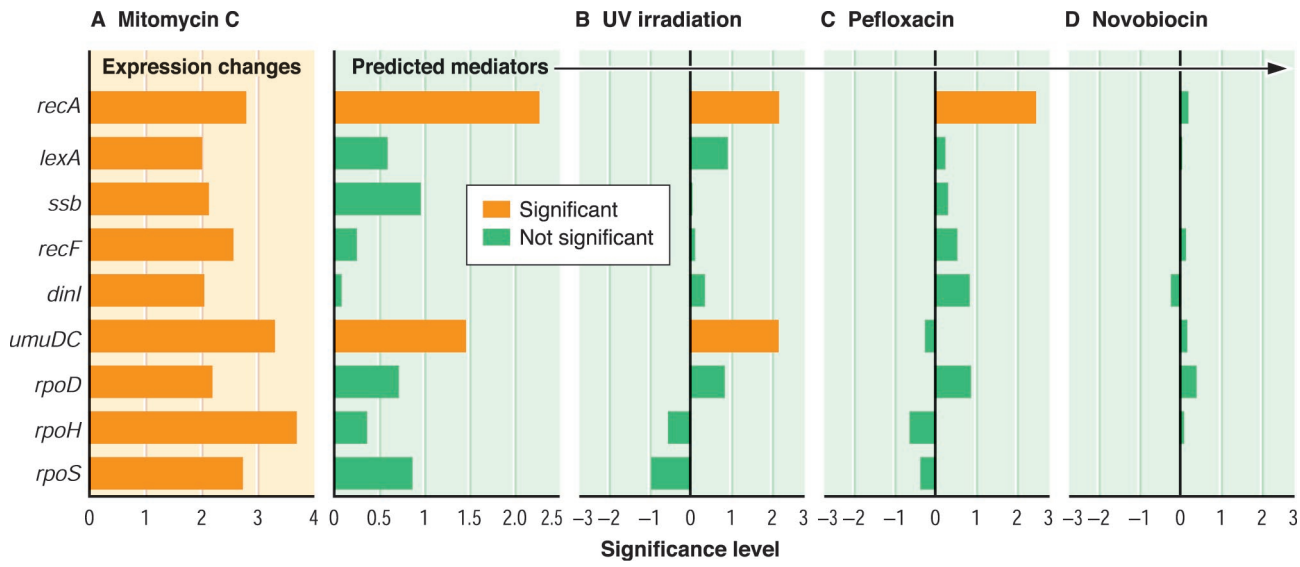
Testing NIR by Studying the SOS Pathway

We tested the NIR method on the SOS pathway in *Escherichia coli*. This extensively studied pathway, which regulates an *E. coli* cell's response to DNA damage and involves more than 100 genes, serves as a good network for validating the NIR method. As a starting point, we applied NIR to a nine-gene subnetwork at the core of the pathway, experimentally altering an inducible plasmid to overexpress each of those genes and measuring RNA responses using real-time PCR.

With the NIR method, we correctly identified 25 already recognized regulatory relationships among those nine genes as well as 14 additional relationships that either are novel regulatory pathways or false-positive findings (Fig. 2A and 2B). Moreover, the NIR-determined network model correctly identified *recA* and *lexA*, the known principal regulators of the SOS response, as having the strongest influence, or largest regulatory weights, on other genes within this network (Fig. 2C). Thus, the model can be used to suggest which genes should be perturbed to elicit a particular response from the network—a capability of great value in optimizing bacteria for environmental remediation or bioproduction of compounds.

We also used the NIR-determined network model together with additional experimental data to identify genes that mediate the network response to a drug or other stimulus. For exam-

FIGURE 3



Prediction of perturbed genes using the network model identified by the NIR method. Cells were perturbed with mitomycin C (MMC) and the resulting expression changes (A) were measured. The *recA* gene is known to mediate the SOS response following DNA damage by MMC. (B) The gene perturbations predicted by the network model to cause the observed expression changes (i.e., the mediators of the response). Only the *recA* and *umuDC* genes were predicted by the model as being mediators, and only *recA* with high confidence. The other predicted perturbations were not statistically significant. Lines denote significance levels: $P = 0.3$ (dashed), $P = 0.1$ (solid). (B–D) The network model was also applied to predict the mediators of expression responses following UV irradiation and antibiotic treatment. The expression data were obtained from public microarray data sets, but are not shown in the figures. In the case of UV irradiation and pefloxacin treatment, both DNA-damaging, *recA* is correctly predicted as the mediator of the expression response. For novobiocin, which does not damage DNA, *recA* is not predicted as the mediator of the expression response.

ple, we treated *E. coli* cells with mitomycin C (MMC), an antibiotic that leads to the formation of single-stranded DNA, thereby activating the RecA protein. Activated RecA subsequently significantly upregulates all genes in the test SOS network (Fig. 3A). When we applied this inferred network model to analyze experimental MMC response data, we found that it correctly identified *recA* as the key mediator of MMC bioactivity (Fig. 3B). The model also identified the DNA-translesion repair polymerase, *umuDC*, as a second mediator, albeit at a lower significance level.

We also applied the model to publicly accessible data that were obtained by using microarrays to assay responses of *E. coli* to various stimuli. Here again, the network model correctly identifies *recA* as the key mediator of the SOS response to DNA damage caused by UV irradiation and also treatments with the quinolone antibiotic pefloxacin. But *recA* is not identified as a mediator of *E. coli* response to novobiocin treatment, which does not damage DNA (Fig. 3D). In addition, *umuDC* is identified as a mediator of the cellular response to DNA damage resulting from UV irradiation, but not to the DNA damage resulting from pefloxacin treatment. The absence of *umuDC* as a mediator of quinolone-induced DNA damage suggests that the genes involved in DNA repair following quinolone damage are different from those associated with damage resulting from MMC and UV irradiation.

Simplifying Complexity

One of our goals in using NIR, as with using most system biology methods, is to understand properties of cellular and biochemical systems that are not apparent from studying individual components. Some researchers presume that to understand such global properties, it is necessary to build extensive computational models that integrate most of the biochemical details of a cell or gene network. However, building such a model is unrealistic both computationally and experimentally because cells are too complex.

Moreover, for most practical purposes building such a comprehensive model is unnecessary.

Traditionally, the complexity of cellular biochemistry is addressed by reducing a particular system to its components, which are then studied in isolation. But in following such an approach, valuable information about system properties is lost, precisely the opposite of what is desired in a systems analysis. Somehow, an alternative approach is needed in which the biochemistry is simplified in a way that preserves the information needed to describe system-wide properties of the network.

In system identification, that simplification is achieved by restricting the complexity of the model chosen to represent the system. For example, with NIR, we use a linear approximation of network interactions because it limits the number of required experiments, but still captures network properties of value in medical and biotechnological applications, namely by identifying major network regulators and mediators of chemical or environmental stresses.

However, this model may not be appropriate for analyzing other behaviors of microorganisms, such as the genetic interactions that orchestrate bacteriophage infection. Other models of varying complexity, including Boolean, Bayesian, and neural network models, may be better suited for analyzing such systems. In fact, one of the major challenges is to select an appropriate model structure that enables analysis of selected global properties while preserving computational tractability, speed, and experimental feasibility. We and other groups are actively exploring criteria for choosing alternative, workable models.

Building Better Drugs and Better Bugs

The NIR method has immediate applicability for improving antibiotics. For example, biofilms are involved in as many as 60% of human infections and are notoriously difficult to eradicate because cells in biofilms survive antibiotic doses several orders of magnitude higher than those sufficient to kill free-floating bacteria.

Although mutations can lead to resistance against anti-infective drugs, genetic changes are not necessarily responsible for the increased antibiotic tolerance that occurs in biofilms. For instance, cells removed from biofilms often prove to be as susceptible to antibiotics as are their free-floating counterparts. Moreover, reduced diffusion of drugs into biofilms does not generally account for their ability to withstand such drugs.

For instance, fluoroquinolone antibiotics such as ofloxacin and ciprofloxacin readily penetrate biofilms and kill many cells. Yet, small numbers of cells within biofilms apparently survive regardless of the concentration of antibiotics applied, according to Kim Lewis and colleagues at Northeastern University in Boston, Mass. These "persistor" cells are believed to repopulate biofilms after antibiotic treatments cease, and thus to cause recurrent infections. In addition, such persistence may permit advantageous mutations to be amplified.

The mechanisms of persistence are not well understood. But persistence is likely a dynamic response of the cell orchestrated by multiple stress response pathways. A better understanding of these stress response networks, obtained using the NIR method, will be of great value in identifying productive targets for novel antibiofilm compounds. The NIR-determined network model can also be used to identify genes that mediate the effects of a particular compound. Thus the network model could be of great value for optimizing candidate antibiotic compounds, and could enable the development of novel classes of drugs that target the complex regulatory properties of genetic networks.

Identifying and analyzing networks with NIR may also be valuable when optimizing microbes used in bioremediation and bioconversion schemes. Bacteria are extraordinarily flexible respirers, possessing multiple and overlapping pathways for obtaining energy by transferring electrons from high-potential compounds, or electron donors, to lower-potential entities that serve as electron acceptors, such as oxygen and metal ions. For example, *Shewanella oneidensis* can reduce solubilized heavy metals, such as uranium(VI), to an insoluble form to decontaminate ground water at waste sites, according to Derek Lovley and colleagues at the University of Massachusetts, Amherst.

While *S. oneidensis* is a remarkably capable organism, the conditions under which it is studied in laboratory cultures differ greatly from those at contaminated sites where it might be used. Indeed, when oxygen is present, *S. oneidensis* will not reduce uranium(VI).

NIR can be used to develop a model of the genetic networks regulating electron transport

pathways in *S. oneidensis* or other microorganisms. In combination with models of bacterial metabolism, such a model could help to manipulate and optimize its performance under field conditions. Similarly, metabolic pathways in other microbes could be manipulated with the help of NIR to enhance their production of valuable compounds, including pharmaceuticals and fuels.

Future Directions

One practical advantage in using NIR is its scalability. Computationally, the NIR algorithm is easily applied to large networks. Experimentally, the scalability of the method depends primarily on the speed with which perturbations can be delivered. So far, we have used only transcriptional overexpressions delivered from episomal expression plasmids, though alternatives such as knockdown approaches based on antisense RNA could also be used. Such perturbations easily could be applied to any gene and require no labor-intensive, biologically unpredictable chromosomal modifications. For example, we are extending our pilot study, using the NIR method to infer the complete *E. coli* DNA-damage response network, including more than 100 genes. We expect this effort to lead to novel or enhanced antibiotics.

Although we have applied NIR only to RNA expression data thus far, the method could just as easily be applied to measurements of proteins and metabolites. However, large-scale measurements of protein concentrations, protein activity states, and metabolite concentrations are still difficult to obtain. When analytic technologies including mass spectrometry, high-resolution electrophoresis, and protein arrays are further developed, it should become possible to use the NIR algorithm to explore the dynamic and quantitative properties of protein signaling cascades and metabolic networks. This capability will be of tremendous value in understanding the mechanisms by which such networks mediate, distinguish, and integrate environmental signals in different organisms.

ACKNOWLEDGMENTS

This work was partially supported by the Defense Advanced Research Projects Agency, the National Science Foundation, and the Office of Naval Research.

SUGGESTED READING

Cheung, K. J., V. Badarinarayana, D. W. Selinger, D. Janse, and G. M. Church. 2003. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. Genome Res. **13:**206–215.

Courcelle, J., A. Khodursky, B. Peter, P. O. Brown, and P. C. Hanawalt. 2001. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. Genetics **158:**41–64.

Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science **301:**102–105.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York.

Lee, T. I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisae*. Science **298:**799–804.

Lewis, K. 2001. Riddle of biofilm resistance. Antimicrob. Agents Chemother. **45:**999–1007.

Liang, S., S. Fuhrman, and R. Somogyi. 1998. REVEAL, a general reverse engienering algorithm for inference of genetic network architectures, Proc. Pacific Symp. Biocomp. **3:**18–29.

Lovley, D. R., E. J. P. Phillips, Y. A. Gorby, and E. R. Landa. 1991. Microbial reduction of uranium. Nature **350:**413–416.

Schwikowski, B., P. Uetz, and S. Fields. 200. A network of protein-protein interactions in yeast. Nature Biotechnol. **18:**1257–1261.

Weng, G., U. S. Bhalla, and R. Iyengar. 1999. Complexity in biological signaling systems. Science **284:**92–96.