

Identification and control of gene networks in living organisms via supervised and unsupervised learning

Michael E. Driscoll^a, Timothy S. Gardner^{b,*}

^a *Bioinformatics Program, Center for BioDynamics, Boston University, Boston, MA, USA*

^b *Department of Biomedical Engineering, Boston University, Boston, MA, USA*

Abstract

Cells efficiently carry out organic synthesis, energy transduction, and signal processing across a range of environmental conditions and at nanometer scales—rivaling any engineered system. In the cell, these processes are orchestrated by gene networks, which we define broadly as networks of interacting genes, proteins, and metabolites. Understanding how the dynamics of gene networks give rise to cellular functions is a principal challenge in biology, and identifying their structure is the first step towards their control. This knowledge has applications ranging from the improvement of antibiotics, the engineering of microbes for environmental remediation, and the creation of biologically-derived energy sources. In this review, we discuss several methods for the identification of gene networks.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Statistical inference; Learning algorithms; Biotechnology

1. Introduction

The investigation of cells from a systems-level perspective is a recent trend, motivated by parallel advances in computational technologies and experimental methods.

The two central tenets of molecular biology are that (i) genes, the fundamental units of heredity, are encoded as sequences of chemical bases in DNA and (ii) a gene is *expressed* when its DNA sequence is *transcribed* into an RNA intermediate and, through directed synthesis of amino acids, is *translated* into a protein. Proteins in turn possess most of the catalytic, mechanical, electrical and other properties needed to execute cell functions (Fig. 1).

For much of the 20th century, genes¹ have been studied in isolation. The dominant perspective in molecular biology viewed genes as independent units of heredity, activity, and function. Mendel's original "one gene,

one trait" thesis was given a biochemical basis when Beadle and Tatum posited their "one gene, one enzyme" doctrine [1]. This principle connected the unit of genetic information, the gene, with a unit of cellular function, a protein (enzymes are proteins).

This discovery motivated much of the work of classical genetics which continues up to the present, whereby a gene's function is characterized by observing organisms with a mutated version of that gene. Genetics spawned the field of genomics where the lens was shifted from genes to whole genomes. In the last two decades the new technologies of genomics have revealed the full DNA sequences of increasingly complex organisms from microbes, worms, fruit flies, and on upward to mice and humans.

These completed genomes, along with years of genetic characterization, have provided a "parts list" for many organisms. But by viewing each gene more or less independently of others, these lists describe neither how these parts work together, nor how networks of genes cooperate in complex functions [2].

* Corresponding author.

¹ In this piece, we will use the term "gene" to refer to any of its forms as a DNA sequence, its RNA intermediate, and its protein product.

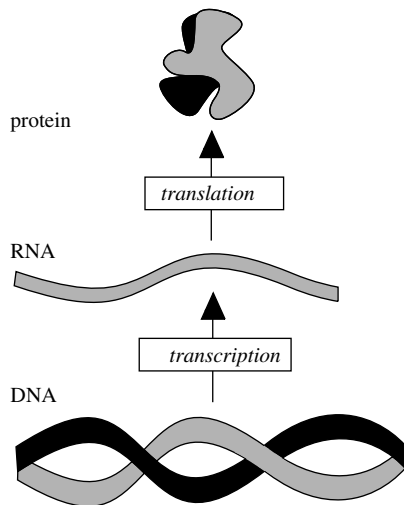


Fig. 1. Schematic illustration of gene expression, the “central dogma” of molecular biology. Genes encoded as sequences of DNA bases are transcribed into RNA transcripts, which are the template for the directed synthesis of amino acids into proteins. Multiple transcripts can be produced from each gene, and multiple proteins from each protein.

Fortunately, there is another perspective in molecular biology which views genes as being inter-dependent and inter-regulated, and has its roots in work done on bacterial phages by Jacob and Monod [3]. These early studies revealed that the expression of a gene coding for a simple sugar enzyme was subject to a sophisticated control system, and depended on the expression of other genes and biomolecules. Gene regulation has since been found to contribute to most processes in cells.

The process of gene expression allows for control at many levels. It can be altered by changing the rate of transcription into RNA, by stalling the process of its translation into protein, or even by cleaving the final protein product into pieces. Cells have evolved to use all of these mechanisms and more, but regulation of transcription is the most common. Transcription factors are an entire class of proteins involved in binding DNA to regulate this process.

Gene, proteins, and metabolites regulate one another in myriad ways (Fig. 2). Proteins bind to DNA to influence the transcription of nearby genes. Proteins also act on one another directly, mediating the addition of catalytically important chemical groups like phosphates, or combine to form multi-protein complexes that act as nanoscale machines, for example unzipping DNA or cleaving RNA. Metabolites can also bind to proteins and alter their activity.

Experimentally, it is often difficult to determine at what level gene regulation is occurring. Moreover, genes are typically embedded in vast networks of regulatory interactions [4–6]. Identifying which genes regulate each other and how they regulate each other is an outstanding challenge.

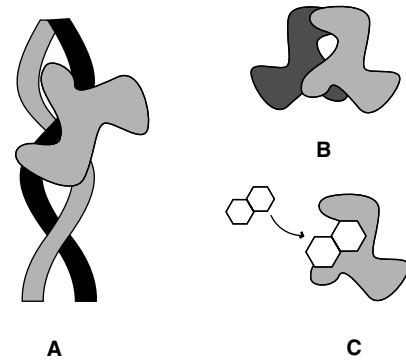


Fig. 2. Examples of gene regulation. (A) Proteins (known as transcription factors) bind to DNA, influencing the transcription of nearby genes. (B) Proteins bind to other proteins, to become active complexes. (C) Proteins bind metabolites, which modulate their activity (for example, as transcription factors).

Below, we survey some existing experimental and analytical approaches for gene network identification, and a few promising applications of this technology. In particular, we focus on methods for decoding how these networks regulate RNA expression. Although much work has been done on the modeling of metabolic networks [7,8], it is not the focus of this review.

2. Methods for gene network identification

Identifying the structure and dynamics of gene networks requires both experiment and analysis. Experimental observation employs a variety of chemical techniques to investigate the inner state of a cell, such as DNA sequencing, mass spectrometry, and labeling with fluorescent dyes. These techniques can be used, respectively, to sequence the DNA to which a transcription factor protein binds, to identify proteins which are bound to other proteins, and to measure how concentrations of RNA or proteins (both metrics of gene expression) change across conditions.

Determining protein–DNA and protein–protein complexes in the cell are examples of a physical approach to network identification [9]. It identifies physical interactions in a cellular network at a given time and condition. A key advantage of these approaches is that they provide direct evidence for a regulatory interaction; if a protein binds to the DNA sequence of a gene, it probably regulates the transcription of this gene. By themselves, however, physical approaches cannot reveal the functional nature of an interaction. It cannot reveal whether the binding of a transcription factor protein to a gene’s DNA sequence has an activating or inhibiting effect on that gene’s transcription.

An alternative to this physical approach is to construct “influence” models of gene networks [9], approaches, which seek to model causal relationships between RNA

transcript changes. The causal relationships may or may not correspond to true molecular interactions.

A chemical assay known as a microarray uses fluorescent labeling to measure the RNA concentrations of all the genes in a cell in a single experiment. A microarray consists of thousands of distinct chemical probes, each specific for a gene's RNA, arranged on a silicon or glass substrate the size of a coin. When the total RNA from an experiment is fluorescently-labeled and washed over it, and the chip is illuminated, each probe will fluoresce according to how much labeled-RNA is bound. Thus the fluorescence pattern on the chip provides a global picture of gene expression for a given experiment. Unlike approaches that measure physical binding interactions between molecules, microarray experiments only provide indirect evidence for gene interactions.

2.1. Network inference via supervised and unsupervised learning

The strategies used for network inference in biology often rely on a form of supervised or unsupervised learning. Supervised learning approaches have alternately been called machine learning and system identification, but they are all examples of parameter estimation, where a set of known inputs are matched to desired outputs. By contrast, in unsupervised learning approaches the inputs to the system are unknown, and the inference task considers only relationships among the set of outputs. Linear regression is an example of supervised learning, while clustering is an example of unsupervised learning.

In a common approach applying supervised learning to gene network inference, a set of genes are perturbed by varying their expression [10–12]. The response of the network is observed—in terms of changes in the expression of all its genes—over many time-points or at a single point when the system is thought to have settled back to a (possibly new) steady state. The gene perturbations are the known inputs, and the learning task is to identify a model and parameters for which these inputs match the observed changes in gene expression within the network (Fig. 3A).

In supervised learning, a model structure must be chosen to represent the gene network. As we will discuss, a wide variety of models have been applied to gene networks. The choice of model is informed by knowledge of the system (such as the physical kinetics of gene regulation), the kind of experimental data available, and by the research questions of interest. The learning process attempts to estimate parameters of the model by minimizing the error between model predictions and experimental data. Nonetheless, because no model is a perfect representation of reality, even the best-fitting models will include error resulting from limitations of the model and experimental noise. Once a network model's para-

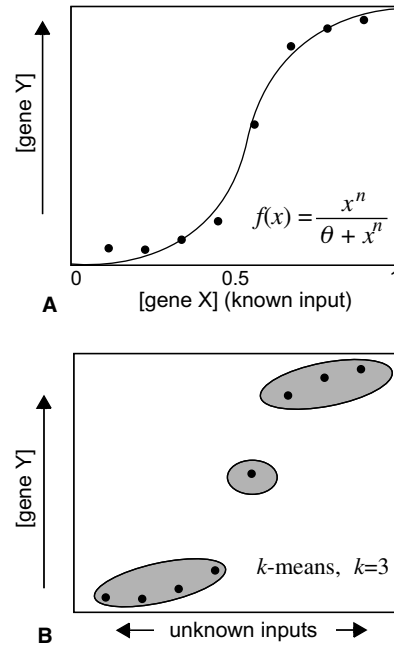


Fig. 3. An illustration of supervised and unsupervised learning. (A) Supervised learning of a gene expression input function, where gene *Y* is a function of the expression of a transcription factor, gene *X*. The chosen model structure is a Hill function, and the parameters to be learned are θ and n . (B) In unsupervised learning, where no knowledge of the input (or conditions) is known, clustering reveals only sets of related conditions.

eters are estimated, it can be used to make predictions about a network's behavior under untested conditions.

In unsupervised learning, only the outputs of the system under study are observed. In the study of gene networks, this is often the case. A typical experiment may observe the expression states of genes across a set of different conditions (such as stages of the cell cycle), but the inputs driving cells into each of these conditions are unknown. A variety of unsupervised approaches have been applied to identify common patterns in the expression of genes, including *k*-means clustering and singular value decomposition [13–15].

A natural goal of clustering is to identify groups of genes that appear to belong together. In a simple example of *k*-means clustering [13], one begins with expression measurements for genes in a single condition (these can be thought of as points scattered along a line). If we group these points into *k* sets, selecting a grouping whose centers (means) are closest to all points, we have achieved a *k*-means clustering (Fig. 3B). Stated formally, *k*-means clustering finds a partitioning that minimizes the sum:

$$D = \sum_{j=1}^k \sum_{i \in G_j} |x_i - \eta_j|^2$$

where x_i is a point and η_j is the center of partition G_j . This clustering method extends easily into higher dimensions, where each gene has expression measurements for multiple conditions.

Clustering approaches can reveal genes that trend together across conditions; the assumption is that highly correlated groups of genes are responding to a common regulatory input. But without further knowledge about the network, this regulatory input remains unknown.

In gene networks, these regulatory inputs are often transcription factors; a cluster of genes sharing an expression patterns are often the target of the same transcription factor protein. Thus, some unsupervised methods examine the DNA sequences of genes in a cluster to identify if there are statistically significant “motifs”, or sequence patterns, that represent the binding site for a common transcription factor. If identified, this method can predict causal connections between a transcription factor and a group of genes. When applied on a larger scale, this method can identify the transcriptional regulatory networks which are active for a given set of conditions [16,17].

The application of supervised learning to gene networks, on the other hand, is more recent and the remainder of our review will focus on such methods. There are two principal challenges in applying supervised learning to gene network inference: (i) the selection of the model structure and (ii) the computational scheme used to estimate parameters. The most important of these is the selection of the model structure, because it influences and ultimately determines the utility of the approach in practical applications.

2.2. Simplifying complexity

It is often presumed that in order to understand cell function at a “system level”, it is necessary to build expansive computational models that integrate much of the nature of the physical details of gene, protein and metabolite interactions in a cellular network. But such a goal is probably unrealistic both computationally and experimentally—cells are too complex. Somehow, the physical interactions must be translated into a simplified model that still preserves properties of the network relevant to a particular application or objective.

The representation of a network or other system by a simplified model is sometimes called *smoothing*. In a sense, each interaction function in the network can be represented as a surface (Fig. 4), and the details of the biochemistry are like bumps or wrinkles on the surface. Model simplifications ignore these bumps, but still capture the general shape and curvature of the surface. As the complexity (roughness) of a model representation increases (and hence its ability to describe the details), so does the amount of data needed to describe it. Thus, there is a trade-off between model scope and realism (complexity) and experimental/computational tractability. This trade-off is exaggerated in multivariate systems (which are the norm in biology). The amount of data needed increases exponentially with the number of

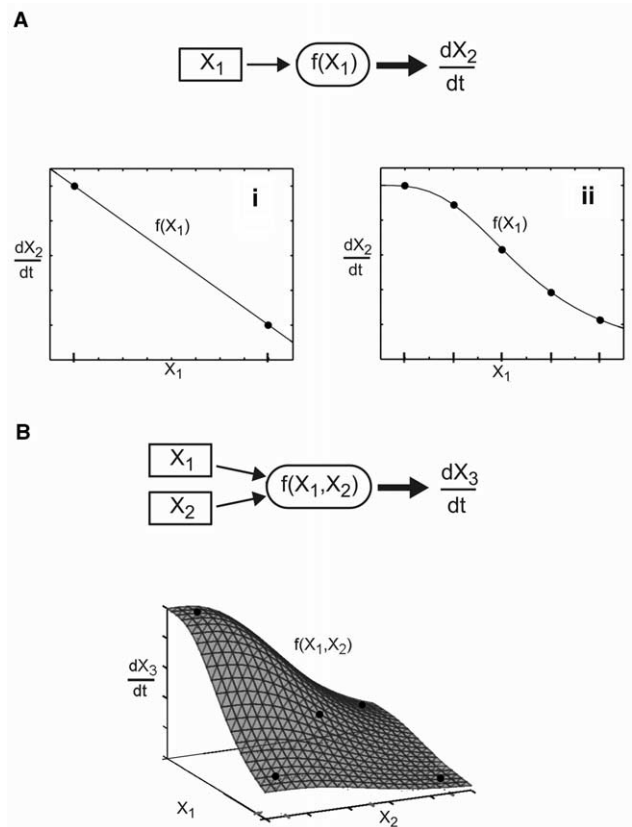


Fig. 4. The quantitative relationship between the concentration of a regulator gene and the expression of a regulated gene can be represented as a function surface f , e.g., the line, curve, and surface in panels A(i), A(ii) and B. A small number of experimental points can fully define simple surfaces such as the line in panel A(i). (Data points in the figures represent experimental measurements of the input/output relationship, which are noise-free for illustration purposes.) But a larger number of experimental points are needed to fully define more complex input/output relationships such as the curve in panel A(ii). In panel A(ii), 5 points are adequate to define relationship for a single-gene input, but for two inputs (B), $5^2 = 25$ points are needed to sample the two-input surface as densely as the one-input surface.

dimensions in the system. This problem is sometimes referred to as the *curse of dimensionality* and is illustrated in Fig. 4. To make the inference of high-dimensional systems tractable, we must use simplified models of input/output relationships, such as a hyperplane or Boolean function. However, such approximations may limit the range of questions addressable by the model.

The choice of model type and simplicity depends on several factors including the nature of the system being studied, the properties that are desired to be studied, and the type and amount of data that can be collected [18]. This choice is the major challenge in applying statistical learning to gene networks.

2.3. Model classes of gene networks

Here we discuss three types of model classes: Boolean functions, Bayesian networks, and systems of ordinary

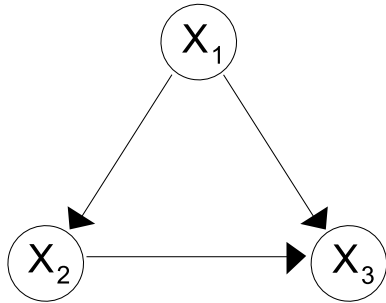


Fig. 5. An idealized three-gene network, with regulatory interactions represented by edges.

differential equations. Each model is illustrated using the simple gene network illustrated in Fig. 5. In each model, the states of the genes (e.g. the concentration or activity of RNA or protein) are represented by the variables X_1 , X_2 , and X_3 .

2.3.1. Boolean models

Boolean models are among the simplest models of gene networks, where the observed expression of genes in the network is considered to be only “on” or “off”. They are often employed when the research question under investigation is not concerned with the dynamics of transitions between “on” or “off” states of gene expression [19]. The learning task is to identify, often through perturbing genes into the “off” state (known inputs), the logical functions which describe the resulting overall state of the gene network.

For example, in research on the developmental network for the sea urchin, a series of genes were systematically deleted from the organism, and based on the observed responses of the full network, the logical circuitry of the network was inferred [10]. This work showed that gene transcription is controlled by complex combinatorial logic, driven by combinations of proteins (transcription factors) binding to DNA and regulating the transcription of genes.

In our illustrative network, based on systematic deletion of genes X_1 and X_2 (inputs) and experimental observation of X_3 (output), we might infer that X_3 is a Boolean AND function of X_1 and X_2 .

$$X_3 = X_1 \wedge X_2$$

In larger gene networks, the task of identifying the logical functions is less than trivial, as the number of experiments (inputs) formally required for n genes in a network grows as $O(n^2)$. This requirement can be greatly reduced, however, if a subset of genes (usually transcription factors, as described previously) is considered to drive the behavior of the network.

Researchers have employed two primary strategies to learn the logical functions in Boolean networks. The first strategy computes the mutual information between sets of two or more genes and tries to find the minimal set of

input genes that provides complete information on the output gene [20]. The second approach looks for the minimal set of input genes whose expression changes are consistent with a gene’s observed output states [21,22].

2.3.2. Bayesian network models

Bayesian networks are used to model regulatory interactions between genes as probabilistic relationships. In such a network, the expression level each gene is represented as a continuous random variable. The probability density function (PDF) for that random variable is assumed to be conditionally dependent upon the concentration of other genes in the network.

In the Bayesian framework, the task of reverse engineering the network is to identify the weights of these dependencies. These parameters are typically learned from large data sets, but occasionally some of these parameters may be supplied as prior information. In our example network of Fig. 5, we would hope to discover a joint probability density function showing that X_2 is dependent on X_1 , whereas X_3 depends on both X_1 and X_2 . To aid the estimation of these relationships, the joint probability of all gene in the network is broken into the product of conditional probabilities which are then estimated. In our three-gene network, this joint probability can be expressed as

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)$$

Further simplifying assumptions, such as using discrete random variables with just two states (corresponding to a gene’s being “on” or “off” as in Boolean models), or limiting the number of edges in the network can simplify the inference task and reduce the requirements for experimental data.

The probabilistic structure of a Bayesian network enables straight-forward incorporation of prior information via application of Bayes rule [23], and thus one can augment an incomplete data set with such prior information. A disadvantage of Bayesian network models is that they cannot include cycles, corresponding to feedback, which is a common property in gene networks.

The network structure is usually determined using a heuristic search, such as a greedy-hill climbing approach or Markov chain Monte-Carlo [23] method. For each network structure visited in the search, an algorithm learns the maximum likelihood parameters for the conditional distribution functions. It then computes a score for each network using the Bayes information criteria [23] or some other metric that measures the overall fit of the model to the data. One may then select the highest-scoring network as the correct network.

The development of Bayesian inference methods for gene networks has received considerable attention in recent years, and has been applied successfully to experimental expression data to identify regulatory links in gene networks [24–27].

2.3.3. Ordinary differential equation models

Systems of ordinary differential equations present a natural and semi-physical model for regulatory gene networks [28,29]. Unlike the previous models, ordinary differential models describe the kinetics of gene expression as functions. Inferring the network is a matter of identifying the parameters or coefficients of these functions.

In the illustrative network of Fig. 5, an ordinary differential model might describe the behavior of the system as:

$$\frac{dx_1}{dt} = f(E) \quad (1)$$

$$\frac{dx_2}{dt} = f(x_1) \quad (2)$$

$$\frac{dx_3}{dt} = f(x_1, x_2) \quad (3)$$

Here x_1 , x_2 , and x_3 represent RNA concentrations which, as before, are considered a measure of gene expression. The concentration of x_1 is considered a function of some external variable E , while the rates of change for x_2 and x_3 are modeled as functions of their regulators.

Introducing constraints on the system can reduce the number of experiments required to learn the parameters of the model. For small networks such as this example, it may be possible to infer parameters of a non-linear differential model that faithfully captures non-linear properties of the system from experimental data. However in larger networks, using existing experimental technologies, this task quickly becomes impractical; the data requirements for estimating many parameters are too large. An additional challenge is that most naturally occurring gene networks are multistable [30,31]. These additional states can be difficult to distinguish experimentally.

To deal with the model complexity of large-scale gene networks, one approach is to examine only the dynamics around a single stable state, in which the cells under observation are assumed to lie. This enables one to approximate the gene network with a simplified model that describes the system response to perturbations around a given steady state. For example, in our work in *E. coli* we employed a linearized model of a nine-gene network [12]. This linearized model represents the first term of a Taylor expansion of the full non-linear representation of the system.

$$\frac{d\bar{x}}{dt} = A\bar{x} + O(\bar{x}^2) \quad (4)$$

Near this steady state, the higher-order terms can be ignored. Qualitatively, the remaining A matrix represents the regulatory influences of the genes upon one another.

In a supervised learning framework, the learning task is to infer the coefficients of A . Generally speaking, the number of experiments required for this estimate scales line-

arly with the size of the network. The introduction of further constraints, such as the assumption that most of these coefficients are zero (representing the biological reality that most genes interact with just a few other genes), can further reduce the requirement for experiments.

As we show below, this approach can be used successfully to identify certain properties and features, but must be applied with caution because (1) it is specific to a particular state of the cells under study and (2) it does not capture the non-linear aspects of the system. In particular, perturbations to this system must be small enough so that these non-linear effects are not too large.

We have developed a method based on a linear model of the network and have used it to correctly infer a model of regulation in *E. coli* controlling DNA damage response and repair [12]. We have shown this model can correctly identify regulatory features and predict behaviors of the network. These results are presented briefly.

3. Network inference via multiple regression (NIR)

In our inference method, called network identification via multiple regression (NIR) gene interactions are represented by an ordinary differential model structure. The rate of synthesis of RNA from each gene is represented as a function the RNA concentrations of the other genes in the network, as described above (Fig. 6).

Experimental data are collected by artificially increasing the level of RNA for individual genes in the network. This perturbation can be represented by a vector of RNA concentrations, \bar{u} , which is added to the system at steady state.

$$\frac{d\bar{x}}{dt} = A\bar{x} - \bar{u} \quad (5)$$

The system re-settles, $d\bar{x}/dt$ goes to zero, and at its new steady state, $A\bar{x} = \bar{u}$. The vector \bar{x} describes the shift in RNA concentrations away from the initially observed steady state; this is the observed response of the network to the perturbation. After enough perturbations, the coefficients of A are identified via multiple regression of the measured RNA concentrations against the known perturbations.

We tested the NIR method on the SOS response network in *E. coli*. This network directs the maintenance and repair of genetic information in *E. coli* in response to DNA damage. When DNA is cleaved by ultraviolet light, this is sensed by proteins which bind to the broken fragments, signaling the expression of over 20 other genes, some of which are involved in mending the breaks of DNA's double-helical strands. The role of networks like this one is to integrate a variety of environmental stimuli and produce an appropriate response in terms of the expression of genes; in this case, structural genes required by the cell to survive. As the SOS network is

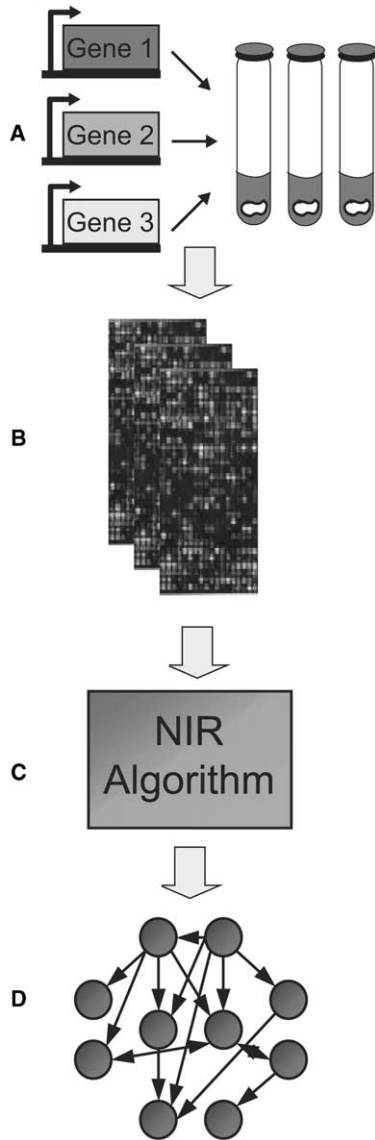


Fig. 6. Overview of the NIR method. (A) A structured set of perturbations is delivered to cells, such as the overexpression of one or more genes in each experiment. (B) Gene expression (or, if possible, protein and metabolite activity) is measured for all genes in the network. (C) This data set is analyzed by the NIR algorithm to infer a model of the perturbed network. (D) The resulting model may then be used for analysis and prediction of network function.

well described in the literature, it serves as a good network for validating the NIR method.

As a starting point, we applied the NIR method to a nine-gene subset at the core of the network. We used an extra copy of each gene to individually alter each gene's expression in nine separate experiments, and we measured the resulting changes in RNA concentrations. The NIR method was able to correctly identify 25 of the previously identified regulatory relationships between the nine genes, as well as 14 relationships that may be novel regulation pathways or possibly false positives (Fig. 7). Moreover, the network model obta-

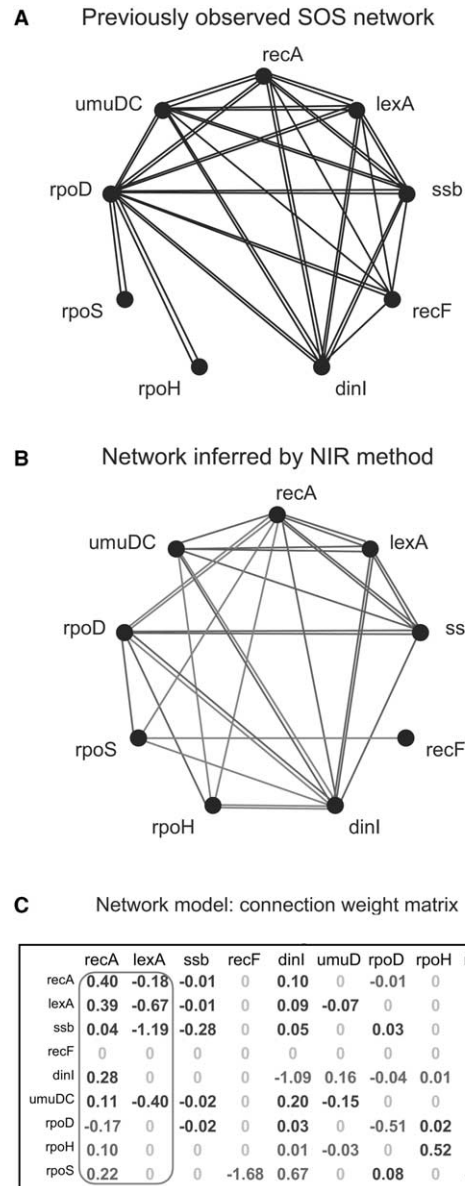


Fig. 7. Inference of *E. coli* subnetwork using the NIR method. (A) Previously known connections of the nine-gene subnetwork of the *E. coli* DNA damage response pathway. (B) The connections identified by the NIR method. For visual clarity, strengths and directions of the identified connections are not labeled. (C) The model is used to calculate the mean influence of each gene on expression changes in the other genes. The model identifies *recA* and *lexA* as the principal regulatory nodes in the network, which is consistent with existing knowledge. (Reproduced with permission from [12].)

ined by the NIR algorithm correctly identified the *recA* and the *lexA* genes, the known principal regulators of the SOS response, as having the strongest influence (largest regulatory weights) on the other genes in the network (Fig. 7C). Thus, the model can be used to suggest which genes should be perturbed to elicit a maximal response from the network—a capability of great value in optimizing bacteria for environmental remediation or bio-production of compounds.

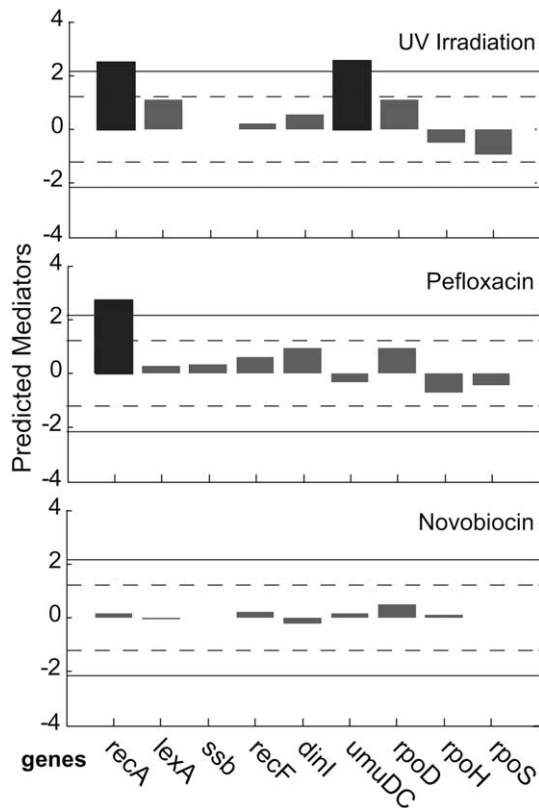


Fig. 8. Prediction of genes mediating response to three different stimuli. The previously recovered SOS network model was used to predict the mediators of expression responses following UV irradiation and treatment with two antibiotics. An independent set of expression data were obtained from public microarray data sets [32] and treated as a known output response. The NIR model was then used to predict the unknown inputs (perturbations) to the SOS network. In the case of UV irradiation and pefloxacin treatment, both DNA-damaging, the *recA* gene 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. Lines denote significance levels: $P = 0.3$ (dashed), $P = 0.1$ (solid). (Reproduced with permission from [12].)

The network model obtained by the NIR algorithm was also used to identify the genes that mediate the network response to a particular stimulus. To assay the response of *E. coli* to various stimuli, the model was applied to a publicly available set of microarrays. As illustrated in Fig. 8, the network model correctly identifies the *recA* gene as the key mediator of the SOS network response to UV irradiation and treatment with the quinolone antibiotic pefloxacin (both cause DNA damage), but not to novobiocin treatment (a quinolone that does not cause DNA damage).

4. Future work

While the development of models for gene network inference has been extensive, experimental evaluation of these methods has been limited. The approaches de-

scribed herein must be rigorously tested against existing data sets to determine where they perform best, and how they should be applied.

The existing data sets for gene expression are themselves limited by current experimental technologies. Inference methods have generally been applied only to RNA concentration data, because genome-wide measurements of protein concentrations, protein activity states, and metabolite concentrations are still difficult to obtain. The networks inferred from RNA concentration data cannot capture, except indirectly, regulation occurring at these layers.

Though still young, a variety of promising technologies are available for genome-wide profiling of proteins and metabolites, including mass-spectrometry, high-resolution electrophoresis, and protein microarrays. With these emerging technologies, it will soon be possible to use inference algorithms to explore the dynamic and quantitative properties of protein signaling cascades and metabolic networks.

The future applications of network identification are broad, even at the level of microbes. Microbes can be used for bioremediation at contaminated waste sites; harnessed to generate an electric current or produce pure hydrogen gas for renewable energy; or prevented from forming biofilms, such as those involved in surgical infections. Identifying the gene networks underlying these processes is the logical first step towards their optimization and control.

As research technologies mature and experimental data becomes available, we expect inference methods will continue to prove valuable in analyzing and predicting the behavior of gene regulatory networks [33]. This capability will be of tremendous value in understanding the mechanisms by which such networks mediate, distinguish and integrate environmental signals in microbes and higher organisms.

Acknowledgments

The authors wish to thank A. Bestavros and J. Byers as well as J.J. Collins for illuminating discussions on these topics. This work was supported by Department of Energy.

References

- [1] G.W. Beadle, E. Tatum, Genetic control of biochemical reactions in neurospora, Proc. Natl. Acad. Sci. USA 27 (1941) 499–506.
- [2] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology, Nature 402 (Suppl. 6761) (1999) 47–52.
- [3] F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins, J. Mol. Biol. 3 (1961) 318–356.
- [4] E.H.E.A. Davidson, A genomic regulatory network for development, Science 295 (5560) (2002) 1669–1678.

- [5] H. McAdams, L. Shapiro, A bacterial cell-cycle regulatory network operating in time and space, *Science* 301 (5641) (2003) 1874–1877.
- [6] U. Alon, M.G. Surette, N. Barkai, S. Leibler, Robustness in bacterial chemotaxis, *Nature* 397 (6715) (1999) 168–171.
- [7] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, E.D. Gilles, Metabolic network structure determines key aspects of functionality and regulation, *Nature* 420 (6912) (2002) 190–193.
- [8] J.A. Papin, N.D. Price, S.J. Wiback, D.A. Fell, B.O. Palsson, Metabolic pathways in the post-genome era, *Trends Biochem. Sci.* 28 (5) (2003) 250–258.
- [9] T. Gardner, J. Faith, Reverse-engineering transcription control networks, *Phys. Life Rev.* 2 (2005) 65–88.
- [10] C.H. Yuh, H. Bolouri, E.H. Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, *Science* 279 (5358) (1998) 1896–1902, comment.
- [11] Y. Setty, A.E. Mayo, M.G. Surette, U. Alon, Detailed map of a cis-regulatory input function, *Proc. Natl. Acad. Sci. USA* 100 (13) (2003) 7702–7707.
- [12] T.S. Gardner, D. di Bernardo, D. Lorenz, J.J. Collins, Inferring genetic networks and identifying compound mode of action via expression profiling, *Science* 301 (5629) (2003) 102–105.
- [13] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* 95 (25) (1998) 14863–14868.
- [14] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Natl. Acad. Sci. USA* 97 (18) (2000) 10101–10106.
- [15] M.A. Beer, S. Tavazoie, Predicting gene expression from sequence, *Cell* 117 (2) (2004) 185–198.
- [16] Y. Pilpel, P. Sudarsanam, G.M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genet.* 29 (2) (2001) 153–159.
- [17] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Systematic determination of genetic network architecture, *Nature Genet.* 22 (3) (1999) 281–285.
- [18] J. Stelling, Mathematical models in microbial systems biology, *Curr. Opin. Microbiol.* 7 (5) (2004) 513–518.
- [19] S.A. Kauffman, *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford, 1989.
- [20] S. Liang, S. Fuhrman, R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Proc. Pacific Sympos. Biocomput.* 3 (1998) 18–29.
- [21] T.E. Ideker, V. Thorsson, R.M. Karp, Discovery of regulatory interactions through perturbation: inference and experimental design, in: *Pacific Sympos. Biocomput.*, 2000, pp. 305–316.
- [22] T. Akutsu, S. Miyano, S. Kuhara, Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, in: *Pacific Sympos. Biocomput.*, 1999, pp. 17–28.
- [23] R.E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2000, Northeastern Illinois University, 1st ed.
- [24] E. Segal, B. Taskar, A. Gasch, N. Friedman, D. Koller, Rich probabilistic models for gene expression, *Bioinformatics* 17 (Suppl. 1) (2001) 243–252.
- [25] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nature Genet.* 34 (2) (2003) 166–176.
- [26] N. Friedman, Inferring cellular networks using probabilistic graphical models, *Science* 303 (5659) (2004) 799–805.
- [27] I. Nachman, A. Regev, N. Friedman, Inferring quantitative models of regulatory networks from expression data, *Bioinformatics* 20 (Suppl. 1) (2004) I248–I256.
- [28] D.C. Weaver, C.T. Workman, G.D. Stormo, Modeling regulatory networks with weight matrices, *Proc. Pacific Sympos. Biocomput.* 4 (1999) 112–123.
- [29] P. D'Haeseleer, X. Wen, S. Fuhrman, R. Somogyi, Linear modeling of mRNA expression levels during CNS development and injury, in: *Pacific Sympos. Biocomput.*, 1999, pp. 41–52.
- [30] D. Angeli, J.E.J. Ferrell, E.D. Sontag, Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems, *Proc. Natl. Acad. Sci. USA* 101 (7) (2004) 1822–1827.
- [31] E.M. Ozbudak, M. Thattai, H.N. Lim, B.I. Shraiman, A. Van Oudenaarden, Multistability in the lactose utilization network of *Escherichia coli*, *Nature* 427 (6976) (2004) 737–740.
- [32] T. Barrett, T. Suzek, D. Troup, S. Wilhite, W. Ngau, P. Ledoux, D. Rudnev, A. Lash, W. Fujibuchi, R. Edgar, NCBI GEO: mining millions of expression profiles—database and tools, *Nucl. Acids Res.* 33 (Database Issue) (2005) 562–566.
- [33] D. di Bernardo, M.J. Thompson, T.S. Gardner, S.E. Chobot, E.L. Eastwood, A.P. Wojtovich, S.J. Elliott, S.E. Schaus, J.J. Collins, Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nat. Biotechnol.* (3) (2005) 377–383.