Systems biology

Reconstructing biological networks using conditional correlation analysis

John Jeremy Rice, Yuhai Tu and Gustavo Stolovitzky*

Computational Biology Center, IBM T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA

Received on February 4, 2004; revised on August 3, 2004; accepted on August 23, 2004

Advance Access publication October 14, 2004

ABSTRACT

Motivation: One of the present challenges in biological research is the organization of the data originating from high-throughput technologies. One way in which this information can be organized is in the form of networks of influences, physical or statistical, between cellular components. We propose an experimental method for probing biological networks, analyzing the resulting data and reconstructing the network architecture.

Methods: We use networks of known topology consisting of nodes (genes), directed edges (gene–gene interactions) and a dynamics for the genes' mRNA concentrations in terms of the gene–gene interactions. We proposed a network reconstruction algorithm based on the conditional correlation of the mRNA equilibrium concentration between two genes given that one of them was knocked down. Using simulated gene expression data on networks of known connectivity, we investigated how the reconstruction error is affected by noise, network topology, size, sparseness and dynamic parameters.

Results: Errors arise from correlation between nodes connected through intermediate nodes (false positives) and when the correlation between two directly connected nodes is obscured by noise, non-linearity or multiple inputs to the target node (false negatives). Two critical components of the method are as follows: (1) the choice of an optimal correlation threshold for predicting connections and (2) the reduction of errors arising from indirect connections (for which a novel algorithm is proposed). With these improvements, we can reconstruct networks with the topology of the transcriptional regulatory network in Escherichia coli with a reasonably low error rate.

Contact: gustavo@us.ibm.com

Supplementary information: Available from our website: www.research.ibm.com/FunGen

INTRODUCTION

A fundamental task in the post-genomic era is to understand the gene regulatory networks controlling cellular functions. Although still a nascent discipline, many attempts have been made to reverse engineer biological networks or pathways (for reviews see Rice and Stolovitzky, 2004; van Someren et al., 2002). Broadly, the goals of these efforts can be placed into three hierarchical levels.

Topology. This level is concerned with mapping interactions among genes, proteins, metabolites and so on without any regard to directionality. Initial work in this area includes methods to identify metabolic network topologies based on correlation (Arkin and Ross, 1995) and later on mutual information (Samoilov et al., 2001) of time-series measurements of species concentration. Other examples include relevance networks connecting genes with similar responses (Butte and Kohane, 2000) and networks based on protein–protein interactions (Uetz et al., 2000).

Topology with qualitative connections. This level includes not only associations between cellular entities but also information on the nature of such associations. The goal is to create a diagram of directional connections (arrows) from the input to the output nodes with some qualitative indicator of how the input affects the output (e.g. a positive or negative arrow). Many methods have been proposed at this level including a more refined correlation-based approach (Arkin et al., 1997); fuzzy logic analysis (Woolf and Wang, 2000); Jacobian matrix methods (Chevalier et al., 1993; Kholodenko et al., 2002); binary network approaches (Liang et al., 1998); Bayesian network inference (Friedman et al., 2000; Hartemink et al., 2001; Smith et al., 2003); and functional module analysis (Sedgell et al., 2003; Tsyvanskaya et al., 2003).

Quantitative connections. This level seeks to determine a quantitative description of the interactions, such as a mathematical relationship that maps outputs in terms of inputs. The goal is to produce a set of equations that could simulate and reproduce the behavior of the actual system. Some examples are linear models of gene regulation (D’Haeseleer et al., 1999; Yeung et al., 2002), perturbation and regression methods (Gardner et al., 2003) and genetic algorithms (Koza et al., 2001).

In this paper, we propose a novel methodology to infer the “wiring” of basic cellular circuits of the second level described above. Our work deals with a conditional correlation-based method in which we seek to deduce directional connections in a network based on gene expression measurements. The directionality comes from the asymmetry of the conditional correlation matrix, which deals with the correlation between two genes given that one of them has been perturbed. As others have done (Koza et al., 2001; Yeung et al., 2002), we also used synthetic networks of known topology as surrogates for real biological networks. Although a study of real data would have been desirable, our method requires several replicates of the same gene perturbations, and this type of data is still unavailable. Indeed our approach indicates the need for new experimental designs for network reconstruction. Until such experiments become available, we will have to assess the performance of our method.
using synthetic networks. In addition, synthetic networks allow precise control of data quantity and quality, and the reconstructed networks can be objectively measured against the original network, a feature not generally possible with real biological networks. The networks consist of nodes (genes), directed edges (gene–gene interactions) and dynamics of the genes’ mRNA concentrations in terms of the gene–gene interactions. We interrogate the network by forcing each node (the ‘mutated’ node) to a zero concentration and measuring the resulting stationary state concentrations of the other nodes. Hence, our method does not require time-series measurements that have been considered by others (Arkin and Ross, 1995; Arkin et al., 1997; Samoilov et al., 2001; Yeung et al., 2002). Replicate measurements of stationary state concentrations are used to compute the correlation between the perturbed gene and all the other genes. Although others have used correlation of experimental expression data to infer undirected connections (Butte and Kohane, 2000), this work attempts to reconstruct whole networks in a node-by-node fashion to predict activating or inhibiting connections using conditional correlations. We test the algorithm on a variety of synthetic networks based on both random placement of connections and on connections found in the transcriptional regulatory network in Escherichia coli.

**METHODS**

We will consider networks with three different types of topologies, as described below.

**Topology I: fully synthetic networks.** In recent years, network theory has become a powerful tool to describe complex interactions used in diverse areas such as the Internet, social networks, food webs, etc. (for an overview see Barabasi, 2002). One important feature of the architecture of many of these networks is that the distribution of the number of links associated with the network nodes does not follow a uniform distribution but rather a power law distribution. These networks are called scale-free (Barabasi and Albert, 1999). A wide set of biological networks have recently been shown to exhibit scale-free behavior, such as metabolic networks (Jeong et al., 2000), protein–protein interaction networks (Jeong et al., 2001) and transcriptional networks (Baba et al., 2004). To reproduce this property in our synthetic networks, we considered random networks in which the connections between nodes are chosen at random using an approximate power law distribution. Specifically, the output degree (OD) of each node $i$, $OD_i$, is drawn from the distribution

$$
P(OD_i) = \begin{cases} \frac{\eta - 1}{\eta \eta_{min}^{\eta} - \eta \eta_{max}^{\eta}} \eta^{OD_i} & \text{if } \eta \eta_{min} \leq \eta \leq \eta \eta_{max} \\ 0 & \text{otherwise}, \end{cases}$$  

(1)

where $\eta_{min}$ and $\eta_{max}$ are the upper and lower bounds on the OD allowed in the network and $\eta$ is a constant chosen to be 2.5 (Wagner, 2002). For each node $i$, we drew $OD_i$ from the distribution of Equation (1), and node $i$ was connected to $OD_i$, of other nodes chosen randomly among the $N - 1$ other nodes. Each connection was randomly assigned to be either excitatory or inhibitory.

**Topology II: E.coli network.** To construct our E.coli transcription network, we start with the interaction topology reported by Shen-Orr et al. (2002). This network contains 423 nodes (operons) and 578 interactions. Next, we removed all the auto-regulation loops. This reduced the number of interactions to 518. In the original dataset, connections between transcription factors and promoters are positive, negative or both (i.e. some connection can be either positive or negative depending on other factors). We removed the third category by randomly assigning these connections to be either positive or negative.

**Topology III: randomized E.coli networks.** Recent studies have suggested that certain network ‘motifs’ are overrepresented in real networks (Milo et al., 2002; Shen-Orr et al., 2002) when compared with randomized networks that preserve the degree properties of the original network. To explore the effect of these motifs in the network reconstruction process, we will consider the networks for which each node has the same number of ingoing and outgoing edges as in the E.coli network, but for which the identity of the nodes connected to and from each node is randomized. This type of network is half way between the two previously discussed ones.

**Network dynamics.** Each network will be equipped with a dynamics for the mRNA concentration of each node. Following Yeung et al. (2002), the kinetic behavior governing the mRNA concentration of gene $i$, $u_i$, is considered to be determined by the concentration of the nodes impinging on node $i$ according to

$$\frac{d}{dt}u_i = -\lambda u_i + \left( \alpha + \sum_{j \in A_i} u_j \right) \eta_{1i} + \left( 1 + \sum_{j \in B_i} u_j \right) \eta_{0i} + \epsilon_i.$$  

(2)

The interpretation of this equation is as follows: the rate of change of the mRNA concentration of gene $i$ is composed of a degradation term $\lambda u_i$, the first term in the right-hand side of Equation (2), a constitutive rate of transcription $\alpha$ and the effects that other genes exert on gene $i$. The nature of this influence can be excitatory (represented by the sum over the set $A$, both in the numerator and in the denominator only) and inhibitory (represented by the sum over $B$, in the denominator only). The time scale is chosen such that the degradation rate $\lambda$ is unity. The baseline rate of transcription $\alpha$ is fixed at 0.5 for all nodes. The sets of excitatory and inhibitory nodes reaching node $i$, $A_i$ and $B_i$, are disjoint. The influence of the excitatory (inhibitory) inputs of node $j$ on node $i$ is incorporated via the Hill-like coefficients $\gamma_{ij}$ ($\beta_{ij}$), which control the non-linear dependence of the output node on the input nodes. Unless otherwise specified, these parameters will be set to 1. Noise is incorporated in the model through $\epsilon_i$, a Gaussian white-in-time noise source independent for each node. The noise strength $\epsilon$ is the same for all nodes, i.e. $\epsilon_i(t) = (t') = \epsilon \delta(t-t')$, where $\delta_{ij}$ is Kroenecker’s delta and $\delta(t-t')$ is Dirac’s delta function.

**Simulation protocol.** The numerical simulation protocol entails simulation of the intact network to produce replicate gene expression measurements of the wild-type system, and subsequently forcing one node at a time (the ‘mutated’ node) to a value of zero and then tracking the output of all nodes in the system at equilibrium. We collected the data by integrating Equation (2). After sufficient time for the initial conditions to decay, $k$ samples are obtained for each node by recording the values of all the $u_i$ ’s for every 100 units of time. This sampling time is much longer than the relaxation time of the system (of the order of 1/$\lambda$, which was set to 1). Thus, the samples can be interpreted as $k$ biological replicates of the system measured in its stationary state. Our protocol can be implemented by measuring the gene expression of all genes of interest in $k$ independent gene expression assays. In our simulations we chose $k = 10$ for each node.

In real biological systems, the type of perturbations that we model with this numerical protocol can be implemented with a variety of techniques that alter the function of single genes. Examples of these technologies are as follows: (1) antisense single-stranded RNA, which hybridizes with its target mRNA leading to reduced levels of its protein; (2) RNA interference, consisting of short double-stranded RNA of specific sequence, known as small interference RNA (siRNA), which can selectively and efficiently inhibit expression of specific genes; (3) gene-specific knockout or transgenic models, in which a gene is either completely silenced or its promoter is manipulated to render the gene inducible under the experimenter’s control; and (4) perturbation of the cellular state with small chemical compounds that target the active site of specific proteins thereby enhancing or diminishing its function. All these technologies could potentially be used singly or in combination for a real implementation of our methodology. In recent years, however, RNA interference (Dykshoorn et al., 2003) has emerged as the predominant tool with valuable properties such as specificity (Chi et al., 2003) and scalability (Luo et al., 2004). Owing to these desirable properties, this technique is simulated.
Reconstructing biological networks

Fig. 1. (A) Data from a 30-node network. Rows are scaled with zero mean and unit variance. Color scale goes from saturated green ($\leq -2$) to saturated red ($>2$). Node 30 is at its stationary state for the first 10 experiments and forced to 0 for the next 10 experiments. Row labels are the node number (left) and the order of connection to node 30 (right). (B) Illustration of the order of connection between two nodes. Node $p$ is connected to node $q$ and node $r$ via first-order connections. Nodes $q$ and $s$ and nodes $p$ and $s$ have a second-order connection.

in our numerical study. However, one must keep in mind that experimental realization of these assays are rarely as perfect as the computational counterpart, since not all genes in a network will have the appropriate siRNA for its perturbation (as is assumed in our method), and that some siRNA can silence non-intended targets (Jackson et al., 2003). Despite these limitations, efforts are underway to produce libraries of thousands of siRNAs (i.e. Paddison et al., 2004) suitable for applications such as large-scale functional genomics studies in the lines of which we are simulating here.

Sample data are shown in Figure 1 for a 30-node random network. The rows correspond to different nodes, and the columns correspond to different experiments. In each row, the data $(u_i)$ have been scaled to obtain mean 0 and variance 1. The color scale goes from saturated green ($\leq -2$) to saturated red ($>2$). Node 30 (lowest row, the mutated gene) is at its steady-state (plus fluctuations) value in the first 10 replicate experiments (corresponding to the wild-type state) and then is forced to zero for the next 10 experiments. The response of the other nodes often shows correlated (e.g. node 26) or anti-correlated (e.g. node 25) behavior with respect to the mutated node. This correlated behavior usually depends on the order of the connection between the mutated node and the other nodes. The order of the connection between a source node and a destination node corresponds to the minimum number of connections that we need to traverse to go from the former to the latter. In Figure 1A, both nodes 25 and 26 have a first-order connection to the mutated node. In Figure 1B, some examples of first-order connections are between nodes $p$ and $q$ and between nodes $p$ and $r$. Similarly, there are second-order connections between nodes $q$ and $s$ and between nodes $p$ and $s$. Such higher order connections can also produce correlations (e.g. nodes 1 and 7 have connections of the orders 2 and 3, respectively, to the mutated node).

Data processing. For each mutated node, the conditional pairwise correlation ($\rho_{lm}$) between node $l$ and node $m$ given that node $l$ was perturbed is computed as

$$
\rho_{lm} = \frac{\sum_{i=1}^{2k} [u_i(l) - \langle u_i \rangle][u_i(m) - \langle u_i \rangle]}{\left( \sum_{i=1}^{2k} [u_i(l) - \langle u_i \rangle]^2 \right)^{1/2} \left( \sum_{i=1}^{2k} [u_i(m) - \langle u_i \rangle]^2 \right)^{1/2}},
$$

where $u_i(l)$ and $u_i(m)$ are the measurements corresponding to nodes $l$ and $m$ in experiment $i$, the first $k$ experiments ($1 \leq i \leq k$) refer to the wild-type network (no mutation to any gene), the second $k$ experiments ($k+1 \leq i \leq 2k$) correspond to the network with gene $l$ mutated and $u_i(l)$ is the average of $u_i(l)$ through the $2k$ experiments. Hence, we compute a total of $N - 1$ conditional correlations for each mutated node. Note that $\rho_{lm}$ is a non-symmetric matrix.
The asymmetry stems from the fact that we were conditioning the correlation on the node corresponding to the first index (node $l$) being perturbed. In what follows we will refer to $\rho_{lm}$ as the correlation between nodes $l$ and $m$, but it should be kept in mind that this is a conditional correlation. As a simple baseline reconstruction algorithm, a positive threshold correlation value ($\rho_{\text{thresh}}$) was chosen. Subsequently, a putative connection was assumed from node $l$ to node $m$ when the correlation between these two nodes exceeds the threshold ($|\rho_{lm}| > \rho_{\text{thresh}}$). The sign of the correlation coefficient was then used to give the edge a sign (i.e. ‘+’ is activating and ‘−’ is repressing).

We can compare the reconstructed connections to the actual connections present in the network used to generate the data. A true positive (TP) corresponds to a correctly reconstructed edge with the correct sign, and a true negative (TN) corresponds to correctly determining that no edge is present. A false negative (FN) refers to a failure to correctly predict the existence of an edge. A false positive (FP) corresponds to a predicted edge that does not exist in the original network, or one that exists but has the opposite sign.

We assess the performance of the reconstruction as a function of $\rho_{\text{thresh}}$ by computing the Receiver Operator Characteristics (ROCs). In these curves, a fractional error is computed as the sum of FP and FN rates and plotted as a function of $\rho_{\text{thresh}}$. The FP rate is computed as the number of FPs divided by the sum of FPs and TNs. Similarly, the FN rate is computed as the number of FNs divided by the sum of TPs and FNs. It is customary to represent ROC curves as plots in which the TP rate is plotted as a function of the FP rate, but in such a representation the parameter used to traverse the ROC curve (in our case $\rho_{\text{thresh}}$) is implicit. As one of our main objectives is the determination of an optimal $\rho_{\text{thresh}}$, we decided to draw the ROC as the sum of the FP rate and the FN to allow for a clear visualization of the effect of the $\rho_{\text{thresh}}$ in the network reconstruction. For this study, each ROC corresponds to the reconstruction of a single network [see Supplementary materials 1 (SM1)].

**RESULTS**

**Effects of noise on network reconstruction**

Figure 2 shows ROCs for reconstructions with different noise strengths $\varepsilon$, for random networks with 100 nodes. OD$_{\text{min}} = 1$, OD$_{\text{max}} = 5$ and $\gamma, \beta = 1$. The case for no noise ($\varepsilon = 0$) is shown by the flat thin trace in Figure 2. This ROC has no $\rho_{\text{thresh}}$ dependence. Here, the number of FNs is zero, but many FPs arise from the perfect correlations between nodes connected at orders of two or more. For $\varepsilon = 0.01$, the FPs greatly increase for low values of $\rho_{\text{thresh}}$, whereas the FNs drop abruptly close to $\rho_{\text{thresh}} = 1$. Interestingly, the small amount of added noise lowers the minimum error below that found with no noise, a result that runs against intuition. This is because a moderate amount of noise strength masks higher order connections that produce FP while affecting the correlation of first-order connections to a lesser extent (see SM2). With increasing noise ($\varepsilon = 0.1$), the ROC is more U-shaped as FNs increase at smaller $\rho_{\text{thresh}}$ values. As the noise level increases, the error rates generally increase over the whole range of $\rho_{\text{thresh}}$ due to a greater number of FNs.

With moderate levels of noise, order 1 connections can be distinguished from the higher order connections, but many connections of order 2–4 produce similar correlations as first-order connections giving rise to FPs. With very noisy data even the correlation between first-order connections becomes smaller making first-order connections harder to distinguish on the basis of correlation, and the FN rate increases markedly ($\varepsilon = 0.5$ and 0.9 in Fig. 2).

**Effects of network properties on reconstruction of synthetic networks**

In the remaining sections of the paper we fix $\varepsilon = 0.1$, as it produces a U-shaped ROC that is typical of higher noise strengths. Figure 3A shows ROCs for reconstructions of networks with 30, 100 and 300 nodes, OD$_{\text{min}} = 1$ and OD$_{\text{max}} = 5$. ROC curves are U-shaped. Error rates tend to increase as noise strength is increased. For noise strengths 0.9 and 0.5, we plotted independently the FN rate dependence as a function of the correlation threshold (‘$\times$’ signs for $\varepsilon = 0.9$ and open diamonds for $\varepsilon = 0.5$) as well as the FP rate dependence on the correlation threshold (plus symbols for $\varepsilon = 0.9$; the similar curve for $\varepsilon = 0.5$ is almost indistinguishable from plus symbols and thus is not represented in the figure).

Figure 3B shows the effects of node OD on ROCs for networks in which each node has the same OD equal to 1, 3 and 5 as labeled. As the degree increases, the error rates increase as a result of the generation of more connections of order $\geq 1$ which are prone to produce FPs. In general, sparse networks with many disconnected nodes tend to be reconstructed with low error rates given that the disconnected nodes are poorly correlated and are thus less likely to generate FPs.

Figure 3C shows the effects of varying $\gamma$ and $\beta$, the exponents on the inputs to each node in Equation (2). For this set of reconstructions, the topology of the connections is the same. Reconstruction is most accurate when all exponents are equal to 1. This result can be explained as follows. Exponents of magnitude 1 produce a fairly linear input-output relationship. Higher exponents produce a steeper and less linear response. It is not surprising that the linear model of interactions provides the best reconstruction results as our method is based on correlations, a statistical measure designed to detect the linear component of dependences. Hence, the ROCs show increasingly higher error rates when the exponent magnitudes are all set to 3 or 5. One other case is shown in which the exponent magnitudes are uniformly distributed between 1 and 5. Unlike the other runs, this network is heterogeneous with respect to the connection exponents.

The ROC for this heterogeneous network with the mean exponent to 3 is quite similar to that where all exponents are set to 3. Hence, there appears to be little effect of homogeneous versus heterogeneous exponents.
Reconstructing biological networks

Methods to reduce false positives

Low error rates are achieved when the correlation of first-order connections are larger than that of higher order connections and unconnected nodes. However, order 1 connections cannot always be easily distinguished from order 2–4 connections on the basis of correlation alone, and hence some number of FPs seems unavoidable. To improve our ability to reconstruct networks, we employed two strategies. First, we attempt to choose an optimal threshold value \( \rho_{opt} \) to distinguish first from higher order connections. Second, we attempt to reject FPs that are likely to be a result of second-order connections masquerading as first-order connections. These two strategies are discussed in the next section.

Fig. 4. Logarithm of the probability density function (pdf) of the correlations between the mutated node and all the other nodes, after mutating all the nodes in the network one by one (jagged thin line) and its theoretical fit [thicker line; Equation (4)]. The fit consists of a weighted mixture of a first component corresponding to random correlations (thin line at the left) and a second component produced by first-order connections in the network (thin parabolic trace at the right). The inset shows a plot of the same two components without the weighting factor term. The intersection of these two curves, shown by the arrow at 0.645 is \( \rho_{opt} \), the optimal threshold that balances FPs and FNs. This \( \rho_{opt} \) differs from \( \rho_{thresh} \), at which the theoretical fit attains its minimum, shown by an arrow in the main plot.

To choose an optimal threshold value, we generate a model of the correlation distribution. First, the logarithm of the normalized histogram of the measured correlation was computed (Fig. 4, thin jagged line). This distribution shows a first mode at \( \rho = 0 \) due to the correlation between weakly connected nodes, and a second mode close to \( \rho \sim 1 \) presumably corresponding to directly connected nodes. The thick line in Figure 4 is a fit to the histogram based on a model with two separate components. The first component is the theoretical distribution of the correlation between two randomly chosen \( k \)-dimensional vectors (where \( k \) will be fitted by enumerative technique from the data) given by the formula

\[
P_{\text{random}}(\rho; k) = \frac{1}{2^{k-1}} \frac{\Gamma(k-1)}{\Gamma(k/2)}
\]

[with \( \Gamma(k) \) being the Gamma function, see SM3]. The second component accounts for the contribution of the highly correlated nodes, and will be modeled with a truncated Gaussian distribution:

\[
P_{\text{corr}}(\rho; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\rho - \mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\rho - \mu)^2}{2\sigma^2}\right); \quad 0 \leq \rho \leq 1.
\]

Our correlation distribution model consists of the mixture of these two components as

\[
P_{\text{mix}}(\rho; k, f, \sigma) = f P_{\text{random}}(\rho; k) + (1 - f) P_{\text{corr}}(\rho; \mu, \sigma),
\]

where \( f \) is a weighing factor. The parameters \( f, k, \mu \) and \( \sigma \) were chosen to minimize the root mean square error between the histogram and \( P_{\text{mix}}(\rho, k, f, \sigma) \) over the range \( 0 \leq \rho \leq 1 \).
Table 1. Reconstruction improvement from the triangle reduction rule. The results before (pre) and after (post) the application of the triangle reduction rule. Unless otherwise specified, $\gamma, \beta = 1$, OD$_{\text{min}} = 1$ and OD$_{\text{max}} = 5$

<table>
<thead>
<tr>
<th>Figure</th>
<th>$\rho_{\text{opt}}$</th>
<th>TP: pre/post</th>
<th>FP: pre/post</th>
<th>FN: pre/post</th>
<th>Network parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.69</td>
<td>159/157</td>
<td>621/10</td>
<td>0/2</td>
<td>$\epsilon = 0.01$; 100 nodes</td>
</tr>
<tr>
<td>2, 3A, 3C</td>
<td>0.68</td>
<td>147/147</td>
<td>157/12</td>
<td>12/12</td>
<td>$\epsilon = 0.1$; 100 nodes</td>
</tr>
<tr>
<td>2</td>
<td>0.59</td>
<td>82/79</td>
<td>101/74</td>
<td>77/80</td>
<td>$\epsilon = 0.5$; 100 nodes</td>
</tr>
<tr>
<td>3A</td>
<td>0.64</td>
<td>34/34</td>
<td>32/2</td>
<td>4/4</td>
<td>$\epsilon = 0.1$; 30 nodes</td>
</tr>
<tr>
<td>3A</td>
<td>0.68</td>
<td>421/420</td>
<td>546/130</td>
<td>23/24</td>
<td>$\epsilon = 0.1$; 300 nodes</td>
</tr>
<tr>
<td>3B</td>
<td>0.66</td>
<td>96/96</td>
<td>87/23</td>
<td>4/4</td>
<td>$\epsilon = 0.1$; OD = 1; 100 nodes</td>
</tr>
<tr>
<td>3B</td>
<td>0.65</td>
<td>248/246</td>
<td>385/24</td>
<td>52/54</td>
<td>$\epsilon = 0.1$; OD = 3; 100 nodes</td>
</tr>
<tr>
<td>3B</td>
<td>0.68</td>
<td>402/388</td>
<td>289/22</td>
<td>98/112</td>
<td>$\epsilon = 0.1$; OD = 5; 100 nodes</td>
</tr>
<tr>
<td>3C</td>
<td>0.63</td>
<td>115/115</td>
<td>268/67</td>
<td>44/44</td>
<td>$\epsilon = 0.1$; $\gamma, \beta = 3$; 100 nodes</td>
</tr>
<tr>
<td>3C</td>
<td>0.65</td>
<td>57/57</td>
<td>154/67</td>
<td>102/102</td>
<td>$\epsilon = 0.1$; $\gamma, \beta = 5$; 100 nodes</td>
</tr>
<tr>
<td>3C</td>
<td>0.67</td>
<td>106/105</td>
<td>299/58</td>
<td>53/54</td>
<td>$\epsilon = 0.1$; $\gamma, \beta = 1$–5; 100 nodes</td>
</tr>
<tr>
<td>5</td>
<td>0.72</td>
<td>518/495</td>
<td>212/85</td>
<td>10/23</td>
<td>$\epsilon = 0.1$; E.coli net; 423 nodes</td>
</tr>
</tbody>
</table>

A maximum-likelihood calculation (see SM4) shows that $\rho_{\text{opt}}$ is the value of $\rho$ for which $P_{\text{random}}(\rho; k) = P_{\text{corr}}(\rho; k, \mu, \sigma)$ (Fig. 4, inset). $\rho_{\text{opt}}$ is the $\rho_{\text{thresh}}$ that just balances competing effects: a smaller $\rho_{\text{thresh}}$ increases the FPs and a larger $\rho_{\text{thresh}}$ increases the FNs. Note that the $\rho_{\text{opt}}$ (shown by the arrow in the inset of Fig. 4) does not correspond to the minimum in the fitted histogram (shown by the arrow in the main plot in Fig. 4).

The second strategy to improve reconstructions is to reduce FPs by looking for second-order connections that can explain high correlations. The triangle reduction construction, can also generate FNs. For example, a TP edge between nodes $p$ and $r$ could be wrongly removed if the second-order path through node $q$ shows higher correlations (Fig. 1B).

The triangle reduction scheme is a heuristic criterion. However, under some assumptions it can be shown that if nodes $q$ and $s$ are not connected but nodes $q$ and $r$ and nodes $r$ and $s$ are, then it must be true that $\min(|\rho_{qs}|, |\rho_{sr}|, |\rho_{qr}|) > |\rho_{qs}|$. This means that the triangle reduction criterion is a necessary, albeit not sufficient, condition for the absence of a connection (see SM5).

We applied the two strategies outlined above sequentially. First, the $\rho_{\text{opt}}$ was determined. Next, the putative first-order connections were determined for this threshold value. Then, the triangle reduction method was applied to remove FPs. The decrease in FPs by application of the triangle reduction is slightly dependent on the order in which the triangles were chosen. For our results, we explored all the triangles in random order. Table 1 shows $\rho_{\text{opt}}$, the number of TPs, FPs and FNs before and after the application of the triangle reduction for all the networks reconstructed in this paper. It can be seen that the number of FPs is dramatically reduced for all the cases considered, without a serious increase in FNs.

Reconstruction of the transcriptional control network of E.coli

Biological networks have very different properties than their random counterparts, with certain ‘motifs’ being over- or under-represented in real versus random networks (Milo et al., 2002; Shen-Orr et al., 2002). To compare the reconstructability of biological versus the random networks considered so far, we created a network using the connection topology of the E.coli transcription network adapted from the study of Shen-Orr et al. Figure 5 shows results for the reconstructed E.coli network as the thick solid trace. Similar to earlier figures, we chose the sum of FP rate and FN rate as the measure of error. The inset in Figure 5 shows the decomposition of this error into the FP rate, which decreases as $\rho_{\text{thresh}}$ increases, and the FN rate, which increases as $\rho_{\text{thresh}}$ increases. It is clear that the FP rate has decreased considerably before the FN rate starts to increase under these conditions, making it relatively easy to separate their respective effects. The minimum error rate is 0.01 which occurs at $\rho_{\text{thresh}} = 0.62$ (shown by the solid arrow). Recall that we are computing the error rates as the sum of FP and FN rates. Therefore, our error rates give equal weight to the FP and FN rates. The minimum error rate of ~1% appears on the surface to be very good, but some care must be taken in its interpretation (see SM6). This value of $\rho_{\text{thresh}}$ produces 929 FPs, far exceeding the 518 actual connections in the network. Our method of computing $\rho_{\text{opt}}$ by optimization of the likelihood (SM4) does not impose equal weighting of the FP and FN rate, and clearly our high rate of FPs illustrates the need to reduce the number of FPs. As shown in Table 1, the number of FPs is reduced to 212 by choosing $\rho_{\text{opt}} = 0.72$ (shown by the dashed arrow in Fig. 5) via the maximization of the likelihood. With the application of the triangle reduction, the number of FPs is further reduced to 85 (a reduction of 127 FPs). This decrease in FPs produced only a moderate increase of 13 FNs. Such an approach may be justified as investigators balance their need to err in the direction of FPs and FNs.

Next, we investigated how the reconstruction of the E.coli network compares to reconstructions of random networks of similar size. The first random network is constructed with OD$_{\text{min}} = 1$ and OD$_{\text{max}} = 3$ to produce an average OD of 1.19 (similar to 1.22 for the reconstructed E.coli network). The error rate is much higher for the E.coli network as the thick solid trace. Similar to earlier figures, we chose the sum of FP rate and FN rate as the measure of error. The inset in Figure 5 shows the decomposition of this error into the FP rate, which decreases as $\rho_{\text{thresh}}$ increases, and the FN rate, which increases as $\rho_{\text{thresh}}$ increases. It is clear that the FP rate has decreased considerably before the FN rate starts to increase under these conditions, making it relatively easy to separate their respective effects. The minimum error rate is 0.01 which occurs at $\rho_{\text{thresh}} = 0.62$ (shown by the solid arrow). Recall that we are computing the error rates as the sum of FP and FN rates. Therefore, our error rates give equal weight to the FP and FN rates. The minimum error rate of ~1% appears on the surface to be very good, but some care must be taken in its interpretation (see SM6). This value of $\rho_{\text{thresh}}$ produces 929 FPs, far exceeding the 518 actual connections in the network. Our method of computing $\rho_{\text{opt}}$ by optimization of the likelihood (SM4) does not impose equal weighting of the FP and FN rate, and clearly our high rate of FPs illustrates the need to reduce the number of FPs. As shown in Table 1, the number of FPs is reduced to 212 by choosing $\rho_{\text{opt}} = 0.72$ (shown by the dashed arrow in Fig. 5) via the maximization of the likelihood. With the application of the triangle reduction, the number of FPs is further reduced to 85 (a reduction of 127 FPs). This decrease in FPs produced only a moderate increase of 13 FNs. Such an approach may be justified as investigators balance their need to err in the direction of FPs and FNs.
random network (thick gray trace in Fig. 5). Two factors increase the error rate in the random network. The first source of error is that the number of connections of order >1 is larger in the random network in which each node has at least one outgoing connection. By construction, every node is connected to at least one other node by a first-order connection, at least one other by a second-order connection, at least one other by a third-order and so on. This increase in the number of connections of higher order is a potential source of FPs. In contrast, the E.coli network is more star-like with a few nodes with high OD and many nodes with an input degree (ID) of 1 and an OD of 0. A second source of error is an increase in the ID of nodes in the random network. The E.coli network has relatively few nodes with high OD and many nodes with an ID of 1. These nodes are the easiest to reconstruct as the correlation tends to be highest when the receiving node has a single input. In contrast, as the number of input nodes increases, each input contributes less to the overall output decreasing the measured correlation. Thus, one expects direct connection to ID 1 nodes to be correctly identified more readily than direct connection to nodes with larger ID. For the examples shown, the E.coli network has 228 connections to nodes of ID 1 and 142 connections to nodes of ID 2. For comparison, the random network considered here had 157 connections to nodes with ID 1 and 186 connections to nodes of ID 2. The net effect of this difference is an increase in the FN rate, which can be clearly seen in Figure 5 as the thick gray line attains larger values than the solid line to the right of the arrows.

To check that the ID of reconstructed nodes has an important effect on the error, we developed another random network where the ID and OD of each node is preserved as the internode connections are rearranged. This degree-preserving randomized network is reconstructed to approximately the same error level as that of the original E.coli network (compare thick solid and dashed traces in Fig. 5). The small differences in traces are within reproducibility bounds. We conclude that the distribution of ID and OD of nodes plays an important role in the ability of this method to reconstruct networks. Note that sparseness is only part of the story: all three networks just considered are similarly sparse if one just computes the ratio of connections to nodes (see also SM7).

**DISCUSSION AND CONCLUSIONS**

We have shown that our method can reconstruct networks with a wide range of sizes, connectivity and noise levels with a reasonable error rate. We now discuss some limitations of our methodology. Foremost, we are only attempting to reconstruct the connections between nodes and whether the connections exert positive or negative influences. While this information can provide the biologist with useful hypotheses, it clearly fails to capture all the richness of biological interactions. Our method does not reconstruct interactions that are activating and inhibiting as is sometimes found in biology. For instance, the original E.coli transcription network contains 24 cases of these dual interactions, 10 of which are associated with a single transcription factor that regulates 72 targets. Hence, at least in this system, dual regulation occurs in a minority of cases and we expect that a majority of connections can be reconstructed.

We chose to ignore self-loops in this initial study. Self-loops may increase the difficulty in reconstructing networks. Positive autoregulation is likely to cause nodes to be more binary-like and are thus less linear. From the results of Figure 3C, reconstructions degrade as the node input-output relationship becomes less linear. The possible effects of negative autoregulation will depend on the function. Negative feedback is thought to decrease the temporal delay between an increase in activating transcription factors and a subsequent increase in the corresponding protein level (Rosenfeld et al., 2002). This type of regulation should have little effect on our methods as the data are collected at equilibrium. In contrast, negative-regulation may also reduce fluctuations in output in the face of changing conditions such as a method proposed to promote robustness (Becskei and Serrano, 2000). In this case, a node may show less of a response to small changes in its input and reconstruction may be hampered, but further study is needed to quantify this effect.

Our continuum-based network dynamics (Equation (2)) may not capture the discrete nature of nanoscale environments (McAdams and Arkin, 1999). However, a similar continuum-based model could reproduce bacterial transcriptional events as reported by high-temporal-resolution measurements in living cells by using green fluorescent protein reporter plasmids (Ronen et al., 2002). Hence, a continuum-based approach appears to be reasonable, at least for a first approximation model of the system.

Noise plays a critical role in our reconstruction method. In the noiseless case, all higher order connections are perfectly correlated giving rise to many FPs. The addition of a small amount of noise decreases the number of FPs as noise masks higher order connections (Fig. 2). However as noise increases further, the error rate increases. When substantial noise is present, first-order connections do not produce correlation magnitudes much larger than those of connections of higher orders (see SM2). In principle one can compensate for this error by considering more data. However, this approach does not fully compensate for high noise levels because more data points makes FPs to be reduced but not the FNs (see SM8). The resulting effect of taking more data is to shift \( \rho_{opt} \) to lower \( \rho_{thresh} \) values with a narrower region around the minimum error in the ROC. In real
biological experiments, collecting large amounts of data may not be feasible.

As mentioned above, more data can be collected to improve reconstruction. Indeed our model for the correlation distribution can be used to estimate the number of measurements needed to obtain a reasonably low error rate. In this paper, we collected 10 wild-type replicates and 10 replicates for each mutated node (a reasonable estimate of what might be collected experimentally) to reconstruct each node for a total of 10(N + 1) samples for the whole network. If each sample corresponds to a gene chip that samples M gene expression levels, then the network reconstruction requires a total of 10(N + 1)M gene expression levels to be measured. Note that the new data are collected to reconstruct each node, but the data collected for one node could be reused for other nodes.

Our triangle reduction scheme has some potential limitations. For example, a source node can have a direct connection as well as an indirect path through an intermediate node. When the indirect path has the same net sign as the direct path, the subnet—example, a source node can have a direct connection to a target as for one node could be reused for other nodes.

new data are collected to reconstruct each node, but the data collected requires a priori assumptions about node connections as well as biological systems. The kinetic equations for small metabolic networks may be limited as it will increase the number of FNs. However, for our reconstruction, we still achieved considerable reduction in FPs with a small increase in FNs (Table 1).

Despite the limitations discussed above, our method has some worthy advantages. We can reconstruct a wide variety of network sizes and topologies with some potential sources of error that we have characterized. While our method only provides qualitative information on the connections, we suspect that this output will be useful to the biologist, even if it contains some errors. For example, a mostly correct set of connections could provide a valuable first step in reconstructing complex biological interactions. While one might desire a full set of kinetic equations to describe the network, this goal does not appear to be tractable in a general sense at present. Reconstructions of synthetic gene networks have been attempted but assumptions of sparseness and linearity are required as well as very well-characterized derivatives of the signals (Yeung et al., 2002). Calculation of derivatives can be problematic with typically noisy biological systems. The kinetic equations for small metabolic networks have been reconstructed (Koza et al., 2001), but the method requires a priori assumptions about node connections as well as tightly controlled manipulation of node inputs that may not be biologically feasible. Clearly, no method is a one-size-fits-all in the world of systems biology. Indeed, untangling complex biological networks will almost certainly require flexible and integrated use of both data and tools. We hope that our pairwise conditional correlation method can be one of the valuable tools in this quest.

ACKNOWLEDGEMENTS

We thank A. Kershenbaum for his valuable advice on many aspects of graph theory. We also thank K. Duggar, T. Syeda-Mahamood, A. Califano and C. Wiggins for their useful discussions.

REFERENCES


Reconstructing biological networks


