A Bayesian Network Driven Approach to Model the Transcriptional Response to Nitric Oxide in \textit{Saccharomyces cerevisiae}

Jingchun Zhu\textsuperscript{1,2,\ast}, Ashwini Jambhekar\textsuperscript{1,2,b}, Aaron Sarver\textsuperscript{1,2,c}, Joseph DeRisi\textsuperscript{1,2,\ast}

1 Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California, United States of America, 2 Howard Hughes Medical Institute, University of California San Francisco, San Francisco, California, United States of America

The transcriptional response to exogenously supplied nitric oxide in \textit{Saccharomyces cerevisiae} was modeled using an integrated framework of Bayesian network learning and experimental feedback. A Bayesian network learning algorithm was used to generate network models of transcriptional output, followed by model verification and revision through experimentation. Using this framework, we generated a network model of the yeast transcriptional response to nitric oxide and a panel of other environmental signals. We discovered two environmental triggers, the diauxic shift and glucose repression, that affected the observed transcriptional profile. The computational method predicted the transcriptional control of yeast flavohemoglobin \textit{YHB1} by glucose repression, which was subsequently experimentally verified. A freely available software application, ExpressionNet, was developed to derive Bayesian network models from a combination of gene expression profile clusters, genetic information and experimental conditions.

Citation: Zhu J, Jambhekar A, Sarver A, DeRisi J (2006) A Bayesian Network Driven Approach to Model the Transcriptional Response to Nitric Oxide in \textit{Saccharomyces cerevisiae}. PLoS ONE 1(1): e94. doi:10.1371/journal.pone.0000094

INTRODUCTION

Nitric oxide (NO\textsuperscript{2-}) is a critical mediator of the cell’s innate immune response that defends against infection caused by a wide range of pathogens including fungi, bacteria, protozoan parasites and viruses [1–5]. Thus, a critical biological counter-measure would include the ability to detoxify or limit the damage from NO\textsuperscript{2-}. However, the mechanism by which pathogens defend against nitric oxide is not well understood. Recently, genomic surveys of NO\textsuperscript{2-} triggered transcriptional responses have been carried out in several fungal organisms (\textit{S. cerevisiae, H. capsulatum}, and \textit{C. albicans}) using microarrays [6–8]. In \textit{S. cerevisiae}, five genes were identified as the NO\textsuperscript{2-} detoxification gene cluster, which are activated through the transcription factor FzTlp when exposed to exogenously supplied NO\textsuperscript{2-}. The five genes are the yeast flavohemoglobin \textit{YHB1}, whose \textit{E. coli} homolog has been shown to convert NO\textsuperscript{2-} to nitrate as a potential mechanism for NO\textsuperscript{2-} detoxification, a putative sulfite pump \textit{SSU1}, and three additional uncharacterized open reading frames [8].

In addition to the detoxification gene cluster activated through FzTlp, the microarray data also revealed alterations of mRNA abundance for many other genes during NO\textsuperscript{2-} treatment [8] including genes involved in the yeast environmental stress response [9]. The pattern of yeast expression profiles was further compounded by the usage of a variety of genetic mutants and conditions that led to a complex and overlapping series of cellular responses. These include large variations in cell culture density, carbon source, and genotypes. Although a preliminary analysis by simple hierarchical clustering was sufficient to identify the NO\textsuperscript{2-} detoxification cluster, we sought to leverage the complexity of the dataset to further dissect additional regulatory mechanisms operating in these experiments.

Bayesian belief networks, a form of graphical probabilistic models encoding dependence relationships among interacting variables as probability distributions, offer a promising computational strategy for elucidating hidden inputs into nitric oxide regulation [10,11]. A Bayesian network uses nodes to model variables of interest such as gene activation or environmental perturbation. The relationship among variables, such as NO\textsuperscript{2-} treatment-triggered gene activation, can be modeled as conditional probability distributions. Given a set of observed data such as a microarray dataset, a probabilistic score can be assigned to each network model. The best models can be derived automatically from experimental data using a network learning procedure [10,11], and the derived model can be viewed as an interpretation for the biological dataset. Similar Bayesian network-driven approaches have been successfully applied to model signaling networks in primary human immune system cells, identify regulatory modules and their condition-specific regulators in \textit{S. cerevisiae}, decode the combinatorial code underlying gene expression, and infer gene regulatory networks by using expression levels of individual genes as network nodes [12–19].
In addition to modeling multivariable systems, a Bayesian network provides a natural platform for incorporating prior biological knowledge. For example, the knowledge of stress-inducible transcriptional responses can help to explain part of the NO- triggered transcription profile. A Bayesian network can incorporate such prior biological knowledge by encoding biological variables such as the presence of stress as network nodes and known biological relationships as probability distributions [20].

The goal of this work was to combine prior biological knowledge, computational modeling, and experimental feedback in an iterative cycle of hypothesis generation and testing to build a network model to decode the relationship of NO-, the transcription factor Fzf1p and other environmental signals using the genome-wide transcriptional output as measured by microarrays. Rather than simply recapitulate what was already known, we sought to develop and use a Bayesian network-driven approach to uncover previously unrecognized mechanisms of control operating within this system. As a result, we discovered two previously unappreciated environmental variables that affect regulation of 1HB1 (flavohemoglobin) mRNA abundance. Through additional experiments, we showed that flavohemoglobin expression, the primary effector of nitric oxide detoxification, is further regulated by glucose repression. Therefore, 1HB1 transcription cannot be viewed as simply constitutive, or even a switch solely controlled by nitric oxide exposure, but rather as receiving a combinatorial input from many signals. As part of this process, we have produced a freely available software package, ExpressionNet, for exploring complex datasets using Bayesian networks.

**Materials and Methods**

**Microarray Experiments**

**E1: NO- Perturbation** Log phase (OD600 1.0) strains were treated with NO- released from 1mM DPTA-NONOate (DBY7283, BY4741 fzf1Δ) and NO- gas bubbling through the media for 10 sec (DBY7283). mRNA isolated from treated (DPTA-NONOate exposure for 10, 20, 40, 80, 120 min; 120 min after gas bubbling) or untreated culture (0 min) was used to generate the Cy5 or Cy3 cDNA probes.

**E2: Glucose to Galactose I** DBY7283 strains with plasmids containing either Gal1p:UacZ or Gal1p:F2F1 were grown to OD600 1.0 in SD-URA, washed by water, then transferred to SGal-URA for continuing growth. mRNA isolated from treated (8, 12 hr after the transfer) or untreated culture (0 hr) was used to generate the Cy5 or Cy3 cDNA probes.

**E3: Glucose to Galactose II** Stationary phase (3-day-old saturated glucose culture) DBY7283 and S288c fzf1Δ strains were inoculated at OD600 0.5 in SCD, grown for 2 hr, washed by water, then transferred to SCGal for continuing growth. Total RNA isolated from treated (4, 8, 12 hr after the transfer) or untreated cultures (0 hr) was used to generate the Cy5 or Cy3 cDNA probes.

**E4: Raffinose to Galactose** DBY7283, F2F1 deleted with KanMX strains were grown to early log phase in SC raffinose. Galactose was added into the media to a final concentration of 2% for continuing growth. Since galactose is the preferred substrate and raffinose is not known to interfere with induction of the galactose pathway, there is no need to remove the remaining raffinose. mRNA isolated from sample (0, 30, 60, 120, 240 min after adding galactose) or reference (DBY7283; combined 0 and 240 min) culture was used to generate the Cy5 or Cy3 cDNA probes.

Differentially labeled cDNA probes were hybridized to yeast cDNA microarrays containing PCR probes of all yeast genes [22].

Microarray production, RNA isolation, cDNA synthesis, amino-alkylic dye coupling, hybridization and data collection were performed as previously described [22]. Microarray data were normalized using the NOMAD database (ucsd-nomad.sourceforge.net). Spots flagged by GenePix® Pro v3 (Axon Instruments) were excluded from analysis. Additional spots excluded from the analysis were both Cy3 and Cy5 signal intensities less than 2 times the background (E1, E2) and with feature intensity less than the background (E3, E4). The E4 dataset was transformed (i.e. normalized) by its 0 min data point. Complete microarray data are available at NCBI Gene expression omnibus database.

**Data Source and Preprocessing for Network Learning**

**Initial Model** The microarray data included those from experiments E1 or E2 [19 arrays] and the published dataset of yeast treated with H2O2 or menadione over 0 to 160 min (21 arrays) [9]. A subset of 130 genes with greater than two fold change in three or more data points in the E1 and E2 experiments were selected. The genes were clustered using data from experiments E1 and E2 [24]. Five major gene clusters were identified using correlation cutoff 0.75 with subsequent manual adjustment: Fzf1p early response, Fzf1p late response, ESR, oxidative phosphorylation, and galactose response clusters (Table 1). The manual adjustment consisted of combining galactose up-regulation and down-regulation clusters to the galactose response cluster and splitting the Fzf1p response cluster into Fzf1p early and late response clusters based on their initial response times (5–15 min vs. 15–45 min).

**Second Model** The microarray data included those in the initial modeling plus array data generated from experiment E3 and a published dataset monitoring the transcriptional response of the diauxic shift [22]. The 150 genes selected in the initial modeling were re-clustered using all the above microarray data [24]. Five major gene clusters were identified using cutoff 0.6 with subsequent manual adjustments: ESR, Fzf1p response, YHB1, galactose utilization and energy clusters (Table 2). The energy cluster included genes in the previously designated oxidative phosphorylation cluster and genes in the glucose utilization pathway that were previously in the galactose response cluster. After obtaining the new data of experiment E3, we realized the crucial separation within the Fzf1p response cluster was that between YHB1 and the rest of the cluster. Therefore, the Fzf1p response clusters were not separated into early and late response clusters as in the initial model. The manual adjustments consisted of separating YHB1 from the Fzf1p response cluster and forming a separate cluster containing only YHB1.

**Third Model** The microarray data included those in the second model as well as those generated from experiment E4. Gene cluster nodes were identical to those defined in the second model.

**Node States and the Learning Datasets**

The discrete states for network nodes in Table 1–3 were defined as the following: Node states for “NO treatment” were defined as the time intervals that best delineated the dynamics of the gene expression response following the treatment. The intervals were as follows: 0–5 minutes (very little transcriptional response), 5–15 minutes (increased transcriptional response), 15–45 minutes (sustained transcriptional changes), and >45 minutes (decreased transcriptional response).

States for the gene expression nodes were defined as up-regulation, down-regulation and unchanged expression, using a 2-fold threshold to convert microarray readouts to discrete values.
Node states for "galactose" were defined as "utilized" and "not utilized" depending on whether the substrate was introduced into the media. Similar states were defined for the node "oxidative agent".

Nodes for "diauxic shift" were defined with three states: aerobic respiration, anaerobic growth and unchanged metabolism; depending on the presence and direction of the shift. Similar node states were defined for the node "glucose derepression".

Table 1. Network nodes defined in the initial nitric oxide response model.

| NODE TYPE            | NODE NAME                  | NODE DEFINITION               | NODE STATES                     |
|----------------------|-----------------------------|-------------------------------|---------------------------------|
| gene cluster node    | Fzf1p early response        | change in transcriptional response | up-regulation;                 |
|                      | Fzf1p late response         |                                | down-regulation;                |
|                      | ESR                         | oxidative phosphorylation     | unchanged expression            |
|                      | galactose response          |                                |                                 |
| experimental perturbation node | nitric oxide     | duration of nitric oxide treatment | 0–5 min;                       |
|                      | galactose                   | galactose utilization         | 5–15 min;                       |
|                      | oxidative agent             | exposure to common oxidative agents | 15–45 min;                    |
|                      | genotype node              | FZF1 genotype                 | wild type;                      |
|                      | protein function node       | Fzf1p activity                | active;                         |

Table 2. Network nodes defined in the second nitric oxide response model.

| NODE TYPE            | NODE LABEL      | NODE DEFINITION               | NODE STATES                     |
|----------------------|-----------------|-------------------------------|---------------------------------|
| gene cluster node    | YHB1            | change in transcriptional response | up-regulation;                 |
|                      | Fzf1p response  |                                | down-regulation;                |
|                      | ESR             | energy                         | unchanged expression            |
|                      | galactose utilization |                                |                                 |
| experimental perturbation node | nitric oxide     | duration of nitric oxide treatment | 0–5 min;                       |
|                      | galactose       | galactose utilization         | 5–15 min;                       |
|                      | oxidative agent | exposure to common oxidative agents | 15–45 min;                    |
|                      | diauxic shift   | change of metabolism between aerobic respiration; | anaerobic growth; |
|                      | genotype node  | FZF1 genotype                 | wild type;                      |
|                      | protein function node | Fzf1p activity                | active;                         |

Node states for "galactose" were defined as "utilized" and "not utilized" depending on whether the substrate was introduced into the media. Similar states were defined for the node "oxidative agent".

Nodes for "diauxic shift" were defined with three states: aerobic respiration, anaerobic growth and unchanged metabolism; depending on the presence and direction of the shift. Similar node states were defined for the node "glucose derepression".
The learning datasets were compiled using 1) average gene expression fold change of each gene cluster converted into discrete values using a 2-fold threshold, and 2) manual annotation of the remaining values (experimental perturbations, FZF1 genotype and Fzf1p activity) based on the experimental conditions and strain genotypes (missing values were allowed). The complete learning datasets and gene membership of each cluster are available at http://derisilab.ucsf.edu/network.

Bayesian network learning and software implementation

A software application, ExpressionNet, was developed to perform Bayesian network learning. We used a Bayesian scoring function to assign a probability score for a network model. The clique-tree technique and the variable-elimination technique were implemented for efficient inference and learning [11,20,25]. The learning process started with random edge combinations, gradually improving the network topology using a greedy search strategy until the score reached a local maximum. The greedy search was iterated to generate a collection of high scoring networks. High scoring networks were subjected to small topology changes by single edge addition, deletion or reversion to expand the collection. Learning was repeated using two different prior probability distributions of the network parameters (priors), both set as a Dirichlet distribution: \( \text{Dir}(1,1, ..., 1) \) and \( \text{Dir}(P_0 \alpha, P_0 \alpha, \ldots, P_0 \alpha) \), where \( P_0 \) is a uniform distribution over the probability space of each CPD and \( \alpha = 5 \). Networks scoring within a percentile cutoff (15% for the initial and second models, 25% for the third model) using both priors were used to construct average Bayesian network models. We defined all environmental and genotype nodes as root nodes and all gene cluster nodes as leaf nodes. Missing values were handled using a Structural Expectation-Maximization algorithm [26]. ExpressionNet is available at http://expressionnet.sourceforge.net/. The derived network models and probability parameters are available at http://derisilab.ucsf.edu/network.

Flow cytometry

The YHB1-GFP strain is a C-terminus fusion of GFP obtained from a genome-wide tagged library [27]. The culture was grown to early log phase in synthetic media with 2% glucose, raffinose or galactose, washed with PBS, then transferred to synthetic medium with 2% glucose (from raffinose or galactose), or raffinose or galactose (from glucose). The cell fluorescence intensities were measured on a Becton Dickinson LSR II flow cytometer at 0, 2, 4.5, 6, 8.25, and 12 hr after the sugar was changed. For each time point, a minimum of 100,000 cells were measured to derive the mean GFP intensity.

TUP1 was deleted with KanMX in the YHB1-GFP strain. Identical experiments were performed as described earlier. Cell fluorescence intensities were measured at 0, 2, 4, 6, and 18 hr after the sugar was changed.

RESULTS

Algorithms

Our approach iterated through four steps: data collection and preprocessing, hypothesis generation, model evaluation and experimental feedback (Figure 1).
In the data collection and preprocessing step, gene expression clusters were identified from a microarray dataset and cluster expression levels were converted to discrete values. Each array was annotated with the strain genotypes, the experimental conditions, and protein functions. The discrete values of cluster expression and the annotated array attributes were combined to form the learning dataset.

In the subsequent hypothesis generation step, expression clusters, environmental signals and genotypes were defined as network nodes. An automatic learning procedure was used to find network connections that best fit the learning data set, measured by a probability score. Networks with the highest scores were collected. The derived model (Bayesian average network) was the average over all the high scoring networks found by the learning process (Materials and Methods). In the derived Bayesian average network model, each edge was associated with a confidence score \( c \), calculated as the percentage of its presence in the high scoring collection [18,21]. As part of the learning process, the conditional probability distribution (CPD) for each node was also automatically inferred.

In the third step, the derived model was compared to the current biological hypotheses and new predictions were then tested experimentally. In the last experimental feedback step, the new experimental data were compared to the data underlying the previous model. If the new data was inconsistent with the previous model, new network nodes such as a new environmental variable or an alternate gene clustering was proposed to attempt to explain the discrepancy. Experiments could also show the new predictions to be incorrect. In either case, we initiated a new iteration of the process to derive a better model, likely with a revised set of network nodes to 1) explain the conflict in the data, 2) predict the role of the new environmental variables on gene expression, and 3) eliminate any incorrect prediction.

In the following sections, we describe three iterations of the algorithm applied to a microarray dataset measuring the nitric oxide transcriptional response, generating increasingly improved transcription network models.

### The initial nitric oxide response network

In order to measure the *S. cerevisiae* transcriptional response to NO- and reactive nitrogen intermediates, and to examine the role of the transcription factor Fzf1p, we exposed wild type and *fzf1*Δ strains to chemically generated NO- (experiment E1, Materials and Methods). To determine whether Fzf1p over-expression could mimic the NO- inducible response, we performed similar experiments with wild type and *GAL1p:FZF1* strains on galactose. We then measured global mRNA levels over time using DNA microarrays (experiment E2, Materials and Methods). These data were combined with a published dataset from a perturbation experiment of yeast treated with common oxidative agents to model the oxidative or environmental stress response (ESR) [9].

A subset of 130 genes with significant expression changes in the combined dataset was selected (Materials and Methods). We defined five major gene clusters: the previously identified detoxification gene clusters [8] were divided into Fzf1p early and late response clusters (which were up-regulated by NO- in an Fzf1p dependent manner, but differed in their initial response time), the ESR cluster, the oxidative phosphorylation cluster, and the galactose response cluster.

We subsequently defined ten network nodes and their discrete states to model the transcriptional response microarray data, which included five gene cluster nodes, three experimental perturbation nodes, one genotype node, and one protein function node, to model the genome-wide transcriptional response to nitric oxide (Table 1).

Given those defined nodes, the learning dataset was compiled by combining discrete values of average cluster expression and manual annotation of experimental attributes for each array. Among them, the values of Fzf1p activity in the learning dataset were inferred based on *FZF1* genotype and the experimental conditions. For example, if the strain was *fzf1*Δ, the value was assigned to “inactive”. Any value that could not be inferred or obtained directly was set as missing values (empty data entries) in the learning dataset.

The initial derived model (10 edges with \( c > 0.9 \)) is shown in Figure 2a. According to this model, exposure to NO- generated two transcription signatures. The core NO- specific response (Fzf1p early and late response clusters), unlike other transcriptional responses, was controlled through the activation of the transcription factor Fzf1p. NO- also triggered a general environmental stress response, which was shared by exposing to oxidative agents.
Those predictions were consistent with the current understanding of the transcriptional response to NO. The edge confidence scores (c) displayed a bimodal distribution in the confidence score histogram (Figure 2b). Since the confidence score was a measurement of the data support for an edge, this distribution showed a clear separation of relationships that were highly supported (c>0.9) or unsupported (c = 0) by the data. This bimodal distribution was significantly different (Kolmogorov-Smirnov normality test; P<0.001) from the normal distribution generated from networks with the randomly assigned network connections (Kolmogorov-Smirnov normality test; P = 0.35), in which all edges showed a low level of support from the data (c = 0.39 ± 0.105).

As part of the network learning process, conditional probability distributions (CPDs) were also computed from the data (supplemental data). This included a CPD of the node “Fzf1p activity”, which had 80% missing values in the learning dataset. The derived CPD of “Fzf1p activity” (the chance of Fzf1p activity in either “active” or “inactive” state given the FZF1 genotype and the duration of NO- treatment) predicted that Fzf1p activity was transiently activated by NO- treatment (Figure 2a CPD table). The prediction was consistent with the biological hypothesis that NO- or NO- derivatives activates Fzf1p leading to transcriptional induction of a discrete set of target genes that function to protect the cell from NO-mediated stress [8].

Two highly supported edges (c>0.9) in the model were unexpected. One was directed from “galactose” to “Fzf1p early response” and the other from “galactose” to “oxidative phosphorylation” (Figure 2a, red edges). The connection from galactose to Fzf1p early response node predicted that the expression of this cluster (containing YHB1 and SSU7) would be up-regulated in response to galactose (as predicted in the probability distribution for the early response node), suggesting another input signal to some of the NO- probability distribution for the early response node), suggesting a hidden connection between the diauxic shift and the energy cluster; 2) re-grouping the five NO- detoxification genes into YHB1 and the Fzf1p response cluster that included the rest of the detoxification genes; and 3) separating out the galactose utilization genes to form the galactose utilization cluster (Table 2).

The second Bayesian network model was expanded to take into account the transcriptional response to the diauxic shift (Figure 2c). The new model maintained the sub-network of the NO- specific response mediated by Fzf1p and the relationship directed from galactose to YHB1. In addition, it revealed the previously hidden connection between the diauxic shift and the energy cluster.

Glucose derepression regulating YHB1 and the third nitric oxide response network

In order to avoid complications due to the diauxic shift in the galactose induction experiments (12 hr), the galactose induction experiments using wild type and fzf1A strains (experiment E3) were repeated using raffinose as the initial sugar source (experiment E4). This allowed a much faster induction and a shorter time course (4 hr). The results showed that the galactose utilization genes such as GAL7 and GAL10 were up-regulated; however, YHB1 induction was not observed (Figure 3c). This result was unexpected since the previous galactose induction experiments had shown that YHB1 was induced by 2–4 fold (Figure 3a wt, 3b). The difference could not be explained by the diauxic shift or other variables considered thus far.

Growth in glucose-rich media represses the transcription of a large number of genes such as enzymes in the TCA cycle, the respiratory chain, sporulation genes and genes needed for the utilization of less efficient sugar sources such as galactose [23]. To address the possibility that YHB1 is partially controlled by glucose repression, a node “glucose repression” was added to account for this effect in a third model (Figure 2e, Table 3). The subsequently derived third model strongly supported the relationship between glucose derepression and YHB1 gene expression; at the same time the overall network structure including the Fzf1p mediated NO- specific response sub-network was maintained. The CPD of node “YHB1” predicted that YHB1 gene expression was up-regulated by either glucose derepression or Fzf1p, but not by galactose.

To verify the prediction of glucose derepression for YHB1, the protein expression levels of GFP tagged Yhb1p were monitored under glucose repression and derepression conditions using flow cytometry. The repression results showed that Yhb1p levels decreased immediately after the sugar source was changed from either raffinose or galactose to glucose, and continued to decrease up to 2–3 fold after 12 hours. This result was confirmed by the
Figure 2. The Bayesian average network representation of the models. (a, b) initial model, (c, d) second model, (e, f) third model. (a, c, e) network graphic representation. The green nodes represent gene expression clusters. Representative genes of each cluster are shown in the box below each node. ESR: environmental stress response cluster. energy: glucose metabolism cluster. oxidative stress: the application of H2O2 or menadione. Nitric oxide: the duration of NO exposure. galactose: galactose utilization. diauxic shift: shift between anaerobic growth and aerobic respiration. Nodes with missing values are colored in gray. The CPD table shows the conditional probability distribution of Fzf1p activity. The red edges represent novel predictions from the first network model. (b, d, f) edge confidence score histogram. The dot-filled columns represent edges excluded from a model by structural constraints.

doi:10.1371/journal.pone.0000094.g002
Our results show YHB1 is important for the survival of yeast under oxidative and nitrosative stress [28,29]. Our results show YHB1 is transcriptionally regulated by both NO· exposure mediated through Fzflp and glucose repression mediated by Tup1p. Taken together, these data indicate that YHB1 is regulated by many environmental signals, highlighting the combinatorial control of this gene. While glucose derepression caused a 2 to 3-fold increase in Yhb1p protein levels, studies have shown a 10-fold increase by NO· treatment, suggesting a more prominent role of Yhb1p in NO· detoxification [8].

Within the context of our Bayesian model, we were able to utilize available biological knowledge to systematically explore the response to nitric oxide by the refinement of the random variables used. Clearly, incorporation of prior biological knowledge has the effect that our results will be biased towards our current understanding of the problem. While this fact represents a caveat, all models make assumptions; and the biological knowledge in this case was extremely useful to uncover the underlying relationships. Indeed, prior knowledge in this case can be considered to be a critical property of the process, since proper definition of the random variables used to model the dataset was essential to arrive at a biologically meaningful conclusion.

A common practice in statistical learning is to select one single model that best fits the data. But in many situations, other models also score very well although not necessarily the best. Using a single highest scoring model to derive a biological conclusion is potentially risky. To circumvent this problem, we used the average of all the high scoring networks found by the searching procedure [21]. An added benefit of this approach is that it yields a confidence score associated with each edge connection [18]. The confidence score is especially useful for filtering out low-confidence connections from a complex network, thus simplifying what might otherwise be a confusing network.

Since Bayesian network edges represent statistical instead of causal relationships, it is possible that a derived edge does not represent a direct biological connection. For example, two gene clusters sharing high mutual information would likely be connected. One method to eliminate such connections is to merge those highly correlated clusters into a single node. Additionally, structural constraints may be used to define gene expression nodes as leaf nodes and the environmental variable nodes as root nodes.

Gene clusters were defined through an automatic hierarchical clustering algorithm with manual interventions. Although it is not purely automatic, this step was where we incorporated prior biological knowledge to interpret the gene expression dataset. Therefore it is critical for ensuring the defined gene cluster nodes that truly represent the underlying gene expression profiles.
The computational framework and experimental approach presented here essentially represents a supervised data exploration system. The overall methodology is a straightforward hypothesis-generating, testing, and refinement cycle. However, complex datasets with large numbers of measurements become increasingly difficult to represent and score with regard to a given hypothesis. The creation and use of Bayesian networks, incorporating prior knowledge, allows for systematic scoring of a given hypothesis, and furthermore, provides an opportunity for automatic learning, which in turn can facilitate the discovery of new relationships within the data.

REFERENCES

1. Shiloh MU, Nathan CF (2000) Reactive nitrogen intermediates and the pathogenesis of Salmonella and mycobacteria. Curr Opin Microbiol 3: 35–42.
2. Nathan C, Shiloh MU (2000) Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. Proc Natl Acad Sci U S A 97: 8841–8848.
3. Misail TA, Lodge JK, McEwen JE (2004) Mechanisms of resistance to oxidative and nitrosative stress: implications for fungal survival in mammalian hosts. Eukaryot Cell 3: 435–446.
4. Fang FC (2004) Antimicrobial reactive oxygen and nitrogen species: concepts and controversies. Nat Rev Microbiol 2: 820–832.
5. Fang FC (1999) Nitric oxide and infection. New York: Kluwer Academic/ Plenum Publishers. xxv, 517 p.
6. Hromatka BS, Noble SM, Johnson AD (2005) Transcriptional Response of C. albicans to Nitric Oxide and the Role of the YHB1 Gene in Nitrosative Stress and Virulence. Mol Biol Cell.
7. Nübler MP, Hocking-Murray D, Foo CK, Sil A (2005) Identification of Histoplasm capsulatum Transcripts Induced in Response to Reactive Nitrogen Species. Mol Biol Cell.
8. Sarver A, Derisi J (2005) Fzf1p Regulates an Inducible Response to Nitrosative Stress in Saccharomyces cerevisiae. Mol Biol Cell.
9. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257.
10. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo Calif.: Morgan Kaufmann Publishers. xix, 552 p.
11. Jordan MI (1999) Learning in graphical models. Cambridge Mass.: MIT Press. xix, 552 p.
12. Sachs K, Perez O, Pe’er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308: 523–529.
13. Segal E, Koller D (2003) Discovering Molecular Pathways from Protein Interaction and Gene Expression data. Bioinformatics 1: 1–9.
14. Segal E, Shapiro M, Regev A, Pe’er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176.
15. Segal E, Yelenksy R, Koller D (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19 Suppl 1: i275–i282.
16. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117: 185–198.
17. Pe’er D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. Bioinformatics 17: S215–224.
18. Friedman N, Linial M, Nachman I, Pe’er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7: 601–620.
19. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput: 422–433.
20. Heckerman D, Geiger D, Chickering D (1995) Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20: 197–243.
21. Horning J, Madigan D, Raftery A, Volinsky T (1999) Bayesian Model Averaging: A Tutorial. Statistical Science 14: 382–417.
22. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680–686.
23. Broach JR, Pringle JR, Jones EW (1994) The Molecular and cellular biology of the yeast Saccharomyces. Cold Spring Harbor N.Y.: Cold Spring Harbor Laboratory Press. v. 1–3 p.
24. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
25. Heckerman D (1999) A tutorial on learning with Bayesian networks. In: Jordan MI, ed. Learning in graphical models. CambridgeMass.: MIT Press. vii, 634 p.
26. Friedman N (1999) The Bayesian structural EM algorithm. In: SCGM, ed. Machine Learning 20: 197–243.
27. Friedman N (1998) The Bayesian structural EM algorithm. In: SCGM, ed. MadisonWI, , USA: Morgan Kaufmann Publishers 24–26 July: 129–138.
28. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686–691.
29. Liu L, Zeng M, Hausladen A, Heitman J, Stamler JS (2000) Protection from nitrosative stress by yeast flavohemoglobin. Proc Natl Acad Sci U S A 97: 4672–4676.
30. Zhao XJ, Raitt D, P VB, Clewell AS, Kwast KE, et al. (1996) Function and expression of flavohemoglobin in Saccharomyces cerevisiae. Evidence for a role in the oxidative stress response. J Biol Chem 271: 25131–25138.