Revising Regulatory Networks: From Expression Data to Linear Causal Models

S. D. Bay,[1] J. Shrager,[1,2] A. Pohorille,[3] and P. Langley[1]

[1]Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

[2]Department of Plant Biology, Carnegie Institute of Washington

[3]Center for Computational Astrobiology and Fundamental Biology
NASA Ames Research Center, M/S 239-4, Moffett Field, CA 94305

Abstract

Discovering the complex regulatory networks that govern mRNA expression is an important but difficult problem. Many current approaches use only expression data from microarrays to infer the likely network structure. However, this ignores much existing knowledge because for a given organism and system under study, a biologist may already have a partial model of gene regulation. We propose a method for revising and improving these initial models, which may be incomplete or partially incorrect, with expression data. We demonstrate our approach by revising a model of photosynthesis regulation proposed by a biologist for Cyanobacteria. Applied to wild type expression data, our system suggested several modifications consistent with biological knowledge. Applied to a mutant strain, our system correctly modified the disabled gene. Power experiments with synthetic data that indicate that reliable revision is feasible even with a small number of samples.

1 Introduction

An important problem in molecular biology is explaining how an organism regulates its levels of gene expression in response to external stimuli. Although scientists understand the basic mechanisms through which DNA produces proteins and thus biochemical behavior, they have yet to determine most of the regulatory networks that control the degree to which each gene is expressed.

DNA microarrays let scientists measure gene activity in terms of mRNA expression levels in an organism. Much recent work in computational biology has focused on inferring a regulatory network that describes how genes influence each other solely from such expression data. However, this approach is rarely pursued by practicing biologists, who bring a wealth of knowledge to the analysis and interpret data about the expression levels in this context.

For a particular organism and system under study, a biologist often has a partial model of gene regulation. Although this model may be incomplete or partially incorrect, it contains much information that could influence an algorithm that infers a model of the regulatory relations between genes. In this paper, we describe an approach that uses gene expression data to drive revision of an initial regulation model. Our goal is to build a computational tool that assists working biologists in constructing models and modifying them in response to observations.

Throughout this paper, we will focus on a model of photosynthesis regulation in Cyanobacteria that a microbiologist proposed to explain physiological adaptation in high light conditions. We discuss how one can map models of this type, which are both qualitative and abstract, into linear causal models, a statistical representation that makes contact with the data. With this connection we can generate qualitative predictions and compare them with the data to guide revision and discovery of causal relations. We demonstrate a computational method that uses expression data for wild type and mutant Cyanobacteria to revise this model of photosynthesis regulation. We also conduct power experiments with synthetic data to determine the reliability of suggested revisions with small sample sizes and with larger models. Finally, we consider limitations of our approach and discuss directions for future work.

2 Background

We focus on a model of photosynthesis regulation that was adapted from a model provided by a microbiologist [9][1]. The model, shown in Figure 1, aims to explain why Cyanobacteria bleaches when exposed to high light conditions and how this protects the organism. Each node in the model corresponds to an observable or theoretical variable; each link stands for a biological process through which one variable influences another. Solid lines in the figure denote internal processes, while dashes indicate processes connected to the environment.

The model states that changes in light level modulate the expression of dspA, a protein hypothesized to serve as a sensor. This in turn regulates NBLR and NBLA proteins, which

[1]The paper describes an initial model for high light response in the Cyanobacterium *S*ynechococcus. This model was modified by the biologist for the Cyanobacterium used in our experiments, *S*ynechocystis PCC6803, by actions such as replacing nblS with its homolog dspA.

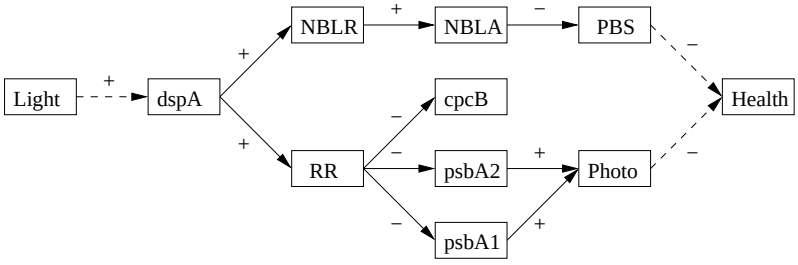

Figure 1: Initial model for photosynthesis of wild type Cyanobacteria.

then reduce the number of phycobilisome (PBS) rods that absorb light. The level of PBS is measured photometrically as the organism's greenness. The reduction in PBS protects the organism's health by reducing absorption of light, which can be damaging at high levels. The organism's health under high light conditions can be measured in terms of the culture density. The sensor dspA impacts health through a second pathway by influencing an unknown response regulator RR, which in turn down regulates expression of the gene products psbA1, psbA2, and cpcB. The first two positively influence the level of photosynthetic activity (Photo) by altering the structure of the photosystem. If left unregulated, this second pathway would also damage the organism in high light conditions.

Although the model incorporates quantitative variables, it is qualitative in that it specifies cause and effect but not the exact numerical form of the relationship. For example, one causal link indicates that increases in NBLR will increase NBLA, but it does not specify the form of the relationship, nor does it specify any parameters.

The model is both partial and abstract. The biologist who proposed the model made no claim about its completeness and clearly viewed it as a working hypothesis to which additional genes and processes should be added as indicated by new data. Some links are abstract in the sense that they denote entire chains of subprocesses. For example, the link from dspA to NBLR stands for a signaling pathway, the details of which are not relevant at this level of analysis. The model also includes an abstract variable RR, an unspecified gene product (or possibly a set of gene products) which acts as an intermediary controller.

3 Methods

Our approach to revising regulatory networks is based on linear causal models, also referred to as structural equation models [1], and methods for learning them from data [8, 17]. Linear causal models provide a statistical representation that connects models provided by biologists with experimental data provided by microarray measurements of mRNA. They make predictions that can be tested against data, and from these tests one can revise the models to better explain the data.

3.1 Linear Causal Models

A linear causal model represents each variable as a linear function of its direct causes plus an error term. For example, the equations below represent a model that states X_1 directly causes X_2, and X_1 and X_2 together cause X_3.

$$X_2 = b_{12}X_1 + e_1 \quad (1)$$
$$X_3 = b_{13}X_1 + b_{23}X_2 + e_2 \quad (2)$$

For a gene regulation model, the variables X_i would correspond to the expression levels of genes or measurements of external quantities, the linear parameters b_{ij} represent the causal effect of variable i on j, and finally the error terms e_i are assumed to be independent and uncorrelated.

There is a direct mapping from the equations to a graphical notation. Each variable becomes a node and linear terms (causal influences) are represented by an arrow from the cause to the effect. For example, the above equations are equivalent to the diagram in Figure 2a. We focus on models where the graph is acyclic (i.e., there are no feedback cycles).

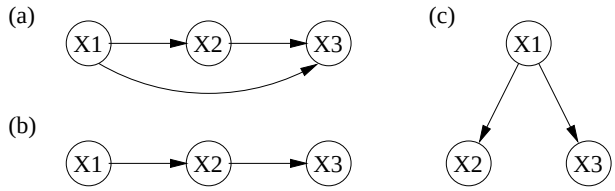

Figure 2: Several alternative models of regulation among variables X_1, X_2, and X_3.

Linear causal models support reasoning at a range of qualitative and quantitative levels, and make predictions that can be scored against data. At the most qualitative level, the model specifies causal interactions, that is, it specifies how the variables directly and indirectly influence each other. For example, the model in Figure 2a states that X_1 directly influences both X_2 and X_3. In contrast, the model in Figure 2b states that X_1 influences X_3 only through the intermediate X_2. Second, at a slightly more detailed level, the model specifies the type of causal interaction. If the sign on a link from $X_1 \rightarrow X_2$ is positive, then X_2 should increase with X_1. Conversely, if the sign is negative, X_2 should decrease as X_1 increases. This analysis can be extended to indirectly connected variables by tracing the connecting paths and multiplying the signs of the link parameters. Finally, at the most detailed level of analysis, we can specify the exact values for the additive parameters, b_{ij}, and obtain a fully quantitative model that predicts numerical values. However, because of limited data we do not pursue this level of analysis.

Linear causal models would clearly be a simplification of any biological system they represent. However, given the extremely limited number of samples available from most microarray experiments, which is often as few as five samples, they are promising because they can use a small number of parameters to represent activation and repression relationships between genes. While more complex models can better represent a wider range of relations such as thresholds or combinatorial interactions, they increase the risk of overfitting with small sample sizes.

Recently, a variety of linear models have been proposed for modeling gene regulation [3, 21, 20]. These approaches all represent the expression (or change in expression) of a gene as a linear function of the expression levels of other genes. Our approach differs from these in two important ways. First, we concentrate on discovering causal relationships between the variables, whereas previous approaches focus on finding predictive but not necessarily causal relations between genes. For example, D'Haeseleer et al. [3] use a multiple regression method that identifies correlations between gene expression levels but cannot determine if genes are linked directly or indirectly connected through other genes. Second, we try to bring as much domain knowledge as possible into the inference process by starting from partial initial models and using constraints on the model structure to limit search.

3.2 Making Predictions and Scoring Models

The structure and parameters of every linear causal model imply predictions about the correlations between variables that can be supported or refuted by observations. We discuss predictions that follow from structure and ones that follow from the signs of the parameters.

3.2.1 Structure and Correlation Constraints

The structure of the model, that is the pattern of directed links between variables, implies certain equality constraints on the correlation values between variables. For example, consider the model in Figure 2b. If we calculate the correlation of X_1 and X_3 from the model's equations, we find that $\rho_{13} = \rho_{12}\rho_{23}$, where ρ_{ij} is the correlation of variables i and j.[2] Note that this relation is true for any values of the parameters b_{ij}, and it provides a testable prediction that can be scored on data without the need to learn the parameter values first.

We can interpret this equality relationship as a zero partial correlation, also known as a *vanishing partial correlation*. Formally, the partial correlation between variables X_1 and X_3 while controlling for X_2 is defined as

$$\rho_{13.2} = \frac{(\rho_{13} - \rho_{12}\rho_{23})}{\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}}\ , \tag{3}$$

where ρ_{ij} is the correlation coefficient for variables i and j. Thus, if $\rho_{13} = \rho_{12}\rho_{23}$ the numerator is zero and the partial correlation must be zero. Like the correlation coefficient, partial correlations vary from -1 to 1. A zero value indicates that the controlling variable k perfectly explains the correlation between i and j. A non-zero value indicates that the correlation is not fully explained.

Partial correlations are significant because they help us determine whether correlated variables are directly linked or whether they are indirectly linked and the correlation is *spurious* [19]. Given two correlated variables, a zero partial correlation means that the variables are connected through the third variable. In contrast, a non-zero partial correlation means that the two variables are connected by paths that do not involve the third variable.[3] This

[2]The calculation involves taking expectation of the variables as defined by the equations. Glymour et al. [8] discuss this in detail.

[3]If the partial correlation is non-zero for all possible sets of controlling variables, then we can infer that the variables are directly connected.

analysis depend on assumptions that variables are uncorrelated to other non-descendents given their parents (causal Markov assumption) and that all common causes are included in the analysis. For example, in Figure 2c the model entails $\rho_{23.1} = 0$ because X_1 is a common cause of X_2 and X_3. In contrast, $\rho_{23.1} \neq 0$ for the models in Figure 2a and 2b because of the direct link connecting X_2 and X_3.

We can determine from a model's equations if it entails a zero partial correlation. However, a more intuitive method involves a path analysis on the graph using the concept of a *trek*. Glymour et al. [8] define a trek between two variables i and j as either a directed path from i to j (or j to i), or as a pair of paths from a third variable u, such that there is a directed path from u to i and from u to j with only u in common. If a variable k appears in all treks between i and j, and either every trek from k to j is a directed path from k to j, or every trek from k to i is a directed path from k to i, then the partial correlation of i and j controlling for k (and only k) is zero ($\rho_{ij.k} = 0$) [8]. In Figure 1 the partial correlation of dspA and PBS given NBLA is zero because NBLA is between them on the pathway and the trek between NBLA and PBS is a direct path in the proper orientation. Similarly, the partial correlation of NBLA and cpcB given dspA is zero. However, the partial correlation of dspA and Health given NBLA is non-zero because NBLA does not appear in the lower paths (dspA, RR, psbA1 or psbA2, PHOTO, Health).

To determine the partial correlations in the data, we test the significance of the observed value of $r_{ij.k}$ (note that we use ρ for population values and r for observed values on data). Specifically, we test the null hypothesis $H_0 : \rho_{ij.k} = 0$, which has three outcomes depending on the p value and two thresholds α and γ. If $p \leq \alpha$, we say that the null hypothesis is rejected and we accept the alternate $H_a : \rho_{ij.k} \neq 0$. If $\alpha < p < \gamma$, then we say that the status of H_0 is ambiguous. Finally, if $p \geq \gamma$, we accept the null hypothesis.

We compare the result of the hypothesis test to the partial correlation implied by the model. If the null hypothesis is clearly rejected or accepted, there are four possible outcomes:

1. the model entails $\rho_{ij.k} = 0$ and the data implies $\rho_{ij.k} = 0$ (true positive)
2. the model entails $\rho_{ij.k} = 0$ and the data implies $\rho_{ij.k} \neq 0$ (false positive)
3. the model entails $\rho_{ij.k} \neq 0$ and the data implies $\rho_{ij.k} = 0$ (false negative)
4. the model entails $\rho_{ij.k} \neq 0$ and the data implies $\rho_{ij.k} \neq 0$ (true negative)

We make this comparison for every combination and ordering of three variables, and from these we develop the score function

$$score = fp + fn - tp - tn, \tag{4}$$

where tp, tn, fp, and fn are the number of true/false positive/negatives. Ambiguous hypothesis tests do not count as evidence for or against a model.

Partial correlation constraints let one recover much of the structure, as most graphs will imply different constraints. However, there are equivalence classes for which several models with the same undirected link structure have identical constraints. For example, the model $X_1 \rightarrow X_2 \rightarrow X_3$ has equivalent partial correlation constraints to $X_1 \leftarrow X_2 \leftarrow X_3$, and $X_1 \leftarrow X_2 \rightarrow X_3$. The correct direction can often be resolved if there is some additional knowledge about the causal ordering. For example, dspA is a known light sensor in the

photosynthesis model presented earlier, so it must come before other genes in the regulation model.

Implicit in our analysis is the assumption that partial correlations in the data are only zero when they are entailed (faithfulness assumption), i.e., true for all possible values of the link parameters b_{ij}. This eliminates models where the partial correlations are zero only for specific values on the links. For example, this could happen for the model in Figure 2c if b_{13} was exactly equal to $-b_{12}b_{23}$.

3.2.2 Parameter Signs and Correlation

We can use knowledge about the signs of the parameters b_{ij} in the model to predict the sign of correlation between any two variables that we can observe in the data. If two variables are directly connected, such as $X_1 \rightarrow X_2$, then we expect that $sign(\rho_{12}) = sign(r_{12})$. When the variables are not directly connected, we can predict the sign by tracing the links in the trek connecting any two variables and by multiplying the signs. For example, in Figure 1 the sign between dspA and PBS should be negative $(1 \times 1 \times -1)$.

When there are multiple treks between two variables the predicted signs could disagree. In a fully quantitative model, each path would have its own degree of influence based on the magnitude of b_{ij}, and one could sum their effects to determine the outcome. In general, we will not assume reasoning at this fully quantitative level. Thus, to obtain unambiguous predictions, we annotate the model with dominance relations that specify the corresponding pathway sign. The dominant pathways can be either specified by the biologist or learned from data.

Given a model and sign assignments, we can score it against data using the function

$$score = \sum_{ij, i \neq j} f(sign_d(i,j), sign_m(i,j)) \ , \tag{5}$$

where $sign_d(i,j)$ and $sign_m(i,j)$ return the sign of ρ_{ij} predicted by the data and the model, and $f(a,b)$ is a function that returns 0 if a and b are equal, 1 otherwise.

3.3 Revising Regulatory Models to Explain Microarray Data

Given an initial model and data, we use a two-stage process to revise the model. The first stage attempts to revise the model structure to find the correct causal relationships. Given the new structure, the second stage attempts to determine the type of regulation between variables (i.e., signs of b_{ij} parameters).

We view the revision process as carrying out heuristic search through the space of candidate models for a network structure that explains the data better. The starting state is the initial model provided by the biologist. The search operators for generating alternative models are the addition, deletion, and reversal of links between genes and external variables. We evaluate the alternative models with the score function in Equation 4, which examines partial correlation constraints, and move through the model space with greedy hillclimbing.

In addition to providing a starting point for the search, biological knowledge comes into play by constraining the link structures that are permitted. For example, in Figure 1 the link from Light is a signaling pathway that should connect to a light sensor, for which

the only candidate is dspA. Our system supports type constraints between variables, where the beginning and end of a link must be variables of a specific type. In our model, these constraints have the effect of fixing the links to and from the external variables. There are many other ways that biological knowledge can constrain the model. For example, Hartemink et al. [10] takes an alternative approach that uses location information to fix links in a Bayesian network.

In general, current gene expression experiments provide only a few data points to score models. This causes severe problems because, with little data, small changes in the data set or algorithm parameters can produce very different revisions. To address this, we use the bootstrap [4] to determine the stability of the suggested revisions. The bootstrap is a resampling method for estimating statistics that would be difficult to infer analytically. Bootstrapping has been used in phylogenetic trees [5] and Bayesian network inference [6].

Our application differs slightly from these previous uses as we attempt to learn stable changes from an initial model as opposed to stable structures inferred from the data alone. Our technique estimates how frequently a change would be suggested during revision with slightly different data sets of fixed size. In particular, it samples with replacement from a data set of size n to create k new data sets also of size n. For each sample, it carries out the revision process and records the suggested changes, then only accepts changes with repeatability greater than a threshold,[4] where we define *repeatability* as the percentage of the samples in which the revision occurs.

Once structure is learned, the system carries out another another search process to determine whether each link should have a positive or negative sign. If there are few links in the model, it exhaustively checks all possible assignments of + or - to the links; otherwise, it resorts to hill-climbing search starting with the assignment given by the initial model whenever the link is present in both models. The system uses Equation 5 to score each candidate assignment and direct the search procedure.

4 Results

In this section, we discuss the results of applying our method to revising the photosynthesis model on real microarray data. We also describe experiments with synthetic data to understand better the properties of our algorithm.

4.1 Results for Wild Type and Mutant Cyanobacteria

We applied our method to revise the regulation model of photosynthesis for wild type Cyanobacteria from Figure 1 and to construct a model for the mutant dspA, which does not bleach in high light conditions. We have microarray data for both organisms which contain measurements for approximately 300 genes believed to play a role in photosynthesis.[5] For this analysis, we focus on the genes in the original model and do not consider links to other genes. The array data were collected at 0, 30, 60, 120, and 360 minutes after high light conditions were introduced, with four replicated measurements at each time point. We treated RR as an unmeasured variable, and Photo, which represents the structure of the

[4] An alternative would be to present the repeatability numbers directly to the biologist and let her decide.

photosystem, was not measured (although it could be in theory). We treated the observations as independent samples and ignored their temporal aspect, as well as the dependencies among the four replicates.

Our system revised the initial model for wild type Cyanobacteria with 20 bootstrap replicates and a repeatability threshold of 75%. Figure 3 shows the revised model. There are four changes from the original model: removing psbA2, changing the signs of the correlation on links between from RR to psbA1 and cpcB, and changing the sign of the link between PBS and Health.

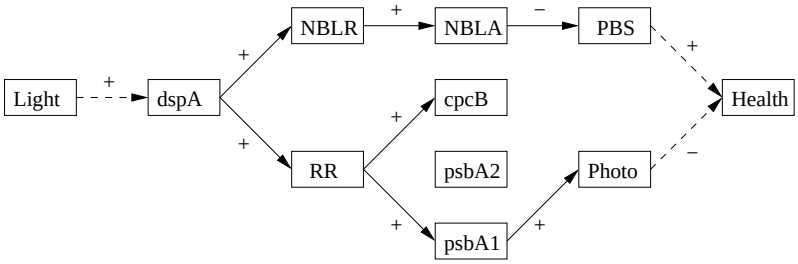

Figure 3: Revised model of photosynthesis in wild type Cyanobacteria.

The revised wild type model dropped all links to the gene psbA2. Discussion with the biologist who proposed the model indicated that the links from RR to the photosystem (psbA1, psbA2, and cpcB) are thought to occur, but the exact configuration and genes involved are uncertain. The presence of one gene product (psbA1) is enough to regulate the structure of the photosynthetic center (Photo), so dropping psbA2 is not problematic.[6] As a check, we can examine correlations of psbA2 with its neighboring genes in the initial model. The gene psbA2 has very low correlations with psbA1 ($r = 0.01$), cpcB ($r = -0.11$), and dspA ($r = -0.21$). In contrast, the genes psbA1 and cpcB are strongly correlated ($r = 0.88$) as expected from their connection through RR.

Although our method suggested several plausible revisions to the wild type model, there were also changes that we did not expect. For example, the revision process changed the sign on the link PBS → Health from negative to positive. The biologist who proposed the model assumed that the light conditions were high enough to cause damage; the revision suggests the opposite, that under high light conditions more PBS is better for the organism. The underlying issue is that the link from PBS to Health is an abstraction that obscures two pathways that compete for dominance, as shown in Figure 4.

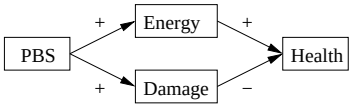

Figure 4: Expansion of the abstract link from PBS to Health.

[5]The data is available at http://www.isle.org/∼sbay/data/cyano.html .

[6]The genes psbA1 and psbA2 both encode variants of the photosystem II D1 protein.

Light provides energy to the organism and this increases viability, but it also damages the organism by increasing the number of oxygen radicals. When light levels are low, the effect of energy dominates. As light levels rise, damage increases and eventually dominates over any gains from energy. The results suggest that the light exposure was not high enough for damage to overcome the benefit from energy.

We also applied our revision process to develop a model that explains why the mutant dspA does not bleach in high light conditions. Presumably, the mutant differs genetically from the wild type organism in at most a few ways, so we used the initial model in Figure 1 as the starting point for revision. The revised model, shown in Figure 5, involves only one change – the removal of the link from dspA to RR.

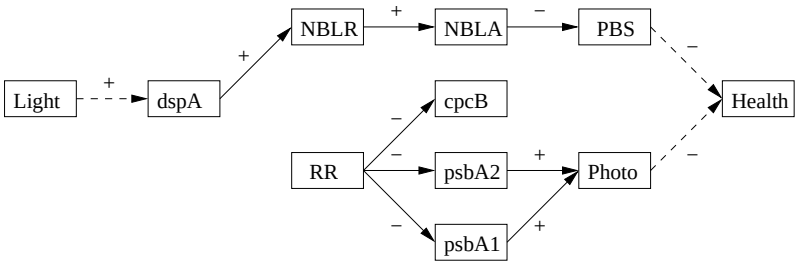

Figure 5: Model of photosynthesis in mutant Cyanobacteria.

Dropping the link from dspA to RR is sensible because the mutant is an experimental knockout in dspA, and thus dspA should not influence other genes in the model. Removing the link moves the model a step closer to an isolated dspA gene. However, the new model does not explain why the mutant fails to bleach in high light conditions. One possibility is that 20 examples do not provide enough statistical power to detect all necessary changes. Specifically, the revised model correctly removed the link from dspA to RR, but did not remove the link from dspA to NBLR. This latter change was suggested in several bootstrap samples, but not frequently enough to meet the 75% repeatability threshold. An additional problem is that the links from RR to psbA2, and psbA2 to Photo are not dropped, as with the wild type model. Again, we believe psbA2 is not removed from the model because of insufficient statistical power.

4.2 Results for Synthetic Data

We used synthetic data to study how well structure could be recovered from incorrect initial models with a small number of samples. We generated data sets of size 20 by treating the structure of the model in Figure 1 as the true model. We assumed values for b_{ij} on the links, and generated e_i according to a random normal distribution ($\mu = 0$, $\sigma = 0.1$). The root causal variable, Light, was assigned a random uniform value between 0 and 1.

We also applied our revision process to synthetic data in which the initial model has been mis-specified by randomly adding or deleting links to the generating model. The results are summarized in Table 1 which lists the number of correct and incorrect suggested revisions given errors in the initial model averaged over 20 runs. The first row represents the situation

in which all variables are observable; the second row represents the situation in which the the synthetic variable corresponding to RR is unobservable.

These results provide an estimate of our method's ability to correct errors in the initial model. For example, the entry in the first row and third column says that if the initial model has two incorrect links, then on average the revision process will correct 1.2 link errors and introduce 0.5 incorrect link changes. The last column represents the empty model, which corresponds to an initial model where nothing is known about the regulatory relations and all variables are unconnected. The number of correct/incorrect revisions are tabulated for a repeatability threshold of 75%. We selected this threshold to control error given a fairly complete initial model. However, this cutoff is too conservative for the empty model where a lower threshold would result in many more correct revisions (on average the top 5.8 suggested revisions are correct when starting with an empty model).

Table 1: Expected number of correct/incorrect revisions on synthetic data.

	errors in initial model					
	0	1	2	4	6	empty
all observable	0/0.4	0.5/0.4	1.2/0.5	1.8/0.3	2.5/0.2	2.1/0.1
RR not observable	0/0.4	0.3/0.2	0.6/0.3	1.3/0.5	1.4/0.2	1.8/0.3

These results suggest that there is enough power to suggest a few revisions reliably, as we found with wild type and mutant Cyanobacteria, even though there were only a small number of samples (20) and unmeasured variables such as RR.

In addition to studying the ability of our algorithm to revise structures such as the photosynthesis regulation model, we also investigated the ability of our algorithm to revise larger models that involve more variables if a greater amount of data were available.

We generated larger models and the corresponding synthetic data by first determining a model structure according to the following procedure. We selected 30 random genes from our real microarray data and used our algorithm to learn a model. We then treated the discovered model as correct and used it to generate new data in the same fashion as in the previous experiment. This approach is similar to a parametric bootstrap.

We then corrupted the model by randomly adding and deleting links between genes (6 changes in total) and measured the ability of our algorithm to suggest correct revisions with varying amounts of data. The results are summarized in Table 2, which lists the average number of correct revisions in the top ten suggestions over five trials. As one might expect, as we increase the number of data points, our algorithm is more likely to suggest correct revisions.

5 Discussion

Although our approach to revising models of gene regulation shows clear promise, we should consider its limitations, as well as its relation to other methods for discovering causal knowledge.

Table 2: Average number of correct suggestions on synthetic data from models with 30 genes. The maximum number of correct revisions is 6. The pooled standard deviation for these results is 0.4.

	Number of data samples		
	50	100	200
correct suggestions	3.4	3.95	4.6

5.1 Limitations

Our approach assumes a linear model that has limited representational power. Although linear models are desirable because they have a small number of parameters, they cannot model combinatorial effects, such as genes X and Y both needing to be highly expressed before Z transcribes. For photosynthesis in Cyanobacteria, the genes were not believed to interact combinatorially and the primary concern was dealing with the small number of samples, making the linear model a natural choice. For other systems that have known combinatorial interactions, we should extend our representation to include interaction terms within the linear framework.

In addition, although our data originated from time-course measurements, we also limited representational power by deliberately choosing not to model time dependent effects for two reasons. First, the data samples were taken far apart in time and we hoped that temporal dependencies would not be significant to the modeling effort at that time scale. Second, proper inference of causal relations in temporal data is an extremely challenging and unsolved problem in microarray data analysis. The main issue is that microarray data are typically sampled with extremely low frequencies (e.g., in our case with a period of at least 30 minutes). The low sampling rate can cause temporal aggregation bias which is well known to lead to spurious causality relationships (e.g., [2, 13]). Essentially, the levels of gene expression between sampling points are unobserved and act as latent variables through which indirectly related variables can have unexplained correlations. These correlations lead to algorithms to incorrectly infer direct causal relationships when none exist in reality.

We restricted the genes that could appear in the model to a small subset of those measured by the microarray chips. The complete set of data contains about 300 variables, from which we used the 11 variables present in the initial model. We restricted the number of variables because we had very few samples, and many variables would have made estimating zero partial correlations unreliable because of the multiple hypothesis testing problem [18]. However, using too few variables means that we may have excluded an important variable from the analysis. Clearly, a tradeoff is involved and we believe a good practical solution is limiting the number of genes to a reasonable set with background knowledge.

Finally, we have focused on mRNA expression levels and did not directly model variables representing biochemical activity, such as the concentration of proteins and their state (e.g., phosphorylated or bound in a complex). Modeling activity at the biochemical level is clearly more realistic, and biologists typically model their regulation system both in terms of mRNA expression and protein activities. However, biochemical activity is not measured by microarrrays and thus the protein levels are generally unobserved.

Partial correlation constraints can distinguish between some, but not all, structures involving unobserved variables [8]. For example, with our photosynthesis regulation model the biologist hypothesized that dspA affects psbA1, psbA2, and cpcB through some unobserved gene RR. However, an alternate hypothesis is that dspA regulates these genes directly. These two hypotheses generate predictions about partial correlations even though RR is unobserved: the model without RR entails $\rho_{cpcB\ psbA1\ .\ dspA} = 0$, while the model with RR entails $\rho_{cpcB\ psbA1\ .\ dspA} \neq 0$. In general, the presence of unobserved variables makes inference about the network structure more difficult and it may not always be possible to distinguish competing models.

5.2 Relation to Bayesian Networks

Linear causal models are closely related to Bayesian networks, which a number of researchers have used to model gene regulation [7, 15, 10, 11, 22]. In fact, a linear causal model is a special case of a Bayesian network that has linear Gaussian conditional densities at each node.

Our method used treks and directed paths to identify zero partial correlations entailed by the model. This approach is very similar to Pearl's [14] notion of *d-separation* in Bayesian networks for determining conditional independence relations. Two variables i and j are d-separated if for all undirected paths between them there is an intermediate variable k such that:

1. k does not have converging arrows (i.e., $\leftarrow k \leftarrow$, $\rightarrow k \rightarrow$, $\leftarrow k \rightarrow$) and k is observed, or

2. k has converging arrows ($\rightarrow k \leftarrow$) and neither k nor its descendents are observed.

In our path analysis, we only considered controlling for a single variable[7] whereas d-separation can be applied when multiple variables are controlled (observed). In addition, treks do not contain paths with converging arrows and thus our analysis to identify partial correlation constraints entailed by the model does not explicitly consider the second condition.

Learning methods for inferring causal Bayesian networks can be divided into two main groups. Constraint-based approaches [8, 16] attempt to find networks whose structures entail the conditional independence relations observed in the data. Note that, for linear models, conditional independence between variables is equivalent to zero partial correlations. Our approach falls into this group, as it attempts to find networks that closely match the observed partial correlation constraints in the data. The other main approach for learning causal Bayesian networks attempts to maximize a Bayesian scoring metric. Methods in this group focus on finding the network model M that produces the best score given the data D, i.e., $P(M|D)$. A central step in computing $P(M|D)$ is determining the likelihood of the data given the model, $P(D|M)$, which is usually decomposed into the score of local models that compute the probability of a variable's observations given its direct causes in the model.

[7] Each additional controlled variable reduces the degrees of freedom available for estimating the partial correlation.

At this point, the advantages and disadvantages of each approach are not completely clear. Constraint-based methods may be sensitive to the test used for conditional independence (zero partial correlations) and violations of test assumptions (e.g., linearity). However, Friedman et al. [7] report that their Bayesian scoring approach is "sensitive to the choice of local model, and in the case of the multinomial model, to the discretization method". Saavedra et al. [16] performed an initial study that attempted to compare constraint-based algorithms [8] with a Bayesian scoring approach [7] on regulation networks developed for the yeast cell cycle. However, they found the methods were difficult to compare because little is known about the true regulatory processes, and differences in handling of missing values and normalization of data can have large effects on the final results, thus masking differences between approaches. In the future, we plan to compare these two general approaches with synthetic data along measures such as the robustness to noise, violations of model assumptions, number of samples, and number of hidden variables.

6 Conclusions and Future Work

In this paper, we have described an approach to combining data-driven search with biological knowledge in order to find better models of gene regulation. We illustrated this method by using it to revise a regulatory model of photosynthesis in Cyanobacteria with expression data.

Our results are encouraging, but we must extend our system in a number of directions to make it a more useful tool for biologists. From the perspective of computational inference, we should expand our analysis techniques to explicitly handle time delay and feedback, both of which are common in gene regulation. One possible approach is to represent interactions between genes with qualitative differential equations. Another issue is incorporating interventional data from knockout experiments into the revision process [15, 22], as so far we have concentrated on analyzing observational data.

From the perspective of modeling domain knowledge, we intend to support many more biological concepts. For example, although biologists often state models in terms of measurable statistical variables, such as gene expression levels, they also describe an organism's behavior in terms of mechanical processes that operate on individual molecules. Karp's [12] work on modeling the tryptophan operon provides one approach to representing such mechanisms. Future work should support the ability to make statistical predictions from such mechanistic models, and thus make better contact with biologists' concepts.

In the longer term, we envision an interactive discovery aide that lets a biologist specify initial models, focus the system's attention on particular data and parts of those models it should attempt to improve, select among candidate models with similar scores, and control high-level aspects of the discovery process.

Acknowledgments

This work was supported by the NASA Ames Director's Discretionary Fund, by the NASA Biomolecular Physics and Chemistry Program, and by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. We thank Arthur Grossman and C.

J. Tu for the initial model, microarray data, and advice on biological plausibility. We thank the anonymous reviewers for their many comments that improved this paper.

References

[1] K. Bollen. *Structural Equations with Latent Variables*. Wiley, 1989.

[2] L. J. Christiano and M. Eichenbaum. Temporal aggregation and structural inference in macroeconomics. Technical Report 60, National Bureau of Economic Research, 1986.

[3] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels durning CNS development and injury. In *Pacific Symposium on Biocomputing*, pages 41–52, 1999.

[4] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[5] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[6] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[7] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3/4):601–620, 2000.

[8] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, 1987.

[9] A. R. Grossman, D. Bhaya, and Q. He. Tracking the light environment by cyanobacteria and the dynamic nature of light harvesting. *The Journal of Biological Chemistry*, 276(15):11449–11452, 2001.

[10] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Combining location and expression data for principled discovery of genetic regulatory models. In *Pacific Symposium on Biocomputing*, pages 437–449, 2002.

[11] S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In *Pacific Symposium on Biocomputing*, pages 175–186, 2002.

[12] P. D. Karp. Hypothesis formation as design. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, 1990.

[13] J. R. McCrorie. Granger causality and the sampling of economic processes. In *Proceedings of the Twelfth Conference on Causality and Exogeneity in Econometrics*, 2001.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.

[15] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, pages S215–S224, 2001.

[16] R. Saavedra, P. Spirtes, R. Ramsey, and C. Glymour. Issues in learning gene regulation from microarray databases. Technical Report IHMC-TR-030101-01, Institute for Human and Machine Cognition, 2001.

[17] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioural Research*, to appear.

[18] J. P. Shaffer. Multiple hypothesis testing. *Annual Review Psychology*, 46:561–584, 1995.

[19] H. Simon. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49:467–479, 1954.

[20] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders. Linear modeling of genetic networks from experimental data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 355–366, 2000.

[21] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. In *Pacific Symposium on Biocomputing*, pages 112–123, 1999.

[22] C. Yoo, V. Thorsson, and G. F. Cooper. Discovery of causal relationships in a gene regulation pathway from a mixture of experimental and observational DNA microarray data. In *Pacific Symposium on Biocomputing*, pages 498–509, 2002.