# Revising Qualitative Models of Gene Regulation

**Kazumi Saito**,[1] **Stephen Bay**,[2] and **Pat Langley**[2]

[1] NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Soraku, Kyoto 619-0237 Japan
saito@cslab.kecl.ntt.co.jp

[2] Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA
sbay@apres.stanford.edu, langley@isle.org

**Abstract.** We present an approach to revising qualitative causal models of gene regulation with DNA microarray data. The method combines search through a space of variable orderings with search through a space of parameters on causal links, with weight decay driving the model toward integer values. We illustrate the technique on a model of photosynthesis regulation and associated microarray data. Experiments with synthetic data that varied distance from the target model, noise, and number of training cases suggest the method is robust with respect to these factors. In closing, we consider related work on inducing causal regulatory models and suggest directions for future research.

## 1 Introduction and Motivation

Like other sciences, biology requires that its models fit available data. However, as the field moves from a focus on isolated processes to system-level behaviors, developing and evaluating models has become increasingly difficult. This challenge has become especially clear with respect to models of gene regulation, which attempt to explain complex interactions in which the expression levels of some genes influence the expression levels of others. A related challenge concerns a shift in the nature of biological data collection from focused experiments, which involve only a few variables, to cDNA microarrays, which measure thousands of expression levels at the same time.

In this paper, we describe an approach that takes advantage of such nonexperimental data to revise existing models of gene regulation. Our method uses these data, combined with knowledge about the domain, to direct search for a model that better explains the observations. We emphasize qualitative causal accounts because biologists typically cast their regulatory models in this form. We focus on model revision, rather than constructing models from scratch, because biologists often have partial models for the systems they study.

We begin with a brief review of molecular biology and biochemistry, including the central notion of gene regulation, then present an existing regulatory model of photosynthesis. After this, we describe our method for using microarray data to improve such models, which combines ideas from learning in neural networks and the notion of minimum description length. Next we report experimental

studies of the method that draws on both biological and synthetic data, along with the results of these experiments. In closing, we consider related work on inducing causal models of gene regulation and directions for future research on this topic.

## 2 Qualitative Causal Models of Gene Regulation

A gene is a fundamental unit of heredity that determines an organism's physical traits. It is an ordered sequence of nucleotides in deoxyribonucleic acid (DNA) located at at specific position on a chromosome. Genes encode functional products, called proteins, that determine the structure, function, and regulation of an organism's cells and tissues.

The gene's nucleotide sequence is used to construct proteins through a multiple stage process. In brief, the enyzme RNA polymerase transcribes each gene into a complementary strand of messenger ribonucleic acid (mRNA) using the DNA as a template. Ribosomes then translate the mRNA into a specific sequence of amino acids forming a protein. Transcription is controlled through the RNA polymerase by transcription factors that let it target specific points on the DNA. The transcription factors may themselves be controlled through signalling cascades that relay signals from cellular or extra-cellular events. Typically, a signalling cascade phosphorylates (or dephosphorylates) a transcription factor, changing its conformation (i.e., physical structure) and its ability to bind to the transcription site. Translation is controlled by many different mechanisms, including repressors binding to mRNA that prevents translation into proteins.

In our work, we focus on revising biological models that relate external cell signals to changes in gene transcription (as measured by mRNA) and, ultimately, phenotype. Specifically, we look at a model of photosynthesis regulation that is intended to explain why Cyanobacteria bleaches when exposed to high light conditions and how this protects the organism. This model, shown in Figure 1, was adapted from a model provided by a microbiologist (Grossman et al., 2001)[1]. Each node in the model corresponds to an observable or theoretical variable that denotes a measurable stimulus, gene expression level, or physical characteristic. Each link stands for a causal 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 expression, which then reduces 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

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

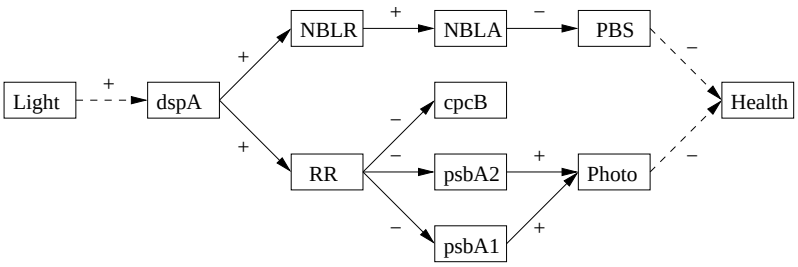


**Figure 1.** Initial model for photosynthesis regulation of wild type Cyanobacteria.

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 a theoretical variable RR, an unspecified gene (or possibly a set of genes) that acts as an intermediary controller.

## 3 An Approach to Revising Qualitative Causal Models

In this paper, we represent causal relationships between variables with a linear model. That is, each quantitative variable $x(i)$ is represented with an equation in the following form:

$$x(i) = \sum_{j=1}^{i-1} A(i,j)x(j) + b(i) \tag{1}$$

where $A(i,j)$ is the causal effect of variable $x(j)$ on $x(i)$, and $b(i)$ is an additive constant. The variables are ordered and variable $x(i)$ can only be influenced by those variables that come before it.

In matrix form we can represent the equations for all $x(i)$, $i = 1..N$, as $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}$. In this formulation $A(i,j) = 0$ if $i \leq j$, where $A(i,j)$ denotes the element in row $i$ and column $j$ of $\mathbf{A}$. This constraint enforces a causal ordering on the variables. A model is completely specified by an ordering of variables in $\mathbf{x}$ and an assignment of values to all elements of $\mathbf{A}$ and $\mathbf{b}$ that satisfy the above constraints.

Let $\mathbf{A}_0$ and $\mathbf{b}_0$ represent the initial model. We transform qualitative models, such as Figure 1, into a matrix $\mathbf{A}_0$ by setting $A(i,j) = 1$ if there is a positive link from variable $j$ to $i$ in the model, $A(i,j) = -1$ if the link is negative, and $A(i,j) = 0$ otherwise. The vector $\mathbf{b}_0$ is set to the zero for all its elements. Given $\mathbf{A}_0$, $\mathbf{b}_0$ and observations on $\mathbf{x}$, we learn new values for $\mathbf{A}$ and $\mathbf{b}$ as follows:

1. Pick an initial ordering for variables in $\mathbf{x}$.
2. Learn the best real valued matrix $\mathbf{A}$ according to a score function that penalizes for differences from $\mathbf{A}_0$, and is subject to the ordering constraints.
3. Swap variables in the ordering and go to step 2 (i.e., perform hill-climbing search in the space of variable orderings). Continue until the score obtained no longer improves.
4. Transform the real matrix $\mathbf{A}$ with the best score into a discrete version with $A(i,j) \in -1, 0, 1$ by thresholding.

Step 1 determines the starting state of the search. Our approach selects a random ordering that is consistent with the partial ordering implied by the initial model.

During Step 2, our method relies on an approach to equation revision that involves transforming the equation $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}$ into a neural network, revising weights in that network, and then transforming the network back into equations in a similar fashion to Saito et al. (2001).

This neural network approach uses a knowledge-based MDL criterion during training to penalize models that differ from the initial model. Specifically, let $\mathbf{w}_0$ be the parameter vector of the neural network that corresponds to the initial model. Our revision task is defined as a problem to find $\mathbf{w}$ fitting to observed data, but it must be reasonably close to $\mathbf{w}_0$. To this end, we consider a communication problem where a sender wishes to transmit a data set to a receiver using a message of the shortest possible length, which is known as the MDL principle proposed by Rissanen (1989). However, unlike the standard MDL criterion, we can naturally assume that the initial model with $\mathbf{w}_0$ is known to the receiver in our revision task. Namely, we try to send message length with respect to $\mathbf{w}_0 - \mathbf{w}$, rather than those of $\mathbf{w}$. Since we can avoid encoding parameter values equal to initial ones, the initial model is preferred. The new parameters $\mathbf{w}_0 - \mathbf{w}$ are regarded as weights of the neural network, and their initial values are set to $\mathbf{0}$. Then, in order to obtain a learning result that is reasonably close to the initial model, the network is trained with weight decay, using a method called the MDL regularizer (Saito & Nakano, 1997).

When there exist some unobserved variables, such as RR in Figure 1, we cannot directly revise links associated with unobserved variables. To cope with

such situations, our method adopts a simple forward-backward estimation based on the initial model. Let $x(i)$ be an unobserved variable, then its value can be forwardly estimated by using an equation, $\hat{x}(i)^{(0)} = \sum_j A(i,j)x(j) + b(j)$. On the other hand, let $S$ be a set of observed variables directly linked from $x(i)$, i.e., $S = \{x(k) : k > i \wedge A(k,i) \neq 0\}$. For $x(k) \in S$, we can obtain an equation for the backward estimation, $x(i) = A(k,i)^{-1}(x(k) - \sum_{j \neq i} A(k,j)x(j) - b(k))$. Thus, let $M$ be the number of elements in $S$, then we have a set of backwardly estimated values, say $\{\hat{x}(i)^{(1)}, ..., \hat{x}(i)^{(M)}\}$. Finally, our method estimates the value of $x(i)$ as their average, by using an equation, $\hat{x}(i) = (M+1)^{-1} \sum_{m=0}^{M} \hat{x}(i)^{(m)}$. Therefore, we can revise all the parameters using these estimated values. Clearly, we can iterate the above pair of procedures, estimation of the unobserved variables and revision of the parameters, although the current implementation makes only one pass.

As stated above, our method performs gradient search through a space of parameters on causal links with weight decay driving the model toward integer values. However, the resulting values are not strictly integers. To overcome this problem, in step 4 we employ a simple thresholding method. After sorting the resulting parameter values to predict one variable $x(i)$, our method divides this sorted list into three portions by using two thresholds, $T_{-1}$ and $T_{+1}$. Namely, parameter value $A(i,j)$ is set to $-1$ if $A(i,j) < T_{-1}$; $+1$ if $A(i,j) > T_{+1}$; 0 otherwise. Note that $T_{-1} \leq T_{+1}$, and we can obtain all possible integer lists within computational complexity of $O(N^2)$, where $N$ denotes the number of parameters. Finally, among these integer lists, our method select the best result which minimizes the MDL cost function defined by $\{0.5 \times (\#samples) \times \log(MSE)\} + \{(\#revised\ parameters) \times \log(N)\}$. Here $MSE$ stands for the mean squared error on the samples. The first term of the cost function is a code length for transmitting data, derived by assuming Gaussian noise for variables, while the second term is a code length for revision information, i.e., multiplication of the number of revised parameters and the code length for an integer to indicate which parameter is revised.

## 4 Experimental Studies of the Revision Method

In this section, we describe experimental studies of our revision method. We take a dual approach of evaluating the system using both natural data obtained from microarrays of Cyanobacteria cultures and synthetic data generated from known mathematical models. Natural data lets us evaluate the biological plausibility of changes suggested by our algorithm. However, because we have an extremely limited number of microarrays, it can be difficult to evaluate the reliability of the suggested revisions even if they appear biologically plausible. Therefore, we also used synthetic data to evaluate the robustness and reliability of our approach. Because we can generate synthetic data from a known model, we can measure the sensitivity and reliability of our algorithm in the presence of complicating factors such as errors in the initial model, small sample sizes, and noise.

### 4.1 Revising the Model of Photosynthesis Regulation

We applied our method to revise the regulatory model of photosynthesis for wild type Cyanobacteria. We have microarray data which includes measurements for approximately 300 genes believed to play a role in photosynthesis. For this analysis, we focus on the genes in the 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 treat both RR and Photo, which represents the structure of the photosystem, as unmeasured variables. We currently treat the data as independent samples and ignore their temporal aspect, along with dependencies among the four replicates.

We implemented our method in the C programming language and conducted all experiments on a 1.3 Ghz Pentium running Linux. Revising the photosynthesis model took 0.02 seconds of CPU time. For each variable, the observed values were normalized to a mean of zero and a standard deviation of one. Figure 2 shows the revised model, which reflects the three changes:

1. dropping the link from dspA to RR;
2. connecting Photo to RR instead of psbA1 and psbA2; and
3. changing the sign of the link from PBS to Health from negative to positive.

The first two changes are difficult to explain from a biological perspective. Because dspA is a light sensor, there should be either a direct or indirect path linking it with the genes cpcB, psbA1, or psbA2. Dropping the link disconnects dspA from those genes and removes it as possible cause. Also, the structure of the photosystem (Photo) is believed to depend on at least one of psbA1 or psbA2, and connecting Photo only to RR removes psbA1 and psbA2 as parents [2].

Changing the sign of the link from PBS to Health is more plausible. The initial model was specified for high light conditions in which excessive light levels damage the organism. However, at lower light levels, increased PBS should aid the organism because it is vital component in energy production. One explanation suggested by the microbiologist is that light levels during the biological experiment may not have been set correctly and were not high enough to reduce health.

### 4.2 Robustness of the Revision Approach

We evaluated the robustness of our approach by generating synthetic data from a known model and varying factors of interest. Specifically, we varied the number of training samples, the number of errors in the initial model, the observability of variables, and the noise level. We expected each of these factors to influence the behavior of the revision algorithm.

[2] The genes psbA1 and psbA2 encode variants of the D1 protein, a necessary and central component of the Photosystem II reaction center (Wiklund et al., 2001).

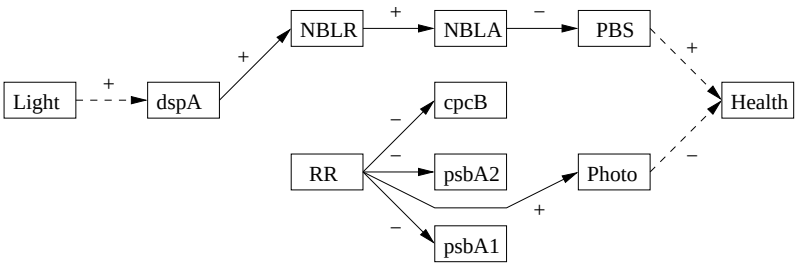


**Figure 2.** Revised model of photosynthesis regulation in Cyanobacteria.

We generated data sets with 25, 50, and 100 examples by treating the structure of the model in Figure 1 as the true model. We assumed that each variable was a linear function of its parents with noise added from a random normal distribution ($\sigma = 0.1$ unless otherwise specified). The root causal variable, Light, has no parents and was assigned a random uniform value between 0 and 1. We generated initial models to serve as starting points for revision by randomly adding links to, or deleting links from, the true model in Figure 1.

Figure 3 shows the experimental results with the $x$ axis representing the number of errors in the initial model and the $y$ axis representing the average number of corrections (i.e., correct changes minus incorrect changes) suggested by the revision process. Each point represents the average of 20 trials. Part (a) shows the ability of our system to correct errors in the model when all variables are observable. In general, there was good performance and even with as few as 25 samples, our system can consistently correct almost all of the errors in the initial model. More training samples tended to improve performance. Part (b) shows the results when a variable, specifically RR, is unobserved. Overall, the performance decreases substantially compared to full observability. However, our system still has enough power to suggest correct revisions improving the model. Parts (c) and (d) show the performance with RR unobserved at greater noise levels with $\sigma = 0.2$ and $\sigma = 0.4$ respectively. The number of corrections is comparable to $\sigma = 0.1$ and suggests that our approach is robust to this type of noise. Note that $\sigma = 0.4$ represents a large noise level in comparison with the range of the variables (e.g., light varies from 0 to 1). Finally, we observe that when the initial model was correct (zero errors), our system never suggested changes to the model.

## 5 Future Research

The results from our experiments on Cyanobacteria data were disappointing, as they were difficult to explain from a biological perspective. However, on synthetic data our system was able to improve incorrect initial models even when there were few training samples, unobserved variables, and noise.

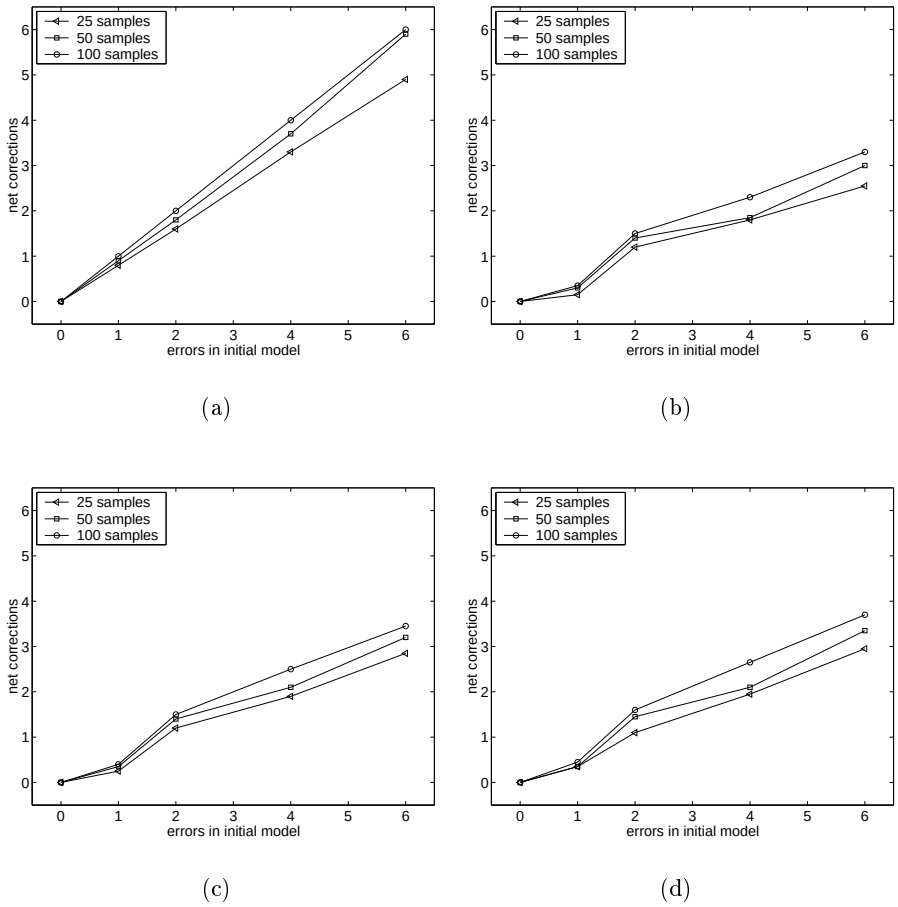


**Figure 3.** Average number of corrections to the initial model for 25, 50, and 100 samples when (a) all variables observed, $\sigma = 0.1$, (b) synthetic variable RR is unobserved, $\sigma = 0.1$, (c) RR unobserved, $\sigma = 0.2$, and (d) RR unobserved, $\sigma = 0.4$.

.

This suggests that our general approach is feasible, but that we may need to address some of the limitations, chosen by design, in our approach. For instance, we modeled the relationships between genes as a linear function. Although linear models are desirable because they have few parameters, they cannot model combinatorial effects among genes or thresholds in which a gene's expression must be above a certain level before it can affect other genes. The neural network approach to revision is not limited to linear models and we could use a more general form to represent relationships between genes.

We also 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. Restricting the number of variables is a tradeoff. Including too many variables for the number of samples makes estimating relationships unreliable because of the multiple hypothesis testing problem (Shaffer, 1995). However, using too few variables increases the likelihood that we may have ignored an important variable from the analysis. Future implementations could minimize this problem by including an operator for adding new genes during the revision process and using domain knowledge to select only the most promising candidates for incorporation into the model.

In addition, we should extend our approach to model revision in various other ways. Since transcriptional gene regulation takes time to occur, a succeeding system should search through an expanded space of models that include time delays on links[3]and feedback cycles. To handle more complex biological processes, it should also be able to represent and revise models with subsystems that have little interaction with each other. Finally, each of these extensions would benefit from incorporation of additional biological knowledge, cast as taxonomies over both genes and regulatory processes, to constrain the search for improved models.

Finally we must test our approach on both more regulatory models and more microarray data before we can judge its practical value. Our biologist collaborators are collecting additional data on Cyanobacteria under more variable conditions, which we predict will provide additional power to our revision method. We also plan to evaluate the technique on additional data sets that we have acquired from other biologists, including ones that involve yeast development and lung cancer.

## 6 Related Work

Although most computational analyses of microarray data rely on clustering to group related genes, we are not the first to focus on inducing causal models of gene regulation. Most research on this topic encodes regulatory models as Bayesian networks with discrete variables (e.g., Friedman et al., 2000; Hartemink, 2002; Ong et al., 2002). Because microarray data are quantitative,

[3] An alternative is to model the regulation between genes with differential equations, possibly with explicit time delays.

this approach often includes a discretization step that may lose important information, whereas our approach deals directly with the observed continuous values.[4] These researchers also report methods that construct causal models from scratch, rather than revising an initial model, though some incorporate background knowledge to constrain the search process.

An alternative approach represents hypotheses about gene regulation as linear causal models, which relate continuous variables through a set of linear equations. Such systems evaluate candidate models in terms of their ability to predict constraints among partial correlations, rather than their ability to predict the data directly. Within this framework, some methods (e.g., Saavedra et al., 2001) construct a linear causal model from the ground up, whereas others (e.g., Langley et al., 2002) instead revise an initial model, as in the approach we report here. One advantage of this constraint-based paradigm is that it can infer qualitative models directly, without the need to discretize or fit continuous parameters. In contrast, our technique combines search through a parameter space with weight decay to achieve a similar end.

We should also mention approaches that, although not concerned with gene regulation, also construct causal models in scientific domains. One example comes from Koza et al. (2001), whose method formulates a quantitative model of metabolic processes from synthetic time series about chemical concentrations. Another involves Zupan et al.'s (2001) GenePath, which infers a qualitative genetic network to explain phenotypic results from gene knockout experiments. Mahidadia and Compton (2001) report an interactive system for revising qualitative models from experimental results in neuroendocrinology. Finally, our approach to revising scientific models borrows ideas from Saito et al. (2001), who transform an initial quantitative model into a neural network and utilize weight learning to improve its fit to observations.

## 7 Conclusions

In this paper, we characterized the task of discovering a qualitative causal model of gene regulation based on data from DNA microarrays. Rather than attempting to construct the model from scratch, we instead assume an existing model has been provided biologists who want to improve its fit to the data. These models require a causal ordering on variables, links between variables, and signs on these links. We presented an approach to this revision task that combines a hill-climbing search through the space of variable orderings and a gradient descent search for weights on links, with the latter using a weight decay method guided by minimum description length to drive weights to integer values.

We illustrated the method's behavior on a model of photosynthesis regulation in Cyanobacteria, using microarray data from biological experiments. However, our experimental evaluation also relied on synthetic data, which let us vary systematically the distance between the initial and target models, the amount of

[4] Imoto et al. (2002) report one way to induce quantitative models of gene regulation within the framework of Bayesian networks.

training data available, and the noise in these data. We found that the method scaled well on each of these dimensions, which suggests that it may prove a useful tool for revising models based on biological data. We noted that our approach has both similarities to, and differences from, other recent techniques for inducing causal models of gene regulation. We must still evaluate the method on other data sets and extend it on various fronts, but our initial experiments on synthetic data have been encouraging.

## Acknowledgements

This work was supported by the NASA Biomolecular Systems Research Program and by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. We thank Arthur Grossman, Jeff Shrager, and C. J. Tu for the initial model, for microarray data, and for advice on biological plausibility.

## References

Friedman, N., Linial, M., Nachman, I., & Peer, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, *7*, 601–620.

Grossman, A. R., Bhaya, D., & He, Q. (2001). Tracking the Light Environment by Cyanobacteria and the Dynamic Nature of Light Harvesting. *The Journal of Biological Chemistry*, *276*, 11449–11452.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., & Young, R. A. (2002). Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. *Pacific Symposium on Biocomputing*, *7*, 437–449.

Imoto, S., Goto, T., & Miyano, S. (2002). Estimation of Genetic Networks and Functional Structures Between Genes by using Bayesian Networks and Nonparametric Regression. *Pacific Symposium on Biocomputing*, *7*, 175–186.

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., & Keane, M. A. (2001). Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing*, *6*, 434–445.

Langley, P., Shrager, J., & Saito, K. (in press). Computational discovery of communicable scientific knowledge. In L. Magnani, N. J. Nersessian, & C. Pizzi (Eds), *Logical and computational aspects of model-based reasoning*. Dordrecht: Kluwer Academic.

Mahidadia, A., & Compton, P. (2001). Assisting model-discovery in neuroendocrinology. *Proceedings of the Fourth International Conference on Discovery Science* (pp. 214–227). Washington, D.C.: Springer.

Ong, I. M., Glasner, J., & Page, D. (2002). Modeling Regulatory Pathways in E.Coli from Time Series Expression Profiles. *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology*.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific, Singapore.

Saavedra, R., Spirtes, P., Scheines, R., Ramsey, J., & Glymour, C. (2001). Issues in Learning Gene Regulation from Microarray Databases. (Tech. Report No. IHMC-TR-030101-01). Institute for Human and Machine Cognition, Unversity of West Florida.

Saito, K., Langley, P., Grenager, T., Potter, C., Torregrosa, A., & Klooster, S. A. (2001). Computational revision of quantitative scientific models. *Proceedings of the Fourth International Conference on Discovery Science* (pp. 336–349). Washington, D.C.: Springer.

Saito, K., & Nakano, R. (1997). MDL regularizer: a new regularizer based on MDL principle. *Proceedings of the 1997 International Conference on Neural Networks* (pp. 1833–1838). Houston, Texas.

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review Psychology*, *46*, 561–584.

Wiklund, R., Salih, G. F., Maenpaa, P., & Jansson, C. (2001) Engineering of the protein environment around the redox-active TyrZ in photosystem II. *Journal of European Biochemistry* , *268*, 5356–5364.

Zupan, B., Bratko, I., Demsar, J., Beck, J. R., Kuspa, A., Shaulsky, G. (2001). Abductive inference of genetic networks. *Proceedings of the Eighth European Conference on Artificial Intelligence in Medicine*. Cascais, Portugal.