

Inducing Biologically Plausible Models from Temporal Expression Data

Pat Langley,¹ Dileep George,¹ Kazumi Saito²
Jeff Shrager,¹ and Stephen Bay¹

¹ Computational Learning Laboratory, CSLI
Stanford University, Stanford, California 94305 USA
langley@csl.i.stanford.edu

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

Abstract. We address the task of inducing biological models from time-series data on gene expressions and background knowledge about candidate biological processes. We describe IPM, an algorithm for inducing quantitative process models from such input, and we demonstrate its use on data and knowledge about the regulation of photosynthesis in Cyanobacteria. We also report experiments with synthetic data on similar problems that study the number of samples needed to find the correct model parameters. In closing, we discuss related work on modeling gene regulation and suggest directions for future research in this area.

1. Introduction and Background

Microbiology aims to understand the mechanisms by which organisms survive, grow, and reproduce. Like other sciences, it collects observations, identifies recurring phenomena, and attempts to explain these phenomena using existing knowledge. Biologists have made great strides in explaining metabolism, energy storage, and related mechanisms in terms of chemical reactions among proteins and other molecules. However, as yet they have only a limited understanding of how these mechanisms are *regulated* so that they become more or less active under different conditions.

In this paper, we describe a computational approach to elucidating such regulatory models. We take advantage of the relatively new technology of cDNA microarrays, which let one measure simultaneously the expression levels of many genes. If one presents the organism with some external stimulus, such as light, and takes samples over time, then one can obtain time-series data about the covariation of expression for many different genes, which should provide hints about their regulation.

As we recount later, the computational biology literature has reported a variety of formalisms for representing regulatory models and methods for inducing them from data. What most approaches lack is some way to encode existing biological knowledge and using it to constrain search through the model space. As a result, models generated by these methods make little contact with domain

concepts, which makes them less comprehensible to biologists. In contrast, the approach we report responds directly to this challenge.

We begin by presenting a motivating problem – the regulation of photosynthesis – and reviewing some experimental results in this area that demand explanation. After this, we propose a formalism for stating quantitative models in this area, which we illustrate with a specific example. Next we present a related formalism for encoding background knowledge about biological processes, then turn to a system for inducing process models from this domain knowledge and time-series data. We demonstrate this method’s behavior on both natural data from the experiment, to show its relevance, and synthetic data, to measure its robustness. We conclude with a review of other approaches to inferring regulatory models and proposals for future work on this topic.

2. A Motivating Problem: Photosynthesis Regulation

Without doubt, photosynthesis is one of the most important mechanisms in the operation of the Earth ecosystem. This process harnesses light energy to produce plant growth, generates the oxygen that we breathe, and removes the carbon dioxide that we produce through natural and artificial means. Thus, a deeper understanding of photosynthesis, and the factors that influence it, would improve our ability to explain and predict crucial changes in our environment.

Photosynthesis is a complex combination of reactions that are catalyzed by a system of protein complexes, most of which are bound into the thylakoid membrane of the chloroplasts of higher plants. There are two sets of reactions, referred to as ‘light’ and ‘dark’. The former, which operate only in the light, use absorbed light energy to produce a variety of biochemical species, which are in turn used by the remainder of the cell as energy. The ‘dark’ reactions, which do not require light, use some of the energy produced by light reactions to combine CO_2 molecules into sugars, which are then either used to produce cellular energy and other products or stored for later utilization.

One side effect of the normal photosynthetic reaction is the creation of ‘reactive oxygen species’ (ROS), which can be very damaging to cellular components, especially those in the photosynthetic apparatus. Cells appear to have systems that aim to minimize creation of ROS, that ‘clean up’ or neutralize ROS, and for repairing damage. For these and other reasons, the complex network of mechanisms for energy production, storage, and utilization in cells includes many regulatory controls.

Although the biochemical reactions involved in photosynthesis, and the general shape of its regulation, are fairly well understood, the details of regulatory signals and mechanisms remain obscure. Biologists know about a variety of abstract regulatory mechanisms that could affect photosynthetic activity, such as signal transduction and transcription, but they are uncertain about which ones are responsible and the detailed forms in which they occur. For instance, the protein produced during translation is known to degrade, but it remains unclear whether this takes place at a constant rate or whether it is regulated.

To further elucidate the details of photosynthesis regulation, Labiosa et al. (2003) carried out an experiment with Cyanobacteria, a unicellular organism, under simulated naturalistic conditions. In particular, they constructed a cyclostat that replicated the light variations that occur with the 24-hour day-night cycle.³ Samples of the organism were collected at times equivalent to 2 AM, 8 AM, 10 AM, noon, 2 PM, 6 PM, and midnight. These were analyzed using cDNA microarray technology to measure mRNA levels for 3000 genes in each sample.

Figure 1 shows the temporal behavior of the 17 genes that were most highly correlated with light intensity. Inspection revealed that each had been implicated in photosynthesis previously, which makes biological sense. However, the shape of their curves (given in logarithmic scale) is somewhat unexpected. Expression levels are low at night, increase rapidly when the sun rises, and decrease again after sunset, but they also exhibit a substantial drop around noon. An adequate model of these genes' regulation should account for all of these regularities in at least qualitative terms, and preferably in quantitative ones as well.

Moreover, in addition to reproducing the shape of these expression curves, an acceptable model of gene regulation should also be consistent with existing knowledge about both photosynthesis and more general biological mechanisms. These requirements set the stage for the coming sections, in which we consider the representation, simulation, and induction of such models for gene regulation.

3. Representing Dynamical Models of Gene Regulation

Before we can assist biologists in constructing models of gene regulation, we must select some formalism in which to represent candidate models. Because biology does not have a tradition, like physics and chemistry, of formal notations, most work along these lines has borrowed frameworks from other fields like computer science, electrical engineering, and physics.

Only some of these formalisms can characterize the behavior of dynamical systems that change over time. These include Boolean networks (e.g., Shmulevich et al., 2002), dynamic Bayesian networks (e.g., Imoto et al., 2002), differential equations (e.g., Tomita et al., 1999), and Petri networks (e.g., Matsuno et al., 2002). Despite their representational power, these frameworks make limited contact with biologists' established concepts, though some fare better along this dimension than others.

The problem is that biologists' papers and talks repeatedly make informal reference to *processes* that operate within living organisms. Research in artificial intelligence has produced formalisms that cast models as sets of interacting processes to explain dynamical behavior, with Forbus' (1984) qualitative process theory being a notable example. This offers a notation for biological mechanisms, but it focuses on qualitative simulations that predict only the directions in which continuous variables change over time.

³ This device was built, and the study was run, in the Carnegie Institute of Washington's Department of Plant Biology.

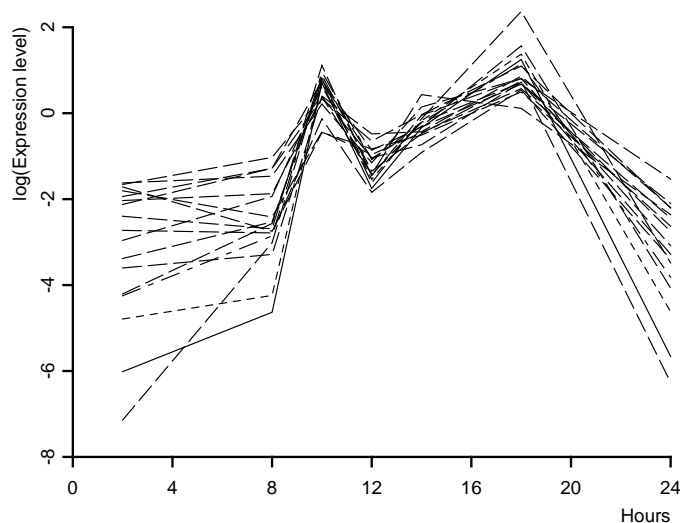


Fig. 1. Observed expression levels of 14 Cyanobacteria genes over a 24-hour period.

We have developed a hybrid representation that embeds numeric equations within the qualitative structures provided by Forbus' approach. A model consists of a set of biological processes, each of which describes the quantitative relations among two or more variables that are cast as one or more algebraic or differential equations. Each process may also include arithmetic conditions on quantitative variables that specify when it is active. Such a quantitative process model must refer to some measurable variables, but it may also include unobservable, theoretical terms.

For example, Table 1 shows one possible model of the expression phenomena from Figure 1. This specifies six quantitative variables – light intensity, energy in the system (redox), rate of mRNA transcription, and the concentrations of mRNA, photosynthetic protein, and reactive oxygen species (ROS). Only two of these variables – light and mRNA – are directly observable, with the remainder being theoretical terms that are biologically plausible.

The model incorporates seven distinct processes. Photosynthesis combines light with proteins to produce energy or redox, but it also increases ROS as a side effect. The photo_translation process increases the concentration of photosynthetic proteins, with the increase depending on the concentration of mRNA. However, another process, protein_degradation_ros, leads to a reduction in both protein and ROS concentration. A fourth process, mRNA_transcription, increases the mRNA concentration by an amount controlled by the variable transcription_rate, which is in turn influenced by two other processes. The first, regulate_light, states that the rate is directly proportional to light, whereas the other process, regulate_redox, states that it is inversely proportional to redox, which is itself reduced. A final process, mRNA_degradation, states that the mRNA concentration decreases by a fixed proportion on every time step.

Table 1. A quantitative process model for photosynthetic regulation.

```
model Photo_Reg;
variables light, mRNA, photo_protein, ROS, redox, transcription_rate;
observables light, mRNA;
process photosynthesis;
  equations  $d[\textit{redox}, t, 1] = 0.01 * \textit{light} * \textit{photo\_protein}$ ;
            $d[\textit{ros}, t, 1] = 0.02 * \textit{light} * \textit{photo\_protein}$ ;
process photo_translation;
  equations  $d[\textit{photo\_protein}, t, 1] = 0.5 * \textit{mRNA}$ ;
process protein_degradation_ros;
  conditions photo_protein > 0;
  equations  $d[\textit{photo\_protein}, t, 1] = -0.01 * \textit{ROS}$ ;
            $d[\textit{ROS}, t, 1] = -0.01 * \textit{ROS}$ ;
process mRNA_transcription;
  equations  $d[\textit{mRNA}, t, 1] = \textit{transcription\_rate}$ ;
process regulate_light;
  equations  $\textit{transcription\_rate} = 0.8 * \textit{light}$ ;
process regulate_redox;
  conditions redox > 0;
  equations  $\textit{transcription\_rate} = -0.5 * \textit{redox}$ ;
            $d[\textit{redox}, t, 1] = -0.03 * \textit{redox}$ ;
process mRNA_degradation;
  equations  $d[\textit{mRNA}, t, 1] = -0.01 * \textit{mRNA}$ ;
```

Like any model, this example makes important simplifying assumptions. For instance, it refers to a single, aggregate measure of mRNA rather than to the amounts for individual genes, and does the same for photo_protein and transcription_rate. Photosynthesis is treated as a single process, rather than as the complex set of activities that we know it involves, and the processes of transcription, degradation, and transcription regulation are abstracted in a similar way. Also, the component processes are all plausible biologically, but some are more so than others. For instance, we know that transcription is regulated and that both protein and mRNA can degrade, but not the details of these activities.

Nevertheless, given such a quantitative process model, we can simulate it to make predictions about how variables will change over time. This involves compiling the process notation into a set of linked algebraic and differential equations, giving them initial values for some variables, and invoking numerical approximation techniques to calculate values for each successive time step. The only complication beyond established methods is that, because conditions can become true or false, one may need to use different equations on each time step. Otherwise, the simulation process is relatively straightforward. However, finding a model that can generate the observed trajectory is a difficult task; in fact, the model in Table 1 provides a poor fit to the data. We would like computational tools that can search the space of model structures and their parameters, ideally taking advantage of biological domain knowledge, to which we now turn.

Table 2. Seven generic processes for gene regulation.

<p>generic process photosynthesis; variables $L\{light\}, P\{photo_protein\}, R\{redox\}, S\{ros\}$; parameters $alpha [0, 1], beta [0, 1]$; equations $d[R, t, 1] = alpha * L * P$; $d[S, t, 1] = beta * L * P$;</p>	
<p>generic process automatic_degradation; variables $C\{concentration\}$; parameters $gamma [0, 1]$; conditions $C > 0$; equations $d[C, t, 1] = -1 * gamma * C$;</p>	
<p>generic process controlled_degradation; variables $D\{concentration\}, E\{concentration\}$; parameters $delta [0, 1]$; conditions $D > 0, E > 0$; equations $d[D, t, 1] = -1 * delta * E$; $d[E, t, 1] = -1 * delta * E$;</p>	
<p>generic process translation; variables $P\{photo_protein\}, M\{mRNA\}$; parameters $rho [0, 10]$; equations $d[P, t, 1] = rho * M$;</p>	<p>generic process transcription; variables $M\{mRNA\}, R\{rate\}$; equations $d[M, t, 1] = R$;</p>
<p>generic process regulate_one; variables $R\{rate\}, S\{signal\}$; parameters $mu [-1, 1]$; equations $R = mu * S$;</p>	<p>generic process regulate_two; variables $R\{rate\}, S\{signal\}$; parameters $nu [-1, 1], pi [0, 1]$; equations $R = nu * S$; $d[S, t, 1] = -1 * pi * S$;</p>

4. Encoding Biological Background Knowledge

A key characteristic of model just described was that it is *explanatory*. An explanation moves beyond a simple description of observations to account for them in terms of other, more basic structures or processes. The explanatory referents are typically unobservable in the current situation, but they make contact with known, familiar mechanisms. The automated construction of such explanatory models requires that we represent the background knowledge to which they refer.

To this end, we utilize the notion of *generic processes*. These are similar in spirit to the specific processes that appear in a model, in that they specify equations and activation conditions, but they do not commit to particular variables or parameter values. Table 2 presents seven generic processes for the domain of plant biochemistry, most of which have direct analogs in Table 1.

Note that each generic process includes a set of generic variables, along with type information that constrains the specific variables against which they can match. Each structure also includes the names of parameters that appear in conditions or equations, along with upper and lower bounds on their values. For instance, the generic process `regulate_two` involves one variable, `R`, that must be

a rate, and another, S , that must be a signal (say light, redox, or ROS), and it refers to two parameters, one of which (π) must fall between zero and one.

Some generic processes are more specific than others. For example, those for photosynthesis, transcription, and translation effectively refer to specific variables, and are generic only in not committing to parameter values. Others, like those for degradation and regulation, refer to classes of variables and can be instantiated in different ways. This lets us encode uncertainty about which variables are actually involved in these processes, but still supports the constrained search for specific models.

Zytkow (1990) has distinguished between general laws or processes that occur in a domain and *models* that hold for a specific situation, with models being cast in terms of known laws or processes. For example, Ohm's and Kirchoff's laws describe general knowledge about the behavior of electric circuits, but they must be combined in particular ways to characterize a specific device. Our generic processes play the role of general biological laws, and we can use them to construct specific models of gene regulation.

5. Inducing Dynamical Models from Time-Series Data

Taken together, time-series data about gene expressions and generic biological processes provide us with the raw material to construct regulatory models. This task is an instance of what we have called *inductive process modeling* (Langley et al., in press). The goal of process model induction is to generate a specific process model, like the one in Table 1, that makes reference to known generic processes and that fits the trajectories of observed variables. Such a model is explanatory, rather than purely descriptive, because it refers to unobserved variables and processes. Moreover, we hold that such a process model will be understandable to domain scientists because it is cast in terms of familiar concepts.

In our current problem, the data concern the expression levels of various genes over time, as shown in Figure 1, along with the associated light intensities. The background knowledge includes plausible forms for processes like photosynthesis, transcription, translation, and degradation, like those in Table 2, including type constraints on their variables and bounds on their parameters. The target is a model like that in Table 1, which contains variants of these generic processes that commit to specific variables and their parameter values. Ideally, this specific model should generate trajectories that match the training data and make accurate predictions about future values.

We have implemented an algorithm, IPM, that decomposes the task of inductive process modeling into two subproblems. The first stage involves a constrained exhaustive search through the space of model structures. To this end, the system finds all ways to instantiate the generic processes with known specific variables that are consistent with the type constraints. Some 24 instantiated processes are generated in this manner from the background knowledge about photosynthesis and gene regulation presented earlier. IPM then composes these instantiated components in all possible ways that involve less than N processes,

Table 3. Model for photosynthetic regulation and initial values induced by IPM.

```
model Photo_Reg;
variables light, mRNA, photo_protein, ROS, redox, transcription_rate;
observables light, mRNA;
initials mRNA = 0.253, photo_protein = 0.836, ROS = 0.059, redox = 0.361;

process photosynthesis;
  equations  $d[\textit{redox}, t, 1] = 0.0155 * \textit{light} * \textit{photo\_protein}$ ;
            $d[\textit{ros}, t, 1] = 0.019 * \textit{light} * \textit{photo\_protein}$ ;

process photo_translation;
  equations  $d[\textit{photo\_protein}, t, 1] = 7.539 * \textit{mRNA}$ ;

process automatic_degradation1;
  conditions photo_protein > 0;
  equations  $d[\textit{photo\_protein}, t, 1] = -1 * 1.905 * \textit{photo\_protein}$ ;

process controlled_degradation1;
  conditions redox > 0, ros > 0;
  equations  $d[\textit{redox}, t, 1] = -1 * 0.0003 * \textit{ros}$ ;
            $d[\textit{ros}, t, 1] = -1 * 0.0003 * \textit{ros}$ ;

process mRNA_transcription;
  equations  $d[\textit{mRNA}, t, 1] = \textit{transcription\_rate}$ ;

process regulate_one_1;
  equations  $\textit{transcription\_rate} = 0.938 * \textit{light}$ ;

process regulate_two_2;
  equations  $\textit{transcription\_rate} = 1.203 * \textit{redox}$ ;
            $d[\textit{redox}, t, 1] = -1 * 0.0002 * \textit{redox}$ ;
```

removing candidates that omit any of the observed variables. For example, when $N = 7$, this scheme produces 288 model structures.

Each such candidate specifies the model's variables and their causal relationships, but it does not include the values for the parameters. Thus, IPM's second stage carries out a gradient descent search through the parameter space defined by each model structure. This search is bounded by the constraints each generic process places on its parameter values, giving a hypercube within which acceptable values can fall. Earlier versions of IPM invoked a version of the Newton method to fit parameters, combined with multiple restarts to mitigate problems with local optima. The current implementation instead utilizes a second-order gradient-descent method that has been described by Saito and Nakano (1997).

Recall that the example model in Figure 2 includes a number of unobserved variables, some of which occur in the left-hand sides of differential equations. This means that, in addition to finding values for the parameters in each process, IPM must also infer the initial values for each such variable. To this end, the system simply treats these as additional parameters that must be fit by the gradient descent mechanism. Elsewhere (Langley et al., in press) we have evaluated this capability on synthetic data, and also shown that one can use a similar approach to induce the thresholds that appear in conditions on processes.

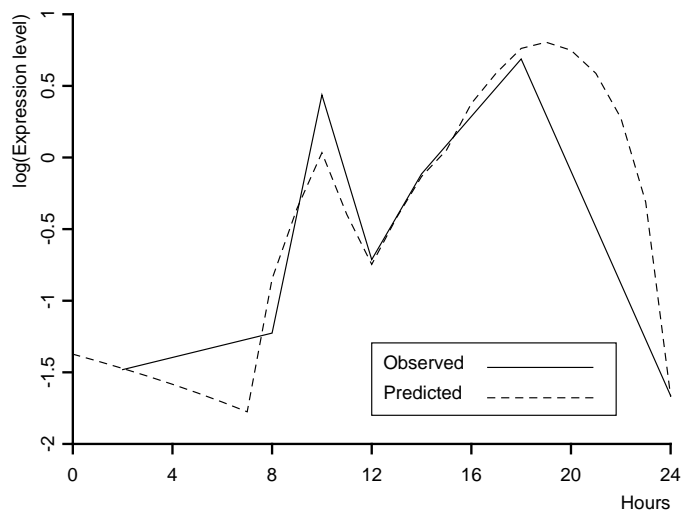


Fig. 2. Predicted and observed levels of average gene expression over a 24-hour period.

To demonstrate that IPM can produce reasonable models of the processes that govern gene regulation, we provided it with the background knowledge from Table 2 and time-series data from the 1cyclostat study. However, because we had only seven samples, we did not attempt to construct a model that predicted separate expression levels for each of the 14 genes. Instead, we averaged the results for these genes at each time step and use the resulting means as the training set for model induction. We also told the system that candidate models should include the observable variables light and mRNA, the unobservable variables photo_protein, ROS, redox, and transcription_rate, and the types for each one.

The process model that IPM generated from these data, shown in Table 3, has similarities and differences from the model presented earlier in Table 1. The new model includes processes for photosynthesis, translation, and transcription, but this is hardly surprising, since their variable types are so constrained as to almost demand their inclusion. More interesting was the inclusion of automatic degradation for photosynthetic proteins, degradation for redox controlled by ROS, and the absence of any degradation process for mRNA. Both models included two distinct processes for regulating transcription_rate, one involving light and the other relying on redox.

Figure 2 shows the trajectories that this model predicts over a 24-hour period, along with the average expression levels computed from the genes measured in the cyclostat experiment. The induced model reproduces the general M shape observed in the data, and the quantitative fit is quite good. The two curves appear to diverge in some places, but this is because the model makes predictions throughout the day, whereas successive observations are simply connected by straight lines. To determine whether the predictions around 6 AM and 10 PM are accurate, we must await further samples from the biologists.

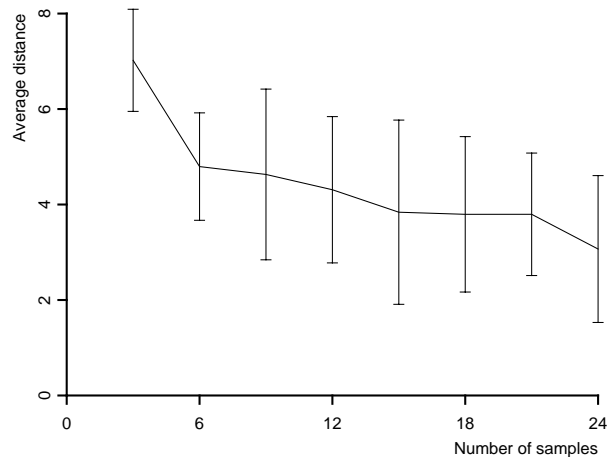


Fig. 3. Distance between induced and target parameters as a function of sampling rate.

Although the process model induced by IPM provides a reasonable fit to the observed expression levels, we cannot determine its correctness because biologists are still uncertain about such issues themselves. However, we can carry out similar runs on analogous data generated from known models to estimate the system’s ability to infer the correct candidate. To this end, we assumed the induced model had the correct structure and used it to generate ten sets of synthetic time-series data, based on different model parameters, over the same 24-hour period, and with the same light levels, as the natural data. Since microarrays are quite noisy, we added ten percent noise to each of these data sets.

A key question is whether the sampling rate used in the actual cyclostat study (seven samples over 24 hours) is high enough to let IPM induce the correct model, so we systematically varied the number of samples (from 3 to 24) provided to the system. Ideally, we would let the system search through the space of 288 model structures for each sampling rate and training set, but this would take an impractical amount of computer time. Instead, we provided the correct model structure and ran the parameter-fitting model instead, on the assumption that, if this finds the correct parameter values, then IPM could find the correct model structure, since it carries out exhaustive search through the structure space.

Our dependent measure was the Euclidean difference between the target parameters and those found by the system, averaged over all parameters in the model and across runs on the ten different training sets. Because we knew the target values and could measure the parametric accuracy directly, we did not need a separate test set. Figure 3 shows the distances as a function of the number of samples. As one might expect, the average distance decreases as more data become available, but the distance is still dropping at one sample per hour. This suggests that we cannot treat the model found on the actual data as especially reliable, and that future experiments on photosynthesis regulation should collect more samples to make the induction task tractable for this biological system.

6. Related and Future Research

Our approach to computational discovery has close connections with other recent efforts on the induction of differential equation models by Todorovski and Džeroski (1997), Bradley et al. (1999), and Koza et al. (2001), which also take advantage of domain knowledge to construct models of dynamical systems. However, we are focused on modeling gene regulation, so we will limit our comments to this area. Much of the research on this topic deals with static models, and thus cannot account for how gene expressions change over time. The remaining work differs from our own by using discrete variables or making little contact with existing knowledge. For instance, Ong et al. (2002) invoke knowledge about promoters to constrain the structure of a dynamic Bayesian network, but they discretize their data. Imoto et al.'s (2002) method induces a quantitative dynamical model, but it makes little use of biological knowledge. Research on Boolean network models (e.g., Shmulevich et al., 2002) suffers from both drawbacks.

Our own previous research on gene regulation (Bay et al., in press) has combined quantitative causal relations with background knowledge in the form of an initial model. However, like most work in this area, our representation of biological processes was quite simplistic (in this case, linear relations) and made limited contact with general biological concepts, such as the distinction among translation, transcription, and degradation. On the other hand, our current approach models regulation only at the aggregate level, whereas most work in this area describes interactions among specific genes.

Although our initial results in this domain are encouraging, it is clear that more work remains ahead. One obvious direction for future research would develop analogous process models for other facets of Cyanobacteria, such as energy storage and utilization, in which specific genes have been implicated. This would require the creation of generic processes for these mechanisms and their use in modeling the expression levels of these genes. We should also expand our studies with synthetic data to better understand how our methods scale to settings with different noise levels, more generic processes, and more complex target models.

In the longer term, we should extend our approach to induce models that explain the expressions of individual genes rather than only their aggregate levels. Such models will have substantially more parameters, so we will also need additional ways to constrain search through the parameter space. Moving in this direction also means we must extend our framework to support larger-scale models of biological systems. A natural approach would rely on hierarchical models that describe the organism in terms of subsystems, which would be based on background knowledge about generic subsystems in addition to generic processes.

In summary, we have described an approach to representing, utilizing, and inducing biological process models from generic background knowledge and time-series data on gene expression and other variables, and we have demonstrated its operation on both natural and synthetic data related to the regulation of photosynthesis. Our formulation has advantages over previous techniques, in that its reliance on domain knowledge should increase interpretability and reduce variance. The results to date are encouraging, and the framework suggests a variety of promising paths to explore in future research.

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 and Kevin Arrigo for use of their laboratory facilities, Lonnie Chrisman for initial models, and Andrew Pohorille for useful discussions.

References

- Bay, S. D., Shrager, J., Pohorille, A., & Langley, P. (in press). Revising regulatory networks: From expression data to linear causal models. *Journal of Biomedical Informatics*.
- Bradley, E., Easley, M., & Stolle, R. (2001). Reasoning about nonlinear system identification. *Artificial Intelligence*, *133*, 139–188.
- Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.
- 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., Mydlowec, W., Lanza, G., Yu, J., & Keane, M. (2001). Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing*, *6*, 434–445.
- Labiosa, R., Arrigo, K., Grossman, A., Reddy, T. E., & Shrager, J. (2003). Diurnal variations in pathways of photosynthetic carbon fixation in a freshwater cyanobacterium. Presented at the *European Geophysical Society Meeting*.
- Langley, P., George, D., & Bay, S. (in press). Robust induction of process models from time-series data. *Proceedings of the Twentieth International Conference on Machine Learning*.
- Matsuno, H., Doi, A., Nagasaki, M., & Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. *Pacific Symposium on Biocomputing*, *5*, 338–349.
- Saito, K., & Nakano, R. (1997). Law discovery using neural networks. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 1078–1083). Yokohama: Morgan Kaufmann.
- Shmulevich, I., Dougherty, E. R., & Zhang, W. (2002). From Boolean to probabilistic Boolean networks as models of gene regulatory networks. *Proceedings of the IEEE*, *90*, 1778–1792.
- Todorovski, L., & Dzeroski, S. (1997). Declarative bias in equation discovery. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 376–384). San Francisco: Morgan Kaufmann.
- Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J., Hutchison, C. (1999). E-CELL: Software environment for whole cell simulation. *Bioinformatics*, *15*, 72–84.
- Żytkow, J. M. (1990). Deriving laws through analysis of processes and equations. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation*. San Francisco, CA: Morgan Kaufmann.