Constructing Explanatory Process Models from Biological Data and Knowledge

Pat Langley, Oren Shiran, Jeff Shrager

Computational Learning Laboratory, CSLI

Stanford University, Stanford, CA 94305 USA

Ljupčo Todorovski

Department of Intelligent Systems, Jozef Stefan Institute

Jamova 39 SI-1000, Ljubljana, Slovenia

Andrew Pohorille

Center for Computational Astrobiology and Fundamental Biology

NASA Ames Research Center, M/S 239-4

Moffett Field, CA 94035 USA

Abstract

We address the task of inducing explanatory models from observations and knowledge about candidate biological processes, using the illustrative problem of modeling photosynthesis regulation. We cast both models and background knowledge in terms of processes that interact to account for behavior. We also describe IPM, an algorithm for inducing quantitative process models from such input, and we demonstrate its use both on photosynthesis and on a second domain, biochemical kinetics. In closing, we consider the generality of our approach, discuss related research on biological modeling, and suggest directions for future work.

Keywords

Computational scientific discovery, Inductive process modeling, Photosynthesis regulation, Biochemical kinetic reactions

1 Introduction and Background

Biology aims to understand the mechanisms by which organisms survive, grow, and reproduce. Like other scientific fields, it collects observations, identifies recurring phenomena, and attempts to explain these phenomena using existing knowledge. However, this endeavor is a complex one, and biologists would benefit from computational tools to assist them in constructing and evaluating their models.

The success of machine learning and data mining in commercial domains has led to increased interest in using similar methods to discover knowledge in biology and other scientific disciplines (Fayyad et al., 1996). However, the best-developed techniques are designed to operate on large data sets and in the absence of background knowledge. Despite rhetoric the contrary,[1] biology remains a data-sparse field, but it has considerable knowledge available to constrain the search for models.

Another drawback of standard induction methods is that they construct descriptive models. These can make accurate predictions on new test cases, which may be sufficient for commercial applications, but biologists typically desire *explanatory* models of behavior. An explanation of some phenomenon is cast in terms of other knowledge, such as structures or processes that are familiar to domain experts.

Finally, traditional induction techniques produce models that are expressed in notations developed by computer scientists, few of which biologists find comprehensible. Even work on inducing causal models, which often have an explanatory flavor, focuses on abstract formalisms that make little contact with concepts from biomedical science. Notations that support the incorporation of domain concepts more directly would presumably be easier to understand and provide additional constraints on model construction.

[1] For example, microarray technology produces many numbers but very few samples, whereas most induction methods assume many of the latter.

In this paper, we describe an approach to inducing biological models that responds to each of these issues. Our models are cast as sets of interacting processes that explain rather than describe the data, and we report a method that constructs such models from background knowledge stated as generic processes, which serve both to constrain search through the model space and make contact with familiar concepts. We illustrate this approach on a problem of central interest to biologists – the regulation of photosynthesis – for which there are limited data but some knowledge. After this, we report results in a second domain – biochemical kinetic reactions. Our approach is a general one that should apply to other biomedical problems, which we discuss in closing along with related research and our plans for future work.

2 The Regulation of Photosynthesis

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. These include 'light' reactions, which operate only in the light and use absorbed energy to produce a variety of biochemical species, which are in turn used by the remainder of the cell as energy. In contrast, 'dark' reactions do not require light but use energy produced by the light reactions to combine CO_2 molecules into sugars, which are then 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 that repair 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 for the observed behavior, as well as the detailed forms in which they occur. For instance, proteins produced during translation are 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. They constructed a cyclodyn which replicated the light variations that occur with the 24-hour day-night cycle.[2] Samples of the organism were collected at nine distinct times throughout the day-night cycle, then analyzed using cDNA microarray technology to measure mRNA levels for 3000 genes in each sample.

Inspection revealed that the 17 genes whose expression levels were most highly correlated with light intensity had each been implicated in photosynthesis previously, which makes biological sense. However, the shape of their curves was somewhat unexpected. Expression levels were low at night, increased rapidly when the sun rose, and decreased again after sunset, but they also exhibited 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. In addition, it should be consistent with existing knowledge about photosynthesis and other biological mechanisms.

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

3 Process Models of Biological Systems

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, yet only some of these formalisms 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., Ong et al., 2002), differential equations (e.g., Tomita et al., 1999), and Petri networks (e.g., Peleg et al., 2002; Matsuno et al., 2002). Despite their representational power, these frameworks make limited contact with established biological concepts.

A fundamental 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.

Instead, we have explored 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 described earlier. This specifies six quantitative variables – light intensity, the concentrations of mRNA, photosynthetic protein, and reactive oxygen species

(ROS), energy in the system (redox), and the rate of mRNA transcription. 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 transcr_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, claims the mRNA concentration decreases by a fixed proportion every time step.

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 protein and transcr_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 trajectories. One complication

is that the conditions on processes may lead different sets of equations to apply during different intervals. Also, if multiple processes influence the same variable, we assume their effects are additive. Otherwise, the simulation process is straightforward. However, finding a model that can generate the observed trajectory is another story, and the model in Table 1 provides a poor fit to the Labiosa et al. data. We would like a computational method that combines knowledge and data to search the space of models, to which we now turn.

4 Encoding Background Knowledge

A key characteristic of the model just described is that it moves beyond a simple description of observations to *explain* 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 incorporate 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 consuming_regulation involves one variable, R, that must be a rate, and another, C, that must be a concentration (such as redox or ROS), and it refers to two parameters, one of which (pi) must fall between zero and one.

Generic processes can co-exist at varying degrees of specificity. 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.

5 Inducing Dynamic Biological Models

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., 2003). The goal 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. The resulting model is explanatory, rather than purely descriptive, because it refers to unobserved variables and processes. Moreover, it should be understandable to domain scientists because it is cast in terms of familiar concepts, much as in Falkenhainer and Forbus' (1991) work on compositional modeling.

In the photosynthesis domain, the data concern the expression levels of photosynthetic genes over time, 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, which stands for *Inductive Process Modeler*, that addresses this task. Its inputs include a set of observable and optional unobservable variables to be included in the model, the types for these variables, a set of generic processes from which to construct candidate models, and a time series of observed values to which models should be fit. As output, the system produces a set of parameterized models ranked their by mean squared error on the training data.

IPM decomposes the task of inductive process modeling into two subproblems, with the first involving 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 14 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 at most U processes, that include all observed variables, and that form a single connected graph. For the run reported below, we used $U = 9$, which produced $2,381$ different model structures.

Each such candidate specifies the model's variables and their causal relationships, but it does not include the values for parameters. Thus, IPM's second stage carries out a search through the parameter space defined by each model structure. This involves a parameter estimation algorithm that uses the entire simulated trajectory. For parameter fitting, IPM invokes a nonlinear least-squares method (Bunch et al., 1993) that utilizes second-order gradient descent. As in other parameter estimation techniques, the system attempts to avoid local minima using multiple restarts. Based on the sum of squared errors, IPM selects the best set of parameters obtained for each model structure. We have found this parameter-estimation method to produce reasonable matches to time series for a variety of domains (e.g., Langley et al., 2003; Todorovski et al., 2005).

Recall that the example model in Table 1 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 terms that must be fit by the parameter estimation module, with the user specifying an acceptable range for each value. Elsewhere (Langley et al., 2003) 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.

To demonstrate that IPM can produce reasonable models of the processes that govern photosynthesis regulation, we provided it with the background knowledge from Table 1 and time-series data from the cyclodyn study. However, because we had only nine samples, we did not attempt to construct a model that predicted separate expression levels for each of the 17 genes. Instead, we averaged the results for these genes at each time step and used 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, along with the optional unobservable variables protein, ROS, redox, and transcr_rate.

The top-ranked process model that IPM generated from these data, shown in Table 3, has similarities to and differences from the one presented earlier in Table 1. The new model includes processes for photosynthesis, translation, and transcription, but this is hardly surprising, since their variable types were so constrained as to demand their incorporation. More interesting was the inclusion of controlled degradation of photosynthetic proteins by ROS, automatic degradation of mRNA, and controlled regulation of transcription rate. The model claims that light affects mRNA transcription, but only indirectly through its influence on redox, rather than through a direct causal link.

Figure 1 shows the expression levels that this model predicts at the times for which samples were taken. Comparison of the average expression levels from the cyclodyn experiment, also given in the figure, with the model's predictions reveals a good quantitative fit that has a correlation coefficient of 0.82. The qualitative match is also good, in that the model reproduces the general M shape that was observed in the study.[3] Equally important, our biologist collaborator believes the model makes sense, as it includes plausible processes for photosynthesis, translation, transcription, regulation, and degradation. However, this is only one opinion and we should also get feedback from other domain experts.

Of course, with only nine samples, we should not be too confident that the model is correct, especially since microarray measurements are often quite noisy. Nor does cross validation seem an appropriate option for evaluating the model, because it is inappropriate for time series that involve different regimes, which we believe holds in this case. Nevertheless, biological experiments typically produce small data sets, and this situation seems unlikely to change in the near future. Our main goal has been to demonstrate that inductive process modeling can construct a model for observed phenomena of scientific interest that is consistent with biological knowledge. We should note that IPM cannot generate models that fit arbitrary curves even in cases where they contain more parameters than the number of observations. The constraints imposed by generic processes, including ranges on parameters and functional forms, should produce relatively low variance even on the small data sets that predominate in biological studies. We believe that, by combining data with knowledge, IPM can produce better results than either in isolation.

[3] One drawback, which did not become apparent until later analysis, is that the model predicts negative values at some intermediate times, which is not biologically possible. This results partly from IPM's assumption that exogenous variables like light are constant unless observed to change. Another model, produced by an earlier version of the system and reported elsewhere (Langley et al., 2004), handled exogenous variables differently and avoided this problem, but clearly future versions of IPM should check candidates against such constraints.

6 Biochemical Kinetics

Although we have shown that IPM can successfully induce explanatory models of photosynthesis regulation, we would also like evidence that its methods have more general applicability. For this purpose, we turned to another key biomedical domain – biochemical kinetics – which studies physiological changes in metabolites over time. In particular, we have examined the glycolysis pathway, which involves the conversion of glucose into pyruvate and which plays an essential role in most life forms. Glycolysis is well understood, with scientists generally agreeing that ten metabolic reactions are responsible. This makes it useful for evaluating our approach to model induction.

To this end, we utilized time-series data collected by Torralba et al. (2003) through an impulse response method that, after a biochemical system has reached steady state, briefly increases the inflow of one substance and measures its effects on others over time. We had access to 14 data points on six metabolites: glucose 6-phosphate ($G6P$), fructose 1,6-biphosphate ($F16BP$), glycerol 3-phosphate ($G3P$), 3-phosphoglycerate ($3PG$), fructose 6-phosphate ($F6P$), and dihydroxyacetone phosphate ($DHAP$). Torralba et al. proposed the model shown in Figure 2, which they produced using a manual analysis method we do not have space to describe here. Their structure differs somewhat from the established glycolysis pathway, primarily because results for $F16BP$ pulses were ambiguous about whether $G3P$ or $3PG$ precedes $G6P$.

Naturally, it seems desirable to automate the construction of biochemical kinetics models that Torralba et al. carried out by hand, and inductive process modeling seems ideally suited for this purpose. Our approach requires background knowledge to define a space of model structures, but readers familiar with the field know biochemists refer to four types of metabolic reaction that appear in their pathway models, which differ in how they affect positive and negative fluxes of the substances involved (Voit et al., 2000). Briefly, the positive flux of a metabolite describes its rate of flow into a reaction pathway, whereas its negative flux instead characterizes its rate of flow out.

Figure 3 depicts the four reaction types, each of which corresponds to a set of ordinary differential equations. An irreversible reaction (a) changes only the positive flux of the reactant $C2$ and the negative flux of $C1$. In contrast, a reversible reaction (b) alters the positive and negative fluxes of both reactants. An inhibition reaction (c) adds an exogenous negative influence on $C2$'s positive flux and $C1$'s negative flux. Finally, an activation reaction (d) includes an exogenous positive influence on $C2$'s positive flux and on $C1$'s negative flux.

However, we must transform these four reaction types into generic processes before we can use them for inductive process modeling. Table 4 shows a generic process library we have developed that incorporates this biochemical knowledge. The generic processes *irreversible*, *reversible*, *inhibition*, and *activation* correspond directly to the reaction types from Figure 3. In addition, the process *flux_combination* states that a metabolite's concentration changes as a weighted sum of its positive and negative fluxes, with each flux term being multiplied by its respective rate.

While formulating the background knowledge for this domain, we encountered a limitation of the process modeling framework as described in previous sections. Recall that, when multiple processes influence the same variable, IPM assumes their effects are additive. However, inspection reveals that, when a metabolite participates in more than one reaction, their effects are multiplicative. To address this issue, we extended the system to accept information for each variable about how it should combine influences on it, specifying for this domain that fluxes have a multiplicative type. This requires slightly more input from the user but increases the system's generality considerably.

When provided with the Torralba et al. data and the generic processes in Table 4, IPM generates 172 distinct model structures. As in the photosynthesis domain, the system attempts to find parameters for each alternative that minimize the squared error between predicted and observed trajectories, then returns a ranked list of parameterized models. Table 5 presents the best-scoring model from this run, whereas Figure 4 depicts it graphically as a reaction pathway.

Figure 5 shows the predictions that this model produces, along with the observations that Torralba et al. reported. The qualitative fit is very good in that it captures the shape of the observed trajectories. For example, the simulated trajectory for $G3P$ peaks and troughs at the same time as does the observed trajectory for this substance. The figure also shows that the flat curves for $F6P$ and $DHAP$ are consistent with the observations. The quantitative fit is also good, with the predictions being highly correlated on average (0.79) with the observations. In fact, this score is lowered substantially by the two nearly flat curves, which have small correlations because they have nearly zero slope.

However, the model structure differs in important ways from both the accepted glycolysis pathway and from the Torralba et al. model. In particular, IPM included no *inhibition* or *activation* processes because we did not provide it with unobserved variables that could plausibly serve as inhibitors and activators. Another possible problem is that reversible reactions dominate, which could produce overfitting because they subsume irreversible reactions as a special case in which some exponents are zero.

Nevertheless, some processes in the induced model have clear correspondences to reactions in the accepted pathway, and the others are biochemically plausible in that there are instances of generic processes relevant to this domain. Thus, although our results are somewhat mixed, we still view the outcome as further evidence that inductive process modeling holds promise for automating the construction of biological models from knowledge and data.

7 Generality, Limitations, and Related Work

Although we have focused here on inducing models in two biomedical domains, the paradigm of inductive process modeling is quite general. Elsewhere (Langley et al., 2003) we have demonstrated that the approach can infer process models of ecosystem behavior, and the basic approach is applicable to any biological domain in which one can identify generic processes with plausible functional

forms and for which quantitative data are available. Here we have emphasized dynamical models and time series, but our methods can handle algebraic models and static data equally well.

One biomedical area that seems a likely candidate is physiology, where there have already been efforts to manually develop quantitative models of behavior using the formalism of differential equations. Another promising topic involves the spread of infectious diseases, for which there already exist numerical models that incorporate ideas from population dynamics. Both fields have considerable knowledge about component processes and functional forms, but data are expensive to collect and the model space is large.

Although our initial results in modeling gene regulation and biochemical kinetics have been encouraging, it is clear that more work still lies ahead. One obvious direction for future research would develop analogous process models for other facets of photosynthesis, such as energy storage and utilization. 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 carry out studies with synthetic data, averaged over different training sets, to better understand how our methods scale to settings with different noise levels, more generic processes, and more complex target models.

More important, we must extend our framework to support larger-scale models of biological systems. Our largest run has involved a space of $3,433$ process model structures, each evaluated on more than 900 observations of one predicted variable and six exogenous variables, which took over 11 hours (averaged across nine runs) on a 2.6 GHz Pentium 4 with one gigabyte of RAM. Both experience and a cursory analysis suggest that the number of candidate models grows exponentially with the number of generic processes. Clearly, we need some way to reduce the size of the space or to constrain search through it substantially.

One promising response would utilize hierarchical models that describe the organism in terms of subsystems and that draw upon background knowledge

about generic subsystems in addition to generic processes. Also, we should adapt our approach to reflect the qualitative nature of many biological models and the fact that biomedical scientists often care only about qualitative fits. In response, we plan to explore methods that induce semi-quantitative process models (e.g., Kay et al., 2000), which can specify ranges on parameters rather than precise values. Such a revised system might direct search based on models' abilities to account for qualitative relations (e.g., one measurement being higher than another) rather than mean squared error.

Any method for inductive learning has the potential for overfitting the training data. We believe that IPM's use of background knowledge reduces this danger over that of knowledge-lean approaches, but experiments with synthetic data indicate that some overfitting can still occur. Bridewell et al. (2005) have described an extension to IPM that mitigates this problem by inducing a number of distinct models from bootstrapped samples and then combining them into a single, more conservative model that includes the most frequent processes. Their approach combines background knowledge with ideas from ensemble methods to reduce overfitting while retaining comprehensibility.

Our approach to biological discovery has close connections with other recent efforts. For example, Bay et al. (2003) present an approach to inducing linear causal models of gene regulation from expression data and background knowledge stated as an initial model. Both Zupan et al. (2001) and Bryant et al. (2001) report systems that infer qualitative genetic networks from biological knowledge and the results of auxotrophic growth experiments, while Mahidadia and Compton (2001) report a similar system that revises qualitative causal models based on experimental results in neuroendocrinology. Ong et al. (2002) describe yet another technique that uses knowledge about promoters to constrain induction of dynamical models for Tryptophan metabolic regulation. However, all have assumed abstract representations that make limited contact with biological concepts like translation, transcription, and degradation.

Another line of research that is closer in its technical details has addressed the induction of quantitative models of dynamical systems. For example, Koza et al. (2001) used evolutionary computation methods to infer the structure and parameters of a metabolic model from time-series data about concentrations. Bradley et al. (1999) describe a different approach to finding differential equation models that draws on knowledge about the behaviors produced by alternative classes of equations. The most similar research comes from Todorovski (2003), whose LaGramge system utilizes domain-specific knowledge, some cast as processes, to guide search for differential equation models. However, his work has focused on environmental domains rather than biomedical ones, such as those we have addressed here.

8 Concluding Remarks

In this paper, we have described an approach to representing, utilizing, and inducing causal biological models. This paradigm – inductive process modeling – supports the construction of explanatory rather than descriptive models, casts these models in terms of familiar biological processes, and takes advantage of background knowledge to constrain search and produce plausible accounts even when there are few samples. We reported a specific system, IPM, that carries out a two-stage search through a space of model structures and their parameters, and we illustrated its operation on background knowledge and time-series data related to the regulation of photosynthesis and biochemical reactions.

The system produced a model for photosynthesis regulation that reproduced both the qualitative shape and the quantitative details of the expression data, while incorporating processes that made biological sense. The small number of samples mean that this result is not entirely reliable, but, we maintain, it is more plausible than ones found without the benefit of background knowledge. In addition, the system produced a reasonable but oversimplified model of biochemical kinetic reactions that matched the trajectories of six metabolites while again remaining consistent with knowledge about the domain.

We argued that inductive process modeling is a general approach that has applications to other biomedical domains like physiology and epidemiology, as well as to other scientific disciplines. However, we also identified some issues that should be addressed in future research, including scaling to larger models, dealing with qualitative phenomena, and mitigating the potential for overfitting. Finally, we noted that our approach incorporates ideas from earlier work on computational discovery that uses domain knowledge to produce interpretable causal models. We believe that such methods, including process model induction, have considerable potential to aid discovery in the biomedical sciences.

Acknowledgements

This work was supported by NSF Grant No. IIS-0326059 and by the NASA Biomolecular Systems Research Program. We thank Arthur Grossman and Kevin Arrigo for use of their laboratory facilities, Rochelle Labiosa for sharing data on photosynthesis, Lonnie Chrisman for developing the initial model of photosynthesis regulation, and Nima Asgharbeygi for contributions to an earlier version of the IPM system. We also thank Stephen Bay, Sašo Džeroski, and Kazumi Saito for many useful discussions.

References

Bay, S. D., Shrager, J., Pohorille, A., & Langley, P. (2003). Revising regulatory networks: From expression data to linear causal models. *Journal of Biomedical Informatics*, *35*, 289–297.

Bradley, E., Easley, M., & Stolle, R. (2001). Reasoning about nonlinear system identification. *Artificial Intelligence*, *133*, 139–188.

Bridewell, W., Bani Asadi, N., Langley, P., & Todorovski, L. (2005). Reducing overfitting in process model induction. *Proceedings of the Twenty-Second International Conference on Machine Learning* (pp. 81–88). Bonn, Germany.

Bryant, C. H., Muggleton, S. H., Oliver, S. G., Kell, D. B., Reiser, P., & King, R. D. (2001). Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions in Artificial Intelligence*, *5 (B1)*, 1–36.

Bunch, D. S.; Gay, D. M.; and Welsch, R. E. (1993). Algorithm 717; subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models. *ACM Transactions on Mathematical Software*, *19*, 109–130.

Falkenhainer, B., & Forbus, K. D. (1991). Compositional modeling: Finding the right model for the job. *Artificial Intelligence*, *51*, 95–143.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). KDD for science data analysis: Issues and examples. *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining* (pp. 50–56). Portland, OR: AAAI Press.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

George, D., Saito, K, Langley, P., Bay, S., & Arrigo, K. (2003). Discovering ecosystem models from time-series data. *Proceedings of the Sixth International Conference on Discovery Science* (pp. 141–152). Saporro, Japan: Springer.

Kay, B., Rinner, B., & Kuipers, B. (2000). Semi-quantitative system identification. *Artificial Intelligence*, *119*, 103–140.

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 *XXVIII General Assembly of the European Geophysical Society*. Nice, France.

Langley, P., George, D., Bay, S., & Saito, K. (2003). Robust induction of process models from time-series data. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 432–439). Washington, DC: AAAI Press.

Langley, P., Shrager, J., Asgharbeygi, N., Bay, S., & Pohorille, A. (2004). Inducing explanatory process models from biological time series. *Proceedings of the Ninth Workshop on Intelligent Data Analysis and Data Mining* (pp. 85–90). Stanford, CA.

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.

Matsuno, H., Doi, A., Nagasaki, M., & Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. *Pacific Symposium on Biocomputing*, *5*, 338–349.

Ong, I. M., Glasner, J. D., & 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* (pp. 241–248).

Peleg, M., Yeh, I., & Altman, R. (2002). Modeling biological processes using workflow and Petri net models. *Bioinformatics*, *18*, 825–837.

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. (2003). *Using domain knowledge for automated modeling of dynamic systems with equation discovery*. Doctoral dissertation, Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

Todorovski, L., Shiran, O., Bridewell, W., & Langley, P. (2005). Inducing hierarchical process models in dynamic domains. *Proceedings of the Twentieth National Conference on Artificial Intelligence* (pp. 892–897). Pittsburgh, PA: AAAI Press.

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.

Torralba, A., Yu, K., Shen, P., Oefner, P., & Ross, J. (2003). Experimental test of a method for determining causal connectivities of species in reactions. *Proceeedings of the National Academy of Sciences*, *100*, 1494–1498.

Voit, E. (2000). *Computational analysis of biochemical systems*. Cambridge: Cambridge University Press.

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* (pp. 304–313). Cascais, Portugal.

Table 1: A process model for photosynthetic regulation.

model photo_regulation;

variables $light, mRNA, protein, ROS, redox, transcr\_rate$;
observables $light, mRNA$;

process photosynthesis;
  equations $d[redox, t, 1] = 1.50 * light * protein$;
  $d[ros, t, 1] = 1.00 * light * protein$;

process photo_translation;
  equations $d[protein, t, 1] = 0.20 * mRNA$;

process protein_degradation_redox;
  conditions $protein > 0, redox > 0$;
  equations $d[protein, t, 1] = -0.05 * redox$;
  $d[redox, t, 1] = -0.05 * redox$;

process mRNA_transcription;
  equations $d[mRNA, t, 1] = transcr\_rate$;

process regulate_light;
  equations $transcr\_rate = 0.80 * light$;

process regulate_redox;
  conditions $redox > 0$;
  equations $transcr\_rate = -2.00 * redox$;
  $d[redox, t, 1] = -1.00 * redox$;

process regulate_mRNA;
  conditions $mRNA > 0$;
  equations $transcr\_rate = -2.00 * mRNA$;
  $d[mRNA, t, 1] = -1.00 * mRNA$;

Table 2: Seven generic processes for plant biochemistry.

process photosynthesis;
variables $L\{light\}, P\{protein\}, R\{redox\}, S\{ROS\}$;
parameters $alpha$ $[0, 1], beta$ $[0, 1]$;
equations $d[R, t, 1] = alpha * L * P$;
$d[S, t, 1] = beta * L * P$;

process controlled_degradation;
variables $D\{degradable\}, E\{degrader\}$;
parameters $delta$ $[0, 1]$;
conditions $D > 0, E > 0$;
equations $d[D, t, 1] = -1 * delta * E$;
$d[E, t, 1] = -1 * delta * E$;

process automatic_degradation;
variables $C\{concentration\}$;
parameters $gamma$ $[0, 1]$;
conditions $C > 0$;
equations $d[C, t, 1] = -1 * gamma * C$;

process translation;
variables $P\{protein\}, M\{mRNA\}$;
parameters rho $[0, 10]$;
equations $d[P, t, 1] = rho * M$;

process transcription;
variables $M\{mRNA\}, R\{rate\}$;
equations $d[M, t, 1] = R$;

process unconsuming_regulation;
variables $R\{rate\}, S\{signal\}$;
parameters mu $[-1, 1]$;
equations $R = mu * S$;

process consuming_regulation;
variables $R\{rate\}, C\{concentration\}$;
parameters nu $[-1, 1], pi$ $[0, 1]$;
equations $R = nu * C$;
$d[C, t, 1] = -1 * pi * C$;

Table 3: Model for photosynthetic regulation induced by IPM.

model photo_regulation;

variables $light, mRNA, protein, ROS, redox, transcr\_rate$;
observables $light, mRNA$;

process photosynthesis;
  equations $d[redox, t, 1] = 8.40 * light * protein$;
  $d[ros, t, 1] = 3.95 * light * protein$;

process photo_translation;
  equations $d[protein, t, 1] = 0.75 * mRNA$;

process protein_degradation_redox;
  conditions $protein > 0, redox > 0$;
  equations $d[protein, t, 1] = -0.89 * redox$;
  $d[redox, t, 1] = -0.89 * redox$;

process mRNA_transcription;
  equations $d[mRNA, t, 1] = transcr\_rate$;

process regulate_light;
  equations $transcr\_rate = 11.94 * light$;

process regulate_redox;
  conditions $redox > 0$;
  equations $transcr\_rate = -8.90 * redox$;
  $d[redox, t, 1] = -7.67 * redox$;

process regulate_mRNA;
  conditions $mRNA > 0$;
  equations $transcr\_rate = -6.58 * mRNA$;
  $d[mRNA, t, 1] = -1.95 * mRNA$;

Table 4: Generic processes for biochemical kinetic reactions.

process flux_combination;
variables $C\{conc\}$, $C_{+flux}\{flux\}$, $C_{-flux}\{flux\}$, $C_{+rate}\{rate\}$, $C_{-rate}\{rate\}$;
equations $d[C,t,1] = C_{+rate} * C_{+flux} + C_{-rate} * C_{-flux}$;

process irreversible;
variables $C1\{conc\}, C1_{-flux}\{flux\}, C2_{+flux}\{flux\}$;
parameters $kinetic\_order1\ [0,1], kinetic\_order2\ [0,1]$;
equations $C1_{-flux} = C1\hat{}kinetic\_order1$;
$C2_{+flux} = C1\hat{}kinetic\_order2$;

process inhibition;
variables $C3_{-flux}\{flux\}, C4_{+flux}\{flux\}, E\{conc\}$;
parameters $kinetic\_order1\ [0,1], kinetic\_order2\ [0,1]$;
equations $C3_{-flux} = E\hat{}(-kinetic\_order1)$;
$C4_{+flux} = E\hat{}(-kinetic\_order2)$;

process activation;
variables $C5_{-flux}\{flux\}, C6_{+flux}\{flux\}, E\{conc\}$;
parameters $kinetic\_order1\ [0,1], kinetic\_order2\ [0,1]$;
equations $C5_{-flux} = E\hat{}kinetic\_order1$;
$C6_{+flux} = E\hat{}kinetic\_order2$;

process reversible;
variables $C7_{+flux}\{flux\}, C7_{-flux}\{flux\}, C8_{+flux}\{flux\}, C8_{-flux}\{flux\}$,
$C7\{conc\}, C8\{conc\}$;
parameters $k\_o1\_1p\ [0,1], k\_o1\_2p\ [0,1], k\_o1\_1n\ [0,1]$,
$k\_o1\_2n\ [0,1], k\_o2\_1p\ [0,1], k\_o2\_2p\ [0,1]$,
$k\_o2\_1n\ [0,1], k\_o2\_2n\ [0,1]$,;
equations $C7_{+flux} = C7\hat{}k\_o1\_1p * C8\hat{}k\_o1\_2p$;
$C7_{-flux} = C7\hat{}k\_o1\_1n * C8\hat{}k\_o1\_2n$;
$C8_{+flux} = C7\hat{}k\_o2\_1p * C8\hat{}k\_o2\_2p$;
$C8_{-flux} = C7\hat{}k\_o2\_1n * C8\hat{}k\_o2\_2n$;

Table 5: Model for biochemical kinetics of glycolysis induced by IPM.

model glycolysis_kinetics;
process flux_combination_G3P;
equations $d[G3P, t, 1] = 2.0828 * G3P_{+flux} + 0.0002 * G3P_{-flux}$;
process flux_combination_3PG;
equations $d[3PG, t, 1] = 1.2251 * 3PG_{+flux} + 4.3892 * 3PG_{-flux}$;
process flux_combination_F16BP;
equations $d[F16BP, t, 1] = 3.2353 * F16BP_{+flux} + 1.2893 * F16BP_{-flux}$;
process flux_combination_F6P;
equations $d[F6P, t, 1] = 9.8457 * F6P_{+flux} + 7.9592 * F6P_{-flux}$;
process flux_combination_DHAP;
equations $d[DHAP, t, 1] = 1.5514 * DHAP_{+flux} + 0.2402 * DHAP_{-flux}$;
process flux_combination_G6P;
equations $d[G6P, t, 1] = 0.1119 * G6P_{+flux} + 0.1557 * G6P_{-flux}$;
process reversible_G3P_F16BP;
equations $G3P_{+flux} = G3P\hat{}0.0824 * F16BP\hat{}0.1451$;
$G3P_{-flux} = G3P\hat{}0.7173 * F16BP\hat{}1$;
$F16BP_{+flux} = G3P\hat{}0.1678 * F16BP\hat{}0.4607$;
$F16BP_{-flux} = G3P\hat{}0 * F16BP\hat{}0.0010$;
process reversible_3PG_G3P;
equations $3PG_{+flux} = 3PG\hat{}0.2755 * G3P\hat{}0.2959$;
$3PG_{-flux} = 3PG\hat{}0.3810 * G3P\hat{}0.6193$;
$G3P_{+flux} = 3PG\hat{}0.2166 * G3P\hat{}0.2742$;
$G3P_{-flux} = 3PG\hat{}0.5907 * G3P\hat{}0.3825$;
process reversible_F16BP_G6P;
equations $F16BP_{+flux} = F16BP\hat{}0.3207 * G6P\hat{}0.0301$;
$F16BP_{-flux} = F16BP\hat{}0.2109 * G6P\hat{}0.0907$;
$G6P_{+flux} = F16BP\hat{}0.6492 * G6P\hat{}0.0855$;
$G6P_{-flux} = F16BP\hat{}0.1937 * G6P\hat{}0.4560$;
process irreversible_DHAP_3PG;
equations $DHAP_{-flux} = 3PG\hat{}0$;
$3PG_{+flux} = 3PG\hat{}0.9455$;
process reversible_G6P_F16BP;
equations $G6P_{+flux} = G6P\hat{}0 * F16BP\hat{}0.1911$;
$G6P_{-flux} = G6P\hat{}0.6751 * F16BP\hat{}0$;
$F16BP_{+flux} = G6P\hat{}0 * F16BP\hat{}0.4132$;
$F16BP_{-flux} = G6P\hat{}0.0072 * F16BP\hat{}0.6080$;
process irreversible_DHAP_G6P;
equations $DHAP_{-flux} = G6P\hat{}0$;
$G6P_{+flux} = G6P\hat{}0.6812$;
process reversible_F16BP_F6P;
equations $F16BP_{+flux} = F16BP\hat{}0.0047 * F6P\hat{}0.0584$;
$F16BP_{-flux} = F16BP\hat{}0.5435 * F6P\hat{}0.2384$;
$F6P_{+flux} = F16BP\hat{}0.6910 * F6P\hat{}0.0453$;
$F6P_{-flux} = F16BP\hat{}0.0210 * F6P\hat{}0.6336$;
process irreversible_F6P_G6P;
equations $F6P_{-flux} = G6P\hat{}0.1442$;
$G6P_{+flux} = G6P\hat{}0.6690$;

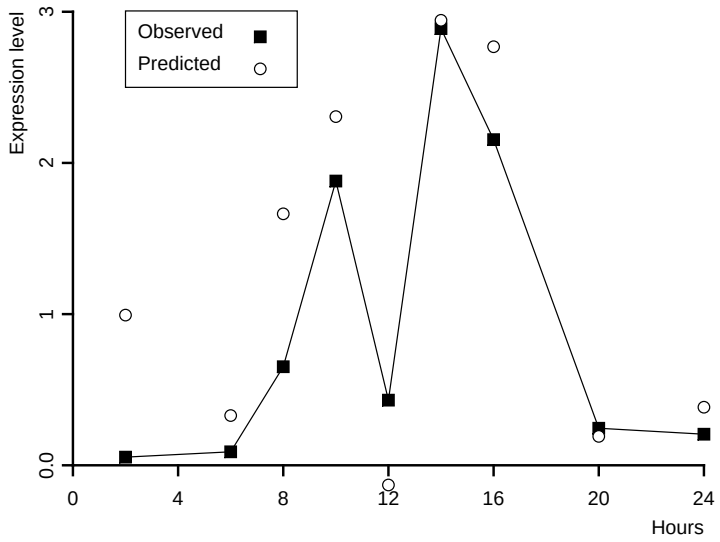

Figure 1: Average expression for 17 genes related to photosynthesis over a 24-hour period and the simulated trajectory produced by the best-scoring induced model. The dependent variable is the ratio of mRNA in each sample to the mRNA in a mixture of all the samples.

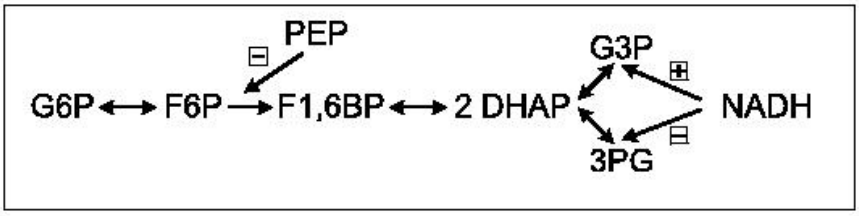

Figure 2: Model from Torralba et al. (2003) that specifies biochemical kinetic reactions among metabolites.

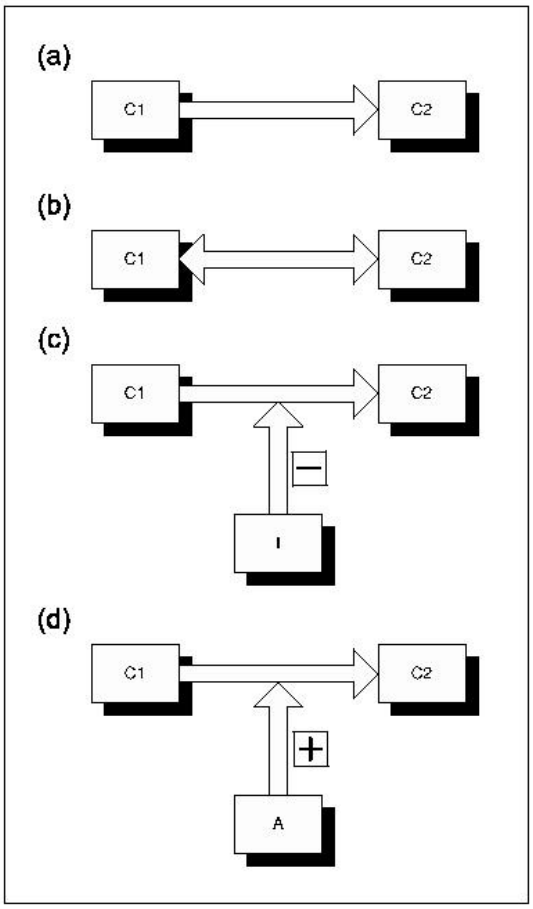

Figure 3: Different connection types for biochemical interactions, including (a) an *irreversible* reaction, (b) a *reversible* reaction, (c) an *inhibition* influence, and (d) an *activation* influence.

Figure 4: Graphical representation of the best-scoring model induced by IPM for biochemical kinetic reactions among metabolites.

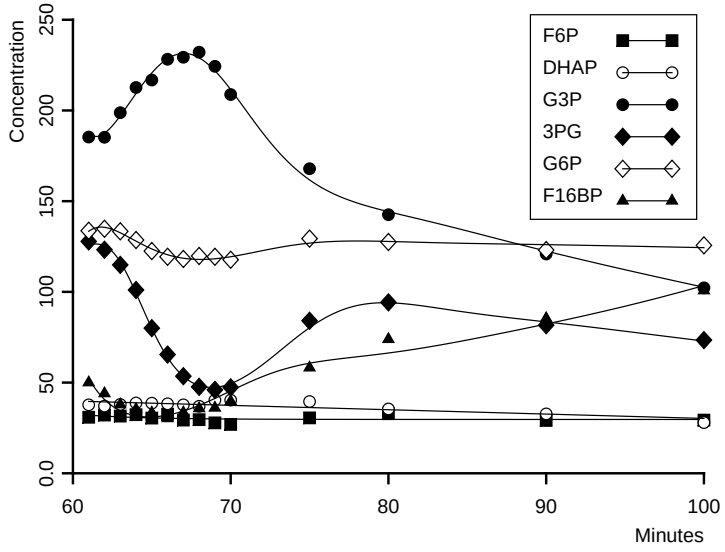

Figure 5: Concentrations for metabolites predicted by the induced model in Table 5 and corresponding concentrations measured by Torralba et al. (2003).