In L. Magnani, N.J. Nersessian, and C. Pizzi (Eds.),
Logical and Computational Aspects of Model-Based Reasoning

Computational Discovery of Communicable Scientific Knowledge

Pat Langley and Jeff Shrager
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306, USA
{langley,shrager}@isle.org

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

Abstract In this paper we distinguish between two computational paradigms for knowledge discovery that share the notion of heuristic search, but differ in the importance they place on using scientific formalisms to state discovered knowledge. We also report progress on computational methods for discovering such communicable knowledge in two domains, one involving the regulation of photosynthesis in phytoplankton and the other involving carbon production by vegetation in the Earth ecosystem. In each case, we describe a representation for models, methods for using data to revise existing models, and some initial results. In closing, we discuss related work on the computational discovery of communicable scientific knowledge and outline directions for future research.

1. Introduction

Scientific discovery is generally viewed as one of the most complex human creative activities. As such, it seems worth understanding for both theoretical and practical reasons. One powerful metaphor treats the discovery process as a form of computation, and in fact work that adopts this metaphor has a long history that dates back over two decades (e.g., Langley, 1979; Lenat, 1977; Lindsay et al., 1980). Research within this framework has advanced steadily until, in recent years, it has led

to new discoveries deemed worth publication in the scientific literature (e.g., see Langley, 2000). However, despite this progress, work on the topic remains subject to important limitations.

In this paper, we describe a new computational approach to discovery of scientific knowledge and illustrate its application to two domains. The first focuses on constructing regulatory models for photosynthesis in phytoplankton using data from DNA microarrays. The second involves finding a quantitative model of the Earth ecosystem that fits environmental data obtained from satellites and ground stations. In both cases, we report our formalism for representing models, a computational technique for producing them from observations, and initial results with actual data.

Although these two applications differ on many dimensions, they also share a reliance on three concerns: the discovered knowledge must be communicable to domain scientists; the new model must be linked to previous domain knowledge; and the model must move beyond a descriptive summary to explain the observations. We should also note that our long-term goal is not to automate the discovery process, but instead to provide interactive tools that scientists can direct and use to aid their model development.

After describing our approaches to discovery in microbiology and Earth science, we discuss related work on computational discovery and outline some likely directions for future research. However, before presenting our computational framework and its application, we must first place it in a broader historical context of work on knowledge discovery.

2. Paradigms for computational discovery

As Kuhn [1962] has noted, the *paradigm* within which scientific research occurs has a major impact on both its content and its method, and computational research on knowledge discovery is no exception. For this reason, we should review the two major frameworks for studying the discovery process in computational terms. These two paradigms hold some important assumptions in common, but they diverge on a key issue.

2.1 The data mining paradigm

A number of developments have made possible the progress on computational approaches to knowledge discovery. The most recent breakthrough, which we may call the *data* revolution, came from the insight that one can benefit by collecting and storing, automatically, vast amounts of data that describe natural, engineering, and social domains of interest. These abilities have been made practical by the availability of

inexpensive computer memory storage, the advent of new measurement techniques that ease data acquisition, and the introduction of communication infrastructure (e.g., the Internet) that supports rapid transfer of data. We can set the date for this revolution around 1995, when these technologies became common, but awareness of the coming situation was widespread five years earlier. Naturally, the access to electronic data sets holds great potential to support knowledge discovery, and many scientists, engineers, and businessmen have focused their energies on fulfilling that potential.

A somewhat earlier development, which we may call the *search* revolution, resulted from the insight that computers are general symbol manipulators and that one can view many tasks which require intelligence as involving search through a space of symbolic structures. This ability became practical with the introduction of computer programming languages that could represent and manipulate symbolic structures, as well as algorithms for carrying out heuristically-guided search through a space of such structures. We can date this revolution to the middle 1950s, when Newell and Simon [1956] created the first list-processing language and used it to automate search for proofs of logical theorems. Notions of heuristic search preceded this achievement, but computationalists began to apply the idea in earnest only after this proof of concept. Simon [1966] was also one of the first authors to view the discovery process in terms of search.

In recent years, these two insights have been combined by researchers and developers in a paradigm known as *data mining* or *knowledge discovery in databases*. Work in this arena emphasizes the availability and potential of large, electronic data sets, as well as computational techniques that can represent and search for knowledge implicit in those data. The data mining community has inherited its key techniques from two parent disciplines – machine learning and databases – that have focused historically on computational processing of data. This approach has become especially popular in the commercial sector, where it has been applied successfully to manufacturing, marketing, and finance, but it has also been put to good effect in a variety of scientific fields.

However, despite its impressive track record, the data mining framework has an important drawback related to its emphasis on the discovery of knowledge in understandable forms. In principle, this concern is perfectly legitimate, since we typically assume that knowledge can be represented explicitly and communicated among humans. Yet the data mining community's efforts along these lines have focused on particular formalisms it has inherited from its parent disciplines, notably decision trees, logical rules, and Bayesian networks. Researchers regularly take

positions about the understandability of such representations, but their stances are based more on popular myths than on careful reasoning or empirical evidence.

One such myth concerns the claim that univariate decision trees, with their logical semantics, are inherently easier to understand than alternative notations, like probabilistic classifiers, that involve numeric weights and degrees of match. Yet Igor Kononenko [personal communication, 1993], who originally believed this intuition, found that medical doctors felt a naive Bayesian classifier, which computes probabilistic summaries, was easier to comprehend than decision trees induced from the same patient data. Presumably, this was because the physicians had more exposure to probability theory than to nonparametric schemes like decision trees. We can draw a tentative conclusion from this result: knowledge is more understandable when cast in a formalism familiar to the recipient.

A similar myth involves the claim that computational methods like backpropagation, which learns weights in a multilayer neural network, produce results that are inherently opaque. Yet Saito and Nakano [1997] have shown that, by carefully structuring the network architecture, one can use backpropagation to discover numeric equations like those central to physics and other sciences, and which, presumably, are interpretable by experts in those domains. We can draw another plausible lesson from this result: whether the discovered knowledge is understandable depends far less on the search algorithm than on the manner in which one uses that algorithm.

2.2 Computational scientific discovery

These observations suggest the relevance of a third, much older, historical development, the *scientific* revolution, which introduced not only the idea of evaluating laws and theories in terms of their ability to fit observations, but also emphasized the casting of such knowledge in some formal notation. We can date this insight to around 1700, when Newton's theory of gravitation became widely accepted, though it was predated by similar formal statements like Kepler's laws. Over the past 300 years, scientists and engineers have developed a variety of formalisms to represent knowledge that bear little resemblance to the notations which dominate the data mining community. We hold that such formalisms from science and engineering are more appropriate targets for knowledge discovery, at least in such domains, than data mining notations.

In fact, there exists an alternative computational paradigm, predating the data mining framework, that combines the representational insights of the scientific revolution with the notion of heuristic search. We will

refer to this framework as *computational scientific discovery*, since its primary focus has been finding laws and theories in scientific domains. This paradigm also assumes the presence of data or observations, but emphasizes their role less than the search metaphor and scientific notations. Research in this area addressed originally the rediscovery of knowledge from the history of science (e.g., Langley et al., 1987; Shrager and Langley, 1990), but the last decade has seen numerous examples of novel discoveries that have led to publications in the relevant scientific literature [Langley, 2000]. We maintain that this approach is more appropriate for the discovery of *communicable* knowledge than the data mining framework precisely because it utilizes formalisms already familiar to domain experts.

Note that there has been considerable work within the KDD tradition on scientific domains. Much of this has focused on applications to molecular biology, such as learning predictors for protein folding, but Fayyad et al. [1996] review similar efforts in astronomy, such as distinguishing stars from galaxies, and planetology, such as detecting volcanoes on Venus. This work has proven valuable to the disciplines involved, but we hold that the knowledge discovered in these cases is not communicable in the same sense as described above. The learned predictors, whether stated as decision trees, neural networks, or probabilistic classifiers, are unlikely to appear as knowledge themselves in scientific papers, and thus would not be *communicated*. Rather, they play the role of measuring instruments, which are essential to scientific progress but which constitute *tacit* knowledge [Polanyi, 1958] rather than the communicable variety.

By this point, we hope to have convinced readers that the task of communicable knowledge discovery differs in important ways from the problems typically pursued in the data mining community, and that this task deserves significantly increased attention among knowledge discovery researchers. For despite the success stories to date, there remain many open problems that require additional effort. For instance, most research on computational scientific discovery has focused on finding knowledge from scratch, but scientists are typically concerned with revising and improving existing theories. Researchers in the field have also concentrated primarily on discovery of descriptive regularities, but scientists often aim for models that explain observed phenomena in terms of unobserved variables and processes. Finally, most work on computational discovery has emphasized automating this activity, but scientists would benefit more from interactive tools that assist them in their efforts rather than ones that aim to replace them.

In the sections that follow, we report progress on these issues in the context of two scientific domains. In both cases, we review an existing

explanatory model that accounts partially for some phenomena, describe a computational system that revises this model to fit these data better, and present some initial results of such improvement. Our research on interactive tools has advanced less, but we have designed our revision techniques to support such a capability. As in other work on computational scientific discovery, the systems cast their discovered knowledge in a familiar scientific notation to ensure communicability.

3. Revising regulatory models in microbiology

Although biologists understand the basic mechanisms through which DNA produces biochemical behavior, they have not yet determined most of the regulatory networks that control the degree to which each gene is expressed. However, for particular organisms under certain conditions, they have developed partial models of gene regulation. The measurement and analysis of gene expression levels, either through Northern blots or cDNA microarrays, has played a central role in the elucidation of regulatory models, as both measures quantify gene activity in terms of RNA concentration.

The most popular computational approach to processing such expression data – clustering genes into coregulated classes – is a clear example of the data mining paradigm. This knowledge-lean method lets one reduce the high dimensionality of microarray data to a manageable level, but the results take the form of descriptions rather than explanations. A second paradigm, more commonly used by practicing biologists, uses data about expression levels to test specific pathway hypotheses. This knowledge-rich approach lets one evaluate proposed explanations, but it generally does not move beyond these hypotheses to suggest improved regulatory models.

In this section, we describe an approach that combines knowledge with data to revise an initial biological model. We focus on the regulation of photosynthesis in Cyanobacteria, an area for which we have both a model proposed by domain scientists and microarray data collected to evaluate this model. As outlined above, our goal is to develop computational methods that can utilize data to improve such a model while retaining its communicability and its links to existing biological knowledge.

3.1 Representing models of gene regulation

Any computational method designed to improve regulatory models must first have some representation for those models. As we have noted, most work in machine learning and data mining draws on representational formalisms that were designed by artificial intelligence researchers

and that make little contact with notations commonly used by practicing scientists. In contrast, we are committed to representing biological models in terms that are familiar to biologists themselves.

Figure 1 presents a regulatory model, obtained from a plankton biologist, that aims to explain why Cyanobacteria bleaches when exposed to high light conditions. Each node corresponds to some variable, either observable or theoretical, whereas each link depicts some biological process through which one variable influences another. Solid lines denote internal processes, whereas dashes indicate processes connected to the environment.

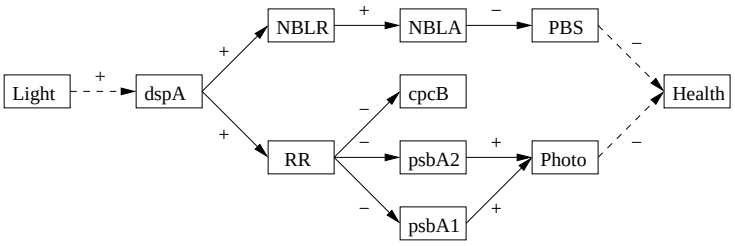

Figure 1. An initial model for regulation of photosynthesis in Cyanobacteria.

The model states that changes in light level modulate the activity of dspA, a protein hypothesized to serve as a sensor. This in turn regulates NBLR and NBLA, which then reduce the number of phycobilisome (PBS) rods that absorb light, which is measurable photometrically as the organism's greenness. The reduction in PBS protects the organism's health because it decreases the absorbtion of light, which can be damaging at high levels. The organism's health under high light conditions can be measured in terms of culture density. The sensor dspA also impacts health through a second pathway by influencing a hypothesized response regulator, RR, which in turn down regulates expression of the gene products psbA1, psbA2, and cpcB. The first two influence positively the level of photosynthetic activity (Photo) by altering the photosystem's structure. If left unaltered, this second pathway would also damage the organism under high light conditions.

Although this model incorporates quantitative variables, it specifies only the directions of influence and not their specific form or their parameters. AI research in qualitative physics (e.g., Forbus, 1984) has used similar notations to support common sense reasoning. We have focused on such qualitative models not because quantitative ones are undesirable, but because biologists usually operate on the former, and we want our computational tools to support their typical reasoning styles.

The example model is also partial and abstract, in that the biologist who proposed it clearly viewed it as a working hypothesis. Some processes are abstract in that they denote entire chains of subprocesses. For instance, the link from dspA to NBLR denotes a complex signaling pathway for which the details are unknown or irrelevant at this level of analysis. The model also includes abstract variables like RR, which refers to an unspecified gene (or set of genes) that acts as an intermediary controller. Thus, our formalism can express partial, abstract, and qualitative models like those often used by biologists.

For the sake of analytical tractability, we also assume that each variable is a linear function of its direct causes plus an error term. This means that we can represent the entire model as a system of linear equations, which Glymour et al. [1987] refer to as a *linear causal model*. This approach to modeling has been used widely in econometrics, where the data are purely observational. Most research in this framework deals with quantitative models that specify the parameters for each equation, but, again, we focus here on the qualitative version.

3.2 Utilizing, evaluating, and revising models

Since our models are qualitative, they cannot predict directly the continuous expression levels one can observe for genes, but they do imply certain relations among variables. In particular, they predict which variables should be correlated and the direction of those relationships. If two variables are connected directly, then we expect their correlation to have the same sign as that on their link. If they are connected indirectly, we multiply the signs on the path that connects them. For instance, the model in Figure 1 predicts that NBLA and cpcB will be negatively correlated, even though neither has a direct causal influence on the other and the path connecting them passes through RR, an unobservable variable.

In some cases, there exist multiple paths between a pair of variables. When the predicted sign for all paths between these nodes agree, the system simply makes that prediction. However, when two or more paths disagree, we assume the model includes an annotation that indicates either the positive or negative paths are dominant, which gives an unambiguous prediction. This extended formalism lets a qualitative model predict a positive or negative correlation for each pair of observed variables, even without information about the quantity of each link's effect.

In addition, casting our regulatory structures as linear causal models lets us make other important predictions about *partial correlations*, which describe the relationship between two variables once the effects of other terms have been factored out. For instance, the partial correlation

$\rho_{12.3}$ denotes the correlation between X_1 and X_2 when controlling for X_3. Simon [1954] has shown that a zero partial correlation $\rho_{12.3}$ implies that X_1 and X_2 are connected through X_3. In contrast, a nonzero partial correlation implies that X_1 and X_2 are connected through paths that do not involve X_3. Thus, the model in Figure 1 predicts that the partial correlation of dspA and PBS given NBLA will be zero, because the variable NBLA lies along the path between them. Glymour et al. have generalized these conditions for more complicated models, but the intution remains the same.

Our approach evaluates a candidate regulatory model by predicting, for each set of three variables, which partial correlations should occur and which ones should not. The system then calculates these partial correlations from the data and determines, for each one, whether it differs significantly from zero. Upon comparing the predicted partial correlations with those supported by the data, it obtains the number of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn). The system combines these counts using

$$score = fp + fn - tp - tn \ ,$$

which provides an overall measure of the model's qualitative fit to the observations. Because most linear causal models imply different partial correlations, this metric lets it discriminate among many alternative regulatory structures.

To revise its model of gene regulation, the system carries out a two-stage heuristic search through a space of candidate models. The first stage, which focuses on the causal structure, starts from the initial model proposed by biologists with the signs on links removed. The operators for generating alternative models include adding a link between variables, removing an existing link, and reversing the direction of a link.[1] The system invokes the *score* metric described above to select among models, and it carries out hill-climbing search through the model space, on each step selecting the revision that most improves the evaluation metric. The search halts after a prespecified number of revision steps.

Because experiments that measure gene expression typically collect few samples, this approach is unstable in that small changes to the data can produce very different models. To offset this, the system generates 20 different training sets by sampling with replacement from the orig-

[1]These operators are constrained by biological knowledge. For instance, the system knows that stimulus variables like Light must serve as causal influences to gene variables, and that behavioral variables like Photo must be caused by the latter.

inal data, then runs its revision algorithm to generate 20 new models. The program then counts how many times each revision occurs in these models and retains only those that appear in at least 75 percent of them.

Once the system has induced the model's causal structure, the second stage carries out another search to determine the signs on links. In this case, the evaluation function measures instead the number of correlations for which the predicted and observed signs agree. If the model involves only a few links, the system considers exhaustively all possible assignments of pluses and minuses on the links, then selects the best-scoring assignment. Otherwise, it resorts to hill climbing through the space of assignments, starting from those in the initial model and halting when no further improvement occurs.

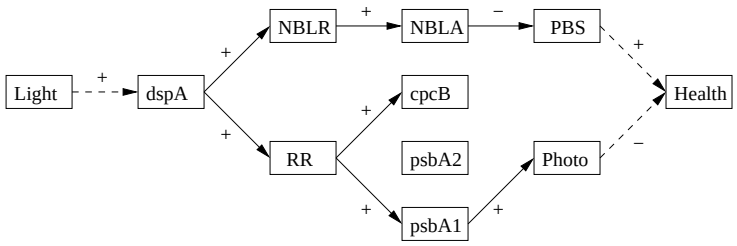

Figure 2. A revised model for regulation of photosynthesis in wild Cyanobacteria.

3.3 Initial results on photosynthetic regulation

We applied our revision method to data for the wild type Cyanobacteria and a mutant that does not bleach under high light conditions. We have data from cDNA microarrays about the expression levels for approximately 300 genes believed to play a role in photosynthesis. For the initial analysis, we focused on genes in the initial model shown in Figure 1 and did not consider links to other genes. The microarray data, which reflects the concentration of mRNA for each gene relative to that in a control condition, were measured at 0, 30, 60, 120, and 360 minutes after high light was introduced, with four replicated measurements at each time point. We treated the data as independent samples, ignoring their temporal aspects and dependencies among the replicates.

Figure 2 shows the revised model that the system produced from these data. There are five differences from the initial regulatory account. Two changes, removal of the links to and from psbA2, involve the model structure. The other three revisions concern changes of signs, in particular for the links from RR to psbA1, from RR to cpcB, and from PBS to Health.

Discussions with the biologist who proposed the original model indicate a strong belief that RR influences Photo, but uncertainty about the exact pathways. This means that the changes which involve RR are not problematic, since the presence of one gene product (psbA1) is enough to regulate the photosynthetic center (Photo). However, the reversed sign on the link from PBS to Health raises a problem, since the belief that excessive light causes damage means this link should be positive. We hypothesize that, in this study, the light exposure was not enough to overcome benefits from the energy it provides, which the model omits.

We also tested the system on expression data for a mutant of Cyanobacteria that does not bleach under high light conditions. Presumably, such a mutant differs genetically from the wild organism in only a few ways, so we started search from the same model as in our first study. In this case, the system removed the link from dspA to RR, but made no other revisions. This is a plausible change, since the mutation involved removal of the dspA gene from the organism. However, the new model does not explain why the mutant fails to bleach when exposed to high light. One possibility is that the 20 samples did not provide enough statistical power to let the system remove the link from dspA to NBLR, which would produce the desired effect. Although these initial results are encouraging, it seems clear that we can still improve our approach to revising qualitative models of gene regulation. Elsewhere [Shrager et al., 2002] we discuss some directions for future research along these lines.

4. Revising quantitative models in Earth science

Earth scientists have reached a broad enough understanding of ecosystem processes to develop models for the entire biosphere. These differ from the microbiological models we considered in the last section in that they are primarily quantitative rather than qualitative. Ecosystem models can also be quite complex, containing tens of equations, many theoretical variables, and parameters for each grid cell, which can number in the thousands. Such models are consistent with high-level ecosystem phenomena, but the availability of new data from satellites and other sources provides the opportunity to refine them further.

One such model, Potter and Klooster's [1997, 1998] CASA, predicts, with reasonable accuracy, the global production and absorption of biogenic trace gases in the Earth's atmosphere, as well as explaining changes in the geographic vegetation patterns on the land. The model's predictive variables include surface temperature, moisture levels, and soil properties, along with global satellite observations of the land surface. CASA incorporates both instantaneous and difference equations that

describe changes over time due to the terrestrial carbon cycle and processes that mineralize nitrogen and control vegetation type. The model operates on gridded input, with typical usage involving grid cells that are eight kilometers square, since this matches the resolution for land surface observations obtained from satellites.

Although CASA has been quite successful at modeling Earth's ecosystem, its predictions still differ from observations in certain ways, and in this section we describe a computational approach to improving its fit to the data available. As before, the result is a revised model, cast in the same notation as the original one, that incorporates changes that are scientifically plausible and, we hope, interesting to Earth scientists.

4.1 A portion of the CASA model

Rather than attempting to refine the complete CASA model, which is quite complex, we decided to focus on a submodel near the 'top' that leads directly to the main dependent variable, NPPc, which denotes the net production of carbon. Table 1 lists the variables that occur in this submodel and summarizes the quantities they represent, whereas Table 2 shows the equations that relate these variables, with indentation reflecting the submodel's logical structure.

The first equation in Table 2 states that NPPc is the product of two unobservable variables, the photosynthetic efficiency at a site, E, and the solar energy intercepted at that site, IPAR. Photosynthetic efficiency is in turn calculated as the product of the maximum efficiency (0.56) and three stress factors that reduce this efficiency. The first stress term, T2, takes into account the difference between the optimum temperature, Topt, and actual temperature, Tempc, for a site. The second factor, T1, involves the nearness of Topt to a global optimum for all sites, reflecting the intuition that plants which are better adapted to harsh temperatures are less efficient overall.

The third term that influences photosynthetic efficiency, W, represents stress that results from lack of moisture as determined by EET, the estimated water loss due to evaporation and transpiration, and by PET, the water loss due to these processes given an unlimited water supply. In turn, PET is influenced by AHI, the annual heat index for a site, and PET-TW-M, another component of potential evapotranspiration.

The model predicts IPAR, the energy intercepted from the sun, as the product of FPAR-FAS, the fraction of energy absorbed through photosynthesis, MONTHLY-SOLAR, the average radiation that occurs during a given month, and SOL-CONVER, the number of days in that month. FPAR-FAS is in turn a function of MON-FAS-NDVI, which indicates

NPPc is the net plant production of carbon at a site during the year.
E is the photosynthetic efficiency at a site after factoring various sources of stress.
T2 is a temperature stress factor ($0 < T2 < 1$), nearly Gaussian in form but falling off more quickly at higher temperatures.
T1 is a temperature stress factor ($0 < T1 < 1$) for cold weather.
W is a water stress factor ($0.5 < W < 1$) for dry regions.
Topt is the average temperature for the month at which MON-FAS-NDVI takes on its maximum value at a site.
Tempc is the average temperature at a site for a given month.
EET is the estimated evapotranspiration (water loss due to evaporation and transpiration) at a site.
PET is the potential evapotranspiration (water loss due to evaporation and transpiration given an unlimited water supply) at a site.
PET-TW-M is a component of potential evapotranspiration that takes into account the latitude, time of year, and days in the month.
A is a polynomial function of the annual heat index at a site.
AHI is the annual heat index for a given site.
MON-FAS-NDVI is the relative vegetation greenness for a given month as measured from space.
IPAR is the energy from the sun that is intercepted by vegetation after factoring in time of year and days in the month.
FPAR-FAS is the fraction of energy intercepted from the sun that is absorbed photosynthetically after factoring in vegetation type.
MONTHLY-SOLAR is the average solar irradiance for a given month at a site.
SOL-CONVER is 0.0864 times the number of days in each month.
UMD-VEG is the type of ground cover (vegetation) at a site.

Table 1. Variables used in the NPPc portion of the CASA ecosystem model.

relative greenness at a site as observed from space, and SRDIFF, an intrinsic property that takes on different numeric values for different vegetation types as specified by the discrete variable UMD-VEG.

Making predictions from this submodel is a straightforward process, in that one simply starts from the observable[2] input variables – Tempc, MONTHLY-SOLAR, SOL-CONVER, MON-FAS-NDVI, UMD-VEG,

[2]Actually, the variables EET, PET-TW-M, and AHI are unobservable terms defined elsewhere in the model. To make the revision task more tractable, we assumed their definitions were correct and treated them as observables, using the model to compute their values.

$$\text{NPPc} = \sum_{month} \max(\text{E} \cdot \text{IPAR}, 0)$$
$$\text{E} = 0.56 \cdot \text{T1} \cdot \text{T2} \cdot \text{W}$$
$$\text{T1} = 0.8 + 0.02 \cdot \text{Topt} - 0.0005 \cdot \text{Topt}^2$$
$$\text{T2} = 1.18/[(1 + e^{0.2 \cdot (\text{Topt} - \text{Tempc} - 10)}) \cdot (1 + e^{0.3 \cdot (\text{Tempc} - \text{Topt} - 10)})]$$
$$\text{W} = 0.5 + 0.5 \cdot \text{EET/PET}$$
$$\text{PET} = 1.6 \cdot (10 \cdot \text{Tempc/AHI})^A \cdot \text{PET-TW-M if Tempc} > 0$$
$$\text{PET} = 0 \text{ if Tempc} \leq 0$$
$$\text{A} = 0.000000675 \cdot \text{AHI}^3 - 0.0000771 \cdot \text{AHI}^2 + 0.01792 \cdot \text{AHI} + 0.49239$$
$$\text{IPAR} = 0.5 \cdot \text{FPAR-FAS} \cdot \text{MONTHLY-SOLAR} \cdot \text{SOL-CONVER}$$
$$\text{FPAR-FAS} = \min((\text{SR-FAS} - 1.08)/\text{SRDIFF(UMD-VEG)}, 0.95)$$
$$\text{SR-FAS} = -(\text{MON-FAS-NDVI} + 1000)/(\text{MON-FAS-NDVI} - 1000)$$

Table 2. Equations used in the NPPc portion of the CASA ecosystem model.

EET, PET-TW-M, and AHI – and calculates values for the variables that depend on them. The resulting quantities are then passed to other equations that compute values for other terms, with this continuing until a value for NPPc is predicted. One repeats this process with each grid cell for which observations are available.

4.2 An approach to quantitative model revision

As before, our approach to scientific discovery involves refining a model like that in Table 2 rather than constructing one from scratch. Thus, this initial model constitutes the starting point for heuristic search through a space of models, with the search process directed by candidates' ability to fit the data. However, in this case our models are quantitative rather than qualitative and, as such, require different operators and a different evaluation function to direct search.

To this end, we assume that the overall structure of the model is correct, but that the specific equations and their parameters can be improved. For example, after the revision process, NPPc would still be defined in terms of E and IPAR, but the functional form of this definition may no longer be $\text{NPPc} = \text{E} \cdot \text{IPAR}$. Moreover, we can utilize parameter revision to mimic revision of equation forms by encoding each expression in the initial model as a multivariate polynomial equation of the form

$$y = w_0 + \sum_{j=1}^{J} w_j \prod_{k=1}^{K} X_k^{w_{jk}},$$

where y is a continuous variable that depends on continuous variables $X_1, \ldots, X_K$. For example, the equation $\mathrm{W} = 0.5 + 0.5 \cdot \mathrm{EET/PET}$ in this scheme becomes $\mathrm{W} = 0.5 + 0.5 \cdot \mathrm{EET}^{1.0} \cdot \mathrm{PET}^{-1.0}$. Such functional relations subsume many of the numeric laws found by earlier quantitative discovery systems like BACON [Langley, 1979] and FAHRENHEIT [Żytkow et al., 1990], as well as the expressions in Table 2.

This encoding transforms our set of equations into the equivalent of a multilayer neural network, with one subnetwork for each relationship in the model. More specifically, each equation becomes a two-layer network with product units in the first level, to encode multiplicative terms, and additive units in the second level, to encode their weighted summation. This transformation maps the set of possible models into a weight space. By adapting Saito and Nakano's [1997] BPQ algorithm for discovering numeric equations, we can implement a gradient descent search through this space. Briefly, this method incorporates a second-order learning technique that calculates both the descent direction and the step size automatically. The search process halts when it finds a set of weights that minimize the squared error on the dependent variable y. The method then transforms the resulting network back into a set of polynomial equations, with weights on product units becoming exponents and weights on linear units becoming coefficients.

We can see readily how this approach can improve the parameters for an equation. Although the NPPc submodel contains some parameterized equations that our Earth science collaborators believe are reliable, like that for computing the variable A from the annual heat index AHI, it also includes equations with parameters about which there is less certainty, like the expression that predicts the temperature stress factor T2 from Tempc and Topt. By fixing the weights that correspond to reliable parameters, as well as the weights that encode exponents, the BPQ algorithm searches through the weight space associated with the other parameters to find settings that reduce predictive error. We can use the same mechanism to revise the form of an equation by specifying that the weights for exponents should not be fixed.

We must extend the approach slightly to support revision of values for an intrinsic property (e.g., SRDIFF) that the model associates with the discrete values for some nominal variable (e.g., the vegetation type UMD-VEG). In such cases, we encode each nominal term as a set of dummy variables, one for each discrete value, setting the dummy variable equal to one if the discrete value occurs and zero otherwise. We introduce one hidden unit for the intrinsic property, with links from each dummy variable and weights that correspond to the intrinsic value associated with each discrete value. We then utilize Saito and Nakano's BPQ

algorithm to search the weight space that corresponds to alternative sets of intrinsic values, using the original model to initialize weights.

Although this approach to model refinement can modify more than one equation or intrinsic property at a time, the results we report in the next section assume that the user focuses the system's attention on one portion of the model. We envision an interactive mode in which the scientist identifies a portion of the model that he thinks could be better, runs the revision method to improve its fit to the data, and repeats this process until he is satisfied.

4.3 Initial results on ecosystem model revision

In order to evaluate our approach to quantitative model revision, we utilized data relevant to the NPPc submodel that were available to the Earth science members of our team. These data consisted of observations from 303 distinct sites with known vegetation type and for which measurements of Tempc, MON-FAS-NDVI, MONTHLY-SOLAR, SOL-CONVER, and UMD-VEG had been recorded for each month of the year. In addition, other portions of CASA were able to compute values for the variables AHI, EET, and PET-TW-M. The resulting 303 training cases seemed sufficient for initial tests of our revision methods, so we used them to drive a variety of changes to the handcrafted model of carbon production.

Discussions with our Earth science collaborators suggested the expression for T2, one of the temperature stress variables, as a likely candidate for revision. As we saw in Table 2, the initial equation for this term was

$$T2 \quad = \quad 1.8/[(1 + e^{0.2(Topt - Tempc - 10)})(1 + e^{-0.3(Tempc - Topt - 10)})] \ ,$$

which generates a Gaussian-like curve, shown in Figure 3, that is slightly assymetrical. This reflects the intuition that the photosynthetic efficiency of vegetation will decrease when the actual temperature (Tempc) is either below or above the optimal temperature (Topt). When we asked our system to improve the parameters in this expression but to retain its original form, it produced

$$T2 \quad = \quad 1.80/[(1 + e^{0.05(Topt - Tempc - 10.8)})(1 + e^{-0.03(Tempc - Topt - 90.33)})] \ ,$$

which has fairly similar values to the initial ones for some parameters but quite different values for others. The root mean squared error for the revised model was 461.466, as measured by leave-one-out cross validation, which was only one percent better than the 467.910 error for the original model.

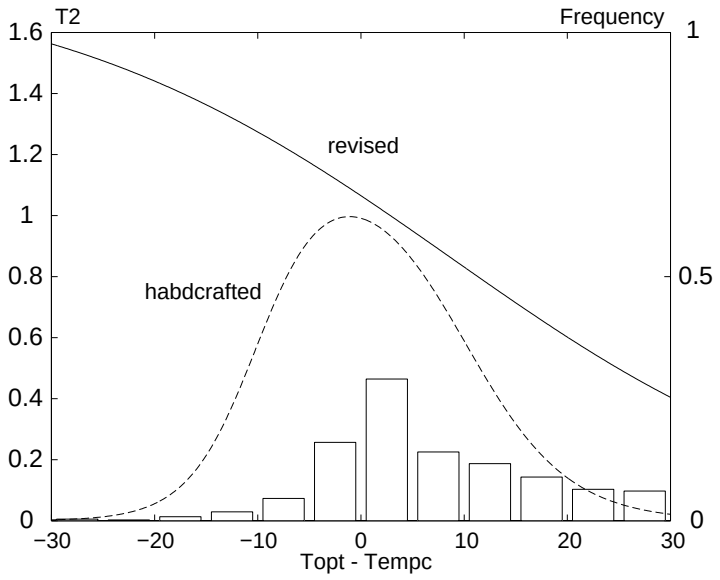

Figure 3. Behavior of handcrafted and revised equations for the stress variable T2.

Although this result seems disappointing at first glance, the curves in Figure 3 reveal a more interesting picture. Plotting the temperature stress factor T2 using the revised equations as a function of the difference Topt − Tempc still gives a Gaussian-like curve, but within the effective range (from −30 to 30 Celsius) its values decrease monotonically. This seems counterintuitive but interesting from an Earth science perspective, as it suggests this stress factor has little influence on NPPc. Because the original equation for T2 was not well grounded in principles of plant physiology, such observations are beneficial to the modeling enterprise even when the empirical improvement is small.

As another candidate for parameter revision, we selected the equation for PET, which calculates potential water loss due to evaporation and transpiration given an unlimited water supply. In this case, the revised parameter values were all very similar to those in the original model's equation and led to no substantial improvement in accuracy. Since the PET equation is based on a method that has been used continuously in Earth science for over 50 years, we should not be overly surprised at this negative result. Indeed, we are encouraged by the fact that our approach did not revise parameters that have stood the test of time.

We also applied our approach to revise values for the intrinsic property SRDIFF that are associated with different vegetation types UMD-VEG. For each site, the latter variable takes on one of 11 nominal values, such

as grasslands, forest, and desert, each with an associated numeric value for SRDIFF that plays a role in the FPAR-FAS equation. As outlined earlier, to revise these intrinsic values, we introduced one dummy variable, UMD-VEG$_k$, for each vegetation type such that UMD-VEG$_k = 1$ if UMD-VEG $= k$ and 0 otherwise.

In this case, the improvement was more substantial, with the revised model reducing error by over four percent, which seems substantial. We have reported the revised intrinsic values elsewhere [Saito et al., 2001], but the most striking result was that the altered values were nearly always lower than the initial values. This result is certainly interesting from an Earth science viewpoint. Our domain experts suspect that measurements of NPPc and related variables from a wider range of sites would produce intrinsic values closer to those in the original model, but such a test must await additional observations.

Because the original 11 intrinsic values were grouped into only four distinct values, we also applied a clustering procedure over the trained neural network to group the revised values in the same manner. We examined the effect on error rate as we varied the number of clusters from one to five. As expected, the training error decreased monotonically, but the cross-validation error was minimized for three clusters. The estimated error for this revised model was better than for the one with 11 distinct values, but only slightly. Again, the clustered values were nearly always lower than the initial ones.

As we noted earlier, our system can also revise the functional forms in a quantitative model. One candidate for such revision was the equation for photosynthetic efficiency, E, which is calculated as a product of three stress terms in

$$E \;=\; 0.56 \cdot T1 \cdot T2 \cdot W \;.$$

Multiplying the stress terms has the effect of reducing photosynthetic efficiency below the maximum 0.56 possible [Potter and Klooster, 1998], since each factor takes on a value less than one.

In this case, a natural extension was to consider the space of equations that included exponents on the stress terms, which we initialized to 1.0, as in the original model, and constrained to be positive. This time, the system produced the revised equation

$$E \;=\; 0.521 \cdot T1^{0.00} \cdot T2^{0.03} \cdot W^{0.00} \;,$$

which reduced error over the original model by almost five percent. The new equation has a similar coefficient, but it also has a small exponent for T2 and zero exponents for T1 and W. These results were very interesting

to our Earth science collaborators, as they suggest that the T1 and W stress terms are not needed for predicting NPPc. One explanation is that the influence of these factors is already being captured by the NDVI measure available from space, for which the signal-to-noise ratio has been steadily improving since CASA was first developed. They are also consistent with our results with the T2 equation, which revealed monotonically changing values for this variable over the relevant range.

5. Related research on computational discovery

As we noted earlier, there is a substantial literature on the computational discovery of communicable scientific knowledge (e.g., Langley et al., 1987; Džeroski and Todorovski, 1993; Washio and Motoda, 1998), but most of this research has focused on the construction of laws and models, rather than on their revision. There also exists a nearly disjoint literature on the computational revision of knowledge bases cast in non-scientific formalisms, most often using Horn clauses and related logical notations (e.g., Ourston and Mooney, 1990). However, there has been some work on the revision of scientific theories, which we should review here briefly.

One body of related research has involved revision of structural models from the history of chemistry and physics. For example, Żytkow and Simon's [1986] STAHL detected inconsistencies in chemical reactions and revised its componential models by adding or removing constituents. Rose and Langley's [1986] STAHLp improved on this approach and applied it to additional historical episodes. Kocabas' [1991] BR-3 system extended this framework to include detection of incomplete theories and postulation of new properties to explain the absence of reactions, then applied these strategies to the history of particle physics. Finally, O'Rorke et al. [1990] developed AbE, an abductive system for model revision which they used to model the shift from the phlogiston to the oxygen theory.

Other work on the revision of qualitative scientific theories, more akin to our own, has focused on process models that explain causal events. Rajamoney's [1990] COAST system incorporated ideas from qualitative physics to represent and revise models about fluid and heat flow, whereas Karp's [1990] HYPGENE used a qualitative biochemical notation to support revision of models about attenuation in gene regulation. Kulkarni and Simon [1990] describe KEKADA, a system that reproduced many steps in Krebs' discovery of the urea cycle. All three systems augmented the revision process with methods for experiment design that aimed to distinguish among competing hypotheses.

There exists less research on the revision of quantitative scientific models. Chown and Dietterich [2000] report an approach that improves an existing ecosystem model's fit to continuous data, but their method only alters parameter values and does not revise equation structure. Džeroski and Todorovski [2001] present LAGRAMGE, a system that revises both the structure of a model's equations and their parameters, using a grammatical formalism to specify domain constraints on acceptable models. They have applied this approach to the same portion of the CASA ecosystem model as we have addressed and obtained similar improvements. Early research by Glymour et al. [1987] addressed revision of linear causal models that took a quantitative form, but their methods are more closely related to those we have used for qualitative model revision.

Our vision for an interactive discovery environment directly derives from Mitchell et al. [1997], who developed a similar environment to support discovery in metallurgy. Their DAVICCAND system let users select pairs of numeric variables to relate, specify qualitative conditions that focus attention on subsets of the data, and find numeric laws that relate variables within a given region. The program also included mechanisms for identifying outliers that violate these numeric laws and for using the laws to infer the values of intrinsic properties. DAVICCAND presented its results using graphical displays and functional forms that were familiar to metallurgists.

We should note that the notion of communicable knowledge discovery is not limited to scientific domains. Another example comes from Rogers et al. [1999], who developed methods for revising the contents of digital maps based on traces from a global positioning system. Their system improved estimates of center lines for road segments, inferred the number of lanes associated with each segment, and added content about the type of traffic signal at intersections. The revised knowledge took the same form as the initial digital map, letting it be displayed in a graphical format familiar to mapmakers and drivers while increasing the map's overall accuracy and detail.

6. Concluding remarks

In this paper, we distinguished between two broad computational approaches to discovery: the paradigm of data mining, which emphasizes the availability of large data sets to drive the search process, and computational scientific discovery, which takes advantage of established scientific formalisms to state the resulting knowledge in a communicable fashion. We argued that the latter is more appropriate for aiding discov-

ery in scientific disciplines, but we also noted the need for more research in this promising framework.

In response, we reported progress on the discovery of communicable scientific knowledge in two domains, one involving gene regulation of photosynthsis in Cyanobacteria, and the other involving carbon production by vegetation as a function of environmental factors. In both cases, we developed algorithms that discovered knowledge in the same formalisms as utilized by domain scientists. Our methods also reflected two additional concerns that have received little attention in the discovery literature: the revision of initial models, rather than their generation from scratch, and the development of explanatory models, with theoretical variables and processes, rather than purely descriptive summaries. We showed that our discovery methods, one designed for qualitative models and the other for quantitative, led to improvements over existing models in terms of their fit to available data.

Although our results to date are encouraging, we must extend our computational discovery techniques in a number of directions before they become useful tools for scientists. For example, both discovery algorithms we presented can alter an initial model's relations among variables, but they cannot introduce new variables during the revision process. Another shared limitation is the methods' support for models with instantaneous relationships among variables but not ones that involve change over time. We should augment both discovery algorithms to consider additional variables during the revision process and to support models that express temporal relations. For quantitative models like CASA, we envision using ordinary differential equations and drawing on methods like Džeroski and Todorovski's [2001] LAGRAMGE for revision; for qualitative models, we will borrow formalisms developed in the qualitative physics community (e.g., Forbus, 1984).

Clearly, such additions will increase the search space that our revision methods must explore, which in turn suggests the need for domain constraints to direct the process. To this end, we intend to introduce a taxonomy of variables and an analogous taxonomy of processes, with the latter making reference to the former. For instance, regarding biochemical models, one might know that metabolic processes are influenced by a certain class of genes and that they involve instantaneous relations, whereas transcription processes are controlled by another class and involve a time delay. Knowledge of this sort can constrain significantly the number of models that are included in the search space, and thus increase the chances of finding a candidate that scientists find acceptable. Analogous knowledge about which types of variables can occur in

which types of equations can place similar constraints on the search for quantitative models.

Another challenge that we have encountered in our research has been the need to translate existing models into a declarative form that our discovery methods can manipulate. In response, we have started to develop a modeling language in which scientists can cast their initial models and carry out simulations, but that can also serve as the declarative representation for our discovery methods. The ability to automatically revise models places novel constraints on such a language. We envision this software developing into an interactive discovery aide that lets a scientist specify initial models, focus the system's attention on particular data sets and on parts of the model it should attempt to improve, and generally control high-level aspects of the discovery process. Thus, future versions will need a graphical interface for creating models, editing them, and marking fragments that can be revised, as well as tools for displaying matches to data, linking to other knowledge bases, and tracking changes in models over time. Taken together, these extensions should produce a valuable aide for practicing scientists.

Naturally, we also hope to evaluate our approach to model revision on other aspects of photosynthesis regulation and other portions of the CASA model as additional data become available. A more serious test of generality would be application of the same methods to other scientific domains in which there already exist formal models that can be revised. In the longer term, we should evaluate our interactive system not only in its ability to increase the predictive accuracy of an existing model, but in terms of the satisfaction the system provides to scientists who use it for model development.

Acknowledgments

This work was supported by Grants NCC 2-1202, NCC 2-5471, and NCC 2-1335 from NASA Ames Research Center, and by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. We thank Arthur Grossman and C. J. Tu for the initial model of photosynthesis regulation and associated microarray data, Stephen Bay for implementing the system that analyzed these data, and Christopher Potter, Alicia Torregrosa, and Steven Klooster for access to their CASA model and related ecosystem data. Portions of this paper have appeared in *Proceedings of the Fourth International Conference on Discovery Science* and *Proceedings of the Pacific Symposium on Biocomputing.*

References

Chown, E. and Dietterich, T.G., 2000, A divide and conquer approach to learning from prior knowledge, in: *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 143–150.

Džeroski, S., and Todorovski, L., 1993, Discovering dynamics, in: *Proceedings of the Tenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 97–103.

Fayyad, U., Haussler, D., and Stolorz, P., 1996, KDD for science data analysis: Issues and examples, in: *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, pp. 50–56.

Forbus, K., 1984, Qualitative process theory, *Artificial Intelligence* 24:85–168.

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

Karp, P.D., 1990, Hypothesis formation as design, in: *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager and P. Langley, eds., Morgan Kaufmann, San Francisco, CA, pp. 275–317.

Kocabas, S., 1991, Conflict resolution as discovery in particle physics, *Machine Learning* 6: 277–309.

Kuhn, T.S., 1962, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, IL.

Kulkarni, D. and Simon, H.A., 1990, Experimentation in machine discovery, in: *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager and P. Langley, eds., Morgan Kaufmann, San Francisco, CA, pp. 255–274.

Langley, P., 1979, Rediscovering physics with BACON.3, in: *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 505–507.

Langley, P., 2000, The computational support of scientific discovery, *International Journal of Human-Computer Studies* 53:393–410.

Langley, P., Simon, H.A., Bradshaw, G.L., and Żytkow, J.M., 1987, *Scientific Discovery: Computational Explorations of the Creative Processes*, MIT Press, Cambridge, MA.

Lenat, D.B., 1977, Automated theory formation in mathematics, in: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 833–842.

Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., and Lederberg, J., 1980, *Applications of Artificial Intelligence for Organic Chemistry: The* DENDRAL *project*, McGraw-Hill, New York.

Mitchell, F., Sleeman, D., Duffy, J.A., Ingram, M.D., and Young, R.W., 1997, Optical basicity of metallurgical slags: A new computer-based system for data visualisation and analysis, *Ironmaking and Steelmaking* 24:306–320.

Newell, A., and Simon, H.A., 1956, The logic theory machine, *IRE Transactions on Information Theory* IT-2:61–79.

O'Rorke, P., Morris, S., and Schulenberg, D., 1990, Theory formation by abduction: A case study based on the chemical revolution, in: *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager and P. Langley, eds., Morgan Kaufmann, San Francisco, CA, pp. 197–224.

Ourston, D. and Mooney, R., 1990, Changing the rules: a comprehensive approach to theory refinement, in: *Proceedings of the Eighth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA, pp. 815–820.

Polanyi, M., 1958, *Personal Knowledge: Towards a Post-Critical Philosophy*, University of Chicago Press, Chicago, IL.

Potter C.S. and Klooster, S.A., 1997, Global model estimates of carbon and nitrogen storage in litter and soil pools: Response to change in vegetation quality and biomass allocation, *Tellus* 49B:1–17.

Potter, C.S. and Klooster, S.A., 1998, Interannual variability in soil trace gas (CO_2, N_2O, NO) fluxes and analysis of controllers on regional to global scales, *Global Biogeochemical Cycles* 12:621–635.

Rajamoney, S., 1990, A computational approach to theory revision, in: *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager and P. Langley, eds., Morgan Kaufmann, San Francisco, CA, pp. 225–254.

Rogers, S., Langley, P., and Wilson, C., 1999, Learning to predict lane occupancy using GPS and digital maps, in: *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, ACM Press, San Diego, pp. 104–113.

Rose, D., and Langley, P., 1986, Chemical discovery as belief revision, in: *Machine Learning* 1:423–451.

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

Saito, K. and Nakano, R., 1997, Law discovery using neural networks, in: *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 1078–1083.

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

Simon, H.A., 1966, Scientific discovery and the psychology of human problem solving, in: *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, R.G. Colodny, ed., University of Pittsburgh Press, Pittsburgh, PA.

Shrager, J., and Langley, P., eds., 1990, *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufmann, San Francisco, CA.

Shrager, J., Langley, P., and Pohorille, A., 2002, Guiding revision of regulatory models with expression data, in: *Proceedings of the Pacific Symposium on Biocomputing*, Lihue, Hawaii, pp. 486–497.

Todorovski, L., and Džeroski, S., 2001, Theory revision in equation discovery, in: *Proceedings of the Fourth International Conference on Discovery Science*, Springer, Heidelberg, Germany, pp. 389–400.

Washio, T. and Motoda, H., 1998, Discovering admissible simultaneous equations of large scale systems, in: *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA, pp. 189–196.

Żytkow, J.M. and Simon, H.A., 1986, A theory of historical discovery: The construction of componential models, *Machine Learning* 1:107–137.

Żytkow, J.M., Zhu, J., and Hussam, A., 1990, Automated discovery in a chemistry laboratory, in: *Proceedings of the Eighth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA, pp. 889–894.