Pacific Symposium on Biocomputing 7:486-497 (2002)

GUIDING REVISION OF REGULATORY MODELS WITH EXPRESSION DATA

JEFF SHRAGER[a] and PAT LANGLEY
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

ANDREW POHORILLE
Center for Computational Astrobiology and Fundamental Biology
NASA Ames Research Center, M/S 239-4, Moffett Field, CA 94035

BioLingua is a computational system designed to support biologists' efforts to construct models, make predictions, and interpret data. In this paper, we focus on the specific task of revising an initial model of gene regulation based on expression levels from gene microarrays. We describe BioLingua's formalism for representing process models, its method for predicting qualitative correlations from such models, and its use of data to constrain search through the space of revised models. We also report experimental results on revising a model of photosynthetic regulation in Cyanobacteria to better fit expression data for both wild and mutant strains, along with model mutilation studies designed to test our method's robustness. In closing, we discuss related work on representing, discovering, and revising biological models, after which we propose some directions for future research.

1 Introduction and Motivation

There is general agreement that scientists need computational tools to assist in analyzing the rapidly increasing amount of biological data. Unfortunately, most existing software makes only limited contact with the methods that practicing biologists use in formulating and evaluating their models. In particular, most computational tools in biology have focused on knowledge-lean methods for data analysis, such as clustering, whereas biologists typically reason in a knowledge-rich manner using models of biological processes.

In this paper, we describe BioLingua, a suite of computational tools designed to assist working biologists in building and reasoning about their process models. Our goal in developing the system has been to match the ways in which biologists think about explanatory models, rather than to apply existing algorithms to available data in ways seldom pursued by biologists themselves. Working biologists, like other scientists, use data and models interactively, utilizing their models to interpret new experimental results and in turn revising these models in response to observations.

[a] Also affiliated with Department of Plant Biology, Carnegie Institution of Washington. Email: jshrager@andrew2.stanford.edu

In the sections that follow, we describe our initial version of BioLingua, which supports data intepretation and model revision in the arena of regulatory models. We start by defining the task of revising an intitial model given expression data and then report on BioLingua's approach to representing models, using them to make predictions, and carrying out heuristic search through the space of candidate models. After this, we discuss related work on representing knowledge about biological processes and discovering models that encode them. In closing, we note some limitations of our system and suggest directions for future research on computational discovery aides for biologists.

2 The Task of Revising Regulatory Models

One important facet of biological theory concerns the regulation of gene expression. Although scientists understand the basic mechanisms through which DNA produces proteins and thus biochemical behavior, they have yet to determine most of the regulatory networks that control the degree to which each gene is expressed. However, for particular organisms under certain conditions, biologists 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.[b]

There are two typical ways in which expression data are used to extend knowledge about regulatory mechanisms. The most common computational approach involves the use of clustering to infer which genes occur in coregulated classes. This knowledge-lean approach lets one reduce the high dimensionality of microarray data to a manageable level, but the result is typically descriptive rather than explanatory in nature. 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.

We have designed BioLingua to combine the best aspects of these two approaches to regulatory model discovery. We can state the task in semi-formal terms as:

- *Given*: a partial model of gene regulation for some organism;
- *Given*: data about the expression levels of relevant genes;
- *Given*: knowledge of biological processes that regulate gene expression;
- *Find*: an improved regulatory model that explains the data better.

[b] The distance between these measures and actual biochemical activity is considerable, but they still provide valuable information to biologists.

Computational tools that support this task will let biologists use microarray data both to test their regulatory models and to revise them in response to relevant observations.

3 An Approach to Regulatory Model Revision

Now that we have stated the task of revising an initial regulatory model based on microarray data, we can describe the approach that BioLingua takes to this discovery problem.

3.1 Representing Models of Gene Regulation

Before we can develop algorithms to improve regulatory models, we must select some representation for those models. Most work in machine learning and data mining, including that in biological domains, draws on representational formalisms like decision trees, logical rules, or Bayesian networks that were designed by artificial intelligence researchers themselves. These formalisms are often adequate for representing complex regularities and making accurate predictions, but they make little or no 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. In biology talks and publications, these models are often depicted graphically. Figure 1 presents one such model, which we obtained from a plankton biologist, that aims to explain why Cyanobacteria bleaches when exposed to high light conditions and how this protects the organism. Each node in the model corresponds to some observable or theoretical variable; each link stands for some biological process through which one variable influences another. Solid lines in the figure denote internal processes, while dashes indicate processes connected to the environment.

The model states that changes in light level modulate the activity of DFR, a protein hypothesized to serve as a sensor. This in turn activates NBLR, which then reduces the number of phycobilisome (PBS) rods that absorb light, which is measurable photometrically as the organism's greenness. The reduction in PBS serves to protect the organism because the reduced PBS array absorbs less light, which can be damaging at high levels. The organism's health under high light conditions can be measured in terms of the culture density. The model also posits that DFR impacts health through a second pathway, by influencing an unknown response regulator RR, which in turn down regulates expression of the gene products psbA1, psbA2, and cpcB. The first two positively influence the level of photosynthetic activity (Photo), which, if left unaltered, would also damage the organism.

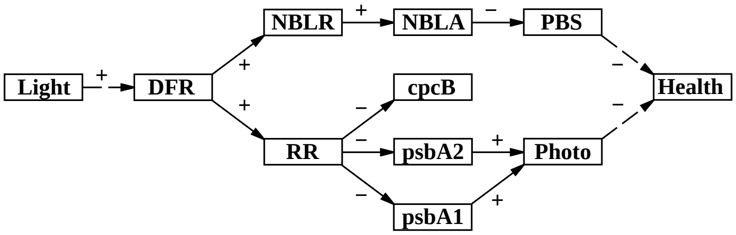

Figure 1: An initial model for regulation of photosynthesis in Cyanobacteria.

Note that this model, although incorporating quantitative variables, is qualitative in that it specifies directions of influence but not their degree. For instance, one causal link indicates that increases in NBLR will increase NBLA, but it does not state whether that relation obeys a linear or some other law, nor does it specify any parameters. We have focused on 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.

Another characteristic of the model is that it is both partial and abstract. The biologist who proposed this model made no claim about its completeness, and clearly viewed it as a working hypothesis to which additional genes and processes should be added as indicated by new data. Some processes are abstract in the sense that they denote entire chains of subprocesses. For instance, the link from DFR to NBLR stands for a signaling pathway, the details of which hold little relevance at this level of analysis. The model also includes abstract variables like RR, which denotes an unspecified gene, or possibly a set of genes, that acts as an intermediary controller. BioLingua's formalism lets it express such partial, abstract, and qualitative models of the type that biologists propose and reason about.

3.2 Microarray Data on Gene Regulation

Like many other researchers, we are excited about the potential of cDNA microarrays for elucidating biochemical processes. Briefly, these devices measure the expression level for hundreds to thousands of an organism's genes, as reflected by the concentration of mRNA for each gene relative to that in a control condition. One can collect such measurements under different environmental conditions (e.g., clean vs. polluted water), for different organisms (e.g., healthy vs. cancerous tissue), or for different points in time.

We have access to such microarray data for several strains of Cyanobacteria under high light conditions that cause the organism to bleach and reduce its photosynthetic activity over a period of hours. These data include measurements of the expression levels for about 300 genes believed to play a role in photosynthesis, although we have focused on those genes mentioned in the model. We have array data collected at 0, 30, 60, 120, and 360 minutes after high light was introduced, with four replicated measurements at each time point. The dimensionality of these data, and thus the number of parameters required in a numeric model, is much higher than the number of observations, providing another reason to favor qualitative models over quantitative ones.

3.3 Making Predictions and Evaluating Models

BioLingua needs some procedure to map a biological model like that in Figure 1 onto the microarray data we have available. Since its models are qualitative, they cannot directly predict the continuous expression levels, but they can predict which variables should be correlated and the direction of those correlations. For each pair of variables (nodes) in a model, the system enumerates the paths that connect those variables. BioLingua transforms each such path into a predicted correlation by multiplying the signs on its links and, when the predictions for all paths between two nodes agree, predicting that correlation.[c]

However, when the correlations predicted by two or more paths disagree, BioLingua must resolve the ambiguity in some manner. In a quantitative model, each path would have its own degree of influence, and one could sum their effects to determine the outcome. Lacking such quantitative information, the system can still annotate the model to indicate that the positive (or negative) paths are dominant, and thus predict a positive (or negative) correlation. This extended formalism lets any qualitative model predict a positive or negative correlation for each pair of observed variables, provided one is willing to pay the cost of adding assumptions about dominance. For example, the model in Figure 1 has three paths between the expression levels for DFR and Health. The product of signs on each path is positive, meaning that they each predict a positive correlation between the two variables. However, if the link from NBLA to PBS were positive, this path would make a different prediction and the model would need a dominance annotation to resolve the ambiguity.

This procedure lets BioLingua generate qualitative correlations between pairs of variables in a given model. Naturally, the system can compare these

[c] Note that some paths pass through unobservable variables like RR; although we cannot measure such terms' values, that does not keep BioLingua from utilizing them in predictive paths between observable variables like DFR and psbA1.

predictions to the observed correlations, which it computes from corresponding expression levels in the arrays across different time steps. BioLingua treats any correlation that fails a significance test, in this case $p < 0.05$, as zero. The system incorporates these matches against the data in its evaluation metric for models. However, it also includes a measure of model complexity which favors simpler models and a term which favors models that make more predictions (i.e., a Popperian bias toward hypotheses that are easier to reject), which we found necessary to guard against degenerate models. The specific function used to evaluate candidates is

$$E = B(variables) + B(links) + B(annotations) + B(errors) - B(predictions) \ ,$$

where $B(X)$ denotes the total number of X (e.g., links or errors) times the number of bits needed to encode X. In this scheme, each variable and each link requires 4 bits, each disambiguation annotation requires 0.1 bit, and each prediction error and each prediction requires 3 bits. The resulting measure, which is similar to minimum description length, gives the overall quality for each model.

3.4 Revising Regulatory Models to Explain Microarray Data

As with most research on computational knowledge discovery, one can view the revision of biological models in terms of heuristic search through a space of candidate models. This framework requires one to make a number of design decisions, including the state from which to initiate the search, the operators used to generate new states, the knowledge used to constrain these operators' application, the evaluation metric used to select among competing states, the overall scheme for search control, and the criterion used to halt the search.

Biologists often have some abstract qualitative model in mind at each stage of their research. BioLingua takes such a model as the starting point for its search process. Some natural operators for revising such a model include adding a signed link, removing a link, and reversing the sign on a link. In the current implementation, BioLingua's evaluation function for selecting among models is simply the measure of model quality E described earlier. The control scheme that utilizes this function is greedy search through the model space, with failure to improve on the evaluation metric as the halting criterion.

For example, to generate an improved regulatory model for the photosynthetic process in Cyanobacteria under high light, BioLingua starts from the model in Figure 1. This model's 11 variables and 12 causal links lead to some 350 one-step revisions that produce distinct models, resulting from link reversals, link additions, and link deletions. The system generates each of these candidates, calculates their E scores given the expression data, and selects the best one as the current model. It then repeats this process, continuing until further changes fail to yield improvements in the evaluation metric.

4 Experimental Results on Photosynthetic Regulation

Ultimately, BioLingua's success as a discovery system will depend on whether it can use expression data to improve biological models. Here we report initial experiments designed to test the program's abilities on this dimension.

4.1 Improving Models of Wild and Mutant Cyanobacteria

We have already described an initial model, shown in Figure 1, of bleaching in Cyanobacteria that we obtained from biologists, along with expression data on the genes that regulate this process over time. The data lead to 18 positive correlations and 10 negative correlations among the observed expression levels.

When given this initial model and these qualitative data, BioLingua's revision module carries out its greedy search through the model space, taking eight steps and examining 2382 candidates along the way. Additional revisions lead to no improvement in the evaluation function, causing the system to halt. Figure 2 shows the final revised model that results from this search process, which matches the observed expression levels better than the starting model and has a better evaluation score ($E = -46$ rather than $E = 12.2$).

This model differs from the initial one in some important ways. These include deletion of the links from DFR to NBLR, from psbA1 to Photo, from RR to psbA2, and from RR to cpcB. The revised model also contains three new links, indicating a positive influence from cpcB to NBLR and negative influences from psbA1 to psbA2 and from psbA2 to cpcB. The revision process has also changed signs on the links from RR to psbA1, from PBS to Health, and from Photo to Health.

In addition to proposing regulatory models for wild strains of an organism, biologists also desire to model mutant strains. We have access to array data for a nonbleaching mutant of Cyanobacteria under the same high light conditions as for the wild strain. Because such a mutant presumably differs genetically from the wild organism in at most a few ways, it seems natural to utilize BioLingua's revision module to formulate a model of the mutant's regulatory processes. In this case, the system considers 2270 candidates while taking nine steps through the model space. Figure 3 presents the resulting model, which has a better score ($E = -24.6$) than the initial one ($E = 12.2$).

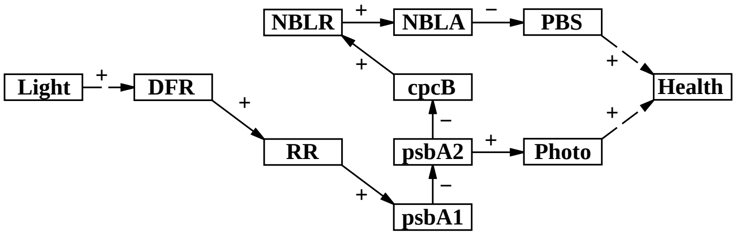

Figure 2: A revised model for regulation of photosynthesis in wild Cyanobacteria.

There are a number of differences between the revised model for the mutant strain and the initial model. These include deletion of the links from DFR to RR, from RR to psbA2, from RR to cpcB, and from psbA1 to Photo. The mutant model also specifies three new links, indicating positive influences from psbA1 to cpcB and from cpcB to psbA2, along with a negative influence from NBLA to RR. The revision mechanism has also changed signs on the links from psbA2 to Photo and from Photo to Health.

These revised models have some biological plausibility, but they also have problematic aspects. Generally speaking, it seems plausible that DFR influences photosynthetic activity through NBLR (in the wild strain) or a psbA1 cascade (in the mutant strain), and additional experiments could test these proposals. On the other hand, in both cases the revision process produced models with cascades whereas the initial model had separate influences, specifically from RR. Although such chains are not impossible, there is no reason to prefer such structures. Additional knowledge, either in the form of biological constraints or an improved evaluation metric, could resolve this ambiguity.

4.2 Robustness of the Approach

Although the previous runs demonstrate BioLingua's relevance to problems in model revision that arise among practicing biologists, they do not provide evidence of its robustness. To evaluate BioLingua's revision module along this dimension, we designed an experiment to determine whether the quality of the final revised model degrades gracefully with decreasing correctness of the initial model. Thus, we took the revised model from Figure 2 as our target T and generated different initial models by taking random steps through the model space. In this manner, we generated ten distinct models that differed from T by one step, another ten that differed by two steps, and so forth, halting at five steps from the target. We then ran the revision algorithm on each initial model with the expression data that produced the model in Figure 2.

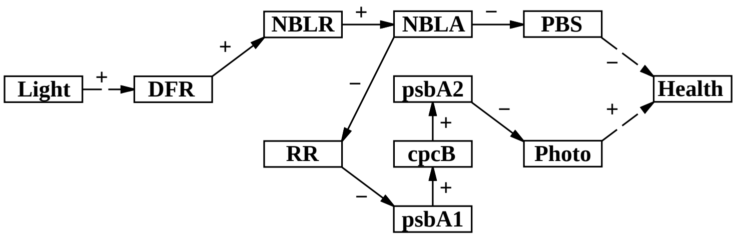

Figure 3: A revised model for regulation of photosynthesis in mutant Cyanobacteria.

We measured two dependent variables as a function of distance from the target model. The first involved the revised model's accuracy at predicting qualitative correlations, specifically the number of correctly predicted correlations or non-correlations over the total number of possible correlations. The second was simply the distance (number of steps in the search space) between the revised model and the target model T. We hypothesized that both measures would get worse, on average, with distance between the initial and target models, but that this degradation would be graceful.

The results were generally consistent with our expectations. The predictive accuracy of the target model on the expression data was 94 percent, whereas the revised models from runs starting one, two, three, four, and five steps from the target had average accuracies of 84, 79, 78, 65, and 63 percent, respectively. Similarly, the average distance of these revised models from the target, in terms of steps through the model space, was 3.5, 3.5, 5.9, 4.4, and 5.0, respectively. Thus, the method's behavior degraded as the revision task became more difficult, but this occurred in a graceful manner.

5 Related Research on Computational Discovery

Our approach to computational biological discovery builds on three previous lines of research. The first framework has focused on the explicit representation of knowledge about biological pathways. For instance, Karp et al.'s EcoCyc[1] encodes most established pathways for E. Coli and lets users display this knowledge graphically. Kanehisa[2] reports another effort that has produced KEGG, which codifies similar knowledge about a range of organisms. The knowledge stored in these systems is impressive, including information about metabolic pathways, regulatory pathways, and molecular assemblies, but their ability to reason over this knowledge remains limited. Tomita et al.[3] describe another framework, E-Cell, which stores similar knowledge and includes mechanisms for predicting behavior, but even E-Cell lacks the ability to revise its models in response to observations, which is BioLingua's central feature.

A second framework has focused explicitly on the discovery of biological knowledge from data. We have already contrasted our approach with the more common technique of clustering microarray data in a knowledge-lean manner, but there exists some other work on constructing process explanations from such data. For example, Koza et al.[4] use heuristic search methods to estimate, from time-series data about concentrations, the structure and parameters of a metabolic model. Zupan et al.[5] describe GenePath, a system that comes somewhat closer to our approach in that it combines biological knowledge and data about the effects of mutations to propose qualitative genetic networks. Hartemink et al.,[6] although not focused on discovery, propose a similar notation for encoding regulatory models and another evaluation metric that could direct search through the model space.

A third research framework has focused not on constructing models from scratch but rather on revising existing theories to improve their fit to data. For example, Ourston and Mooney[7] present a method that uses data to revise models stated as sets of propositional Horn clauses, whereas Towell[8] reports a related approach transforms such models into multilayer neural networks, then uses backpropagation to improve their fit to observations. Our technique comes closer to Karp's HypGene,[9] which uses qualitative phenomena to revise a model cast in biological terms, but which differs considerably in its formalism and reasoning mechanisms. This framework has emphasized supervised rather than unsupervised data, but it shares the notion of revising an initial model.

Each of these frameworks has clear merits. Our research is novel in that it combines these three themes into a single system for the computational discovery of biological knowledge.

6 Concluding Remarks

BioLingua is a computational tool kit designed to assist biologists in stating process models, using those models to make predictions, interpreting observations in light of those predictions, and improving their models in response. Our initial work has focused on revising a given regulatory model to better fit observed expression levels, an approach that differs considerably from the knowledge-lean methods typically applied to such data.

We illustrated BioLingua's application to this task in the context of a particular model of photosynthetic regulation in Cyanobacteria and expression data collected for that organism. We presented the system's formal representation for biological process models, a method that uses such models to predict qualitative correlations between expression levels, and an algorithm that carries out heuristic search through the space of regulatory models, guided by data and a bias toward simpler models. In addition, we demonstrated the system's revision of an initial model of photosynthetic regulation, given expression data for wild and mutant Cyanobacteria. We also studied BioLingua's ability to recover a model's structure after mutilating it to varying degrees, and the system exhibited reasonable robustness on this task.

Although our results to date are encouraging, we must extend BioLingua in a number of directions before it can become a useful tool for biologists. For example, the current system can add, remove, and reverse causal links to the initial model, but it cannot introduce new variables that correspond to observed expression levels for known genes, which seems desirable. Achieving this functionality means adding a new revision operator and thus enlarging the space of candidate models, which in turn will require an improved search mechanism. This expanded search process would benefit from interaction with biologists, who could help to guide the decision process in cases where different models have similar scores.

Future versions of the system should support link types that correspond to additional biological concepts. For example, BioLingua should distinguish between metabolic processes, which are effectively instantaneous, and regulatory processes, which typically take place over time. This distinction will also mean extending our formalism and prediction mechanism to support time-delayed effects. One response to this challenge comes from qualitative physics, which describes dynamic systems in terms of qualitative differential equations. This approach is consistent with our bias toward qualitative models.

A more fundamental issue concerns BioLingua's current modeling formalism. Although biologists state some models in terms of measurable statistical variables, such as gene expression levels, they often describe an organism's behavior in terms of mechanical processes that operate on individual molecules. Karp's work[9] on modeling the Tryptophan operon provides one approach to representing such mechanisms. Future versions of BioLingua should support the ability to make statistical predictions from such mechanical models, and thus make better contact with biologists' conceptual repertoire.

In the longer term, we envision BioLingua developing into an interactive discovery aide that lets a biologist specify initial models, focus the system's attention on particular data and parts of those models it should attempt to improve, select among candidate models with similar scores, and generally control high-level aspects of the discovery process. Combined with other planned extensions, this facility should make BioLingua a more valuable tool for practicing biologists.

Acknowledgements

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

References

1. P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, S. Paley, and A. Pellegrini-Toole, "EcoCyc: Electronic Encyclopedia of E. coli genes and metabolism." *Nucleic Acids Research*, **28**, 56 (2000).
2. M. Kanehisa, "A database for post-genome analysis." *Trends in Genetics*, **13**, 375–376 (1997).
3. M. Tomita, K. Hashimoto, K. Takahashi, T. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. Hutchison, "E-Cell: Software environment for whole cell simulation." *Bioinformatics*, **15**, 72–84 (1999).
4. J. R. Koza, W. Mydlowec, G. Lanza, J. Yu, and M. A. Keane, "Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming." *Pacific Symposium on Biocomputing*, **6**, 434–445 (2001).
5. B. Zupan, I. Bratko, J. Demsar, J. R. Beck, A. Kuspa, and G. Shaulsky, "Abductive inference of genetic networks."*Proceedings of the Eighth European Conference on Artificial Intelligence in Medicine* (Cascais, Portugal, 2001).
6. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks." *Pacific Symposium on Biocomputing*, **6**, 422–433 (2001).
7. D. Ourston and R. Mooney, "Changing the rules: A comprehensive approach to theory refinement." *Proceedings of the Eighth National Conference on Artificial Intelligence*, 815–820 (AAAI Press, Boston, 1990).
8. G. Towell, *Symbolic Knowledge and Neural Networks: Insertion, Refinement, and Extraction.* Doctoral dissertation, Computer Sciences Department, University of Wisconsin, Madison (1991).
9. P. D. Karp, "Hypothesis formation as design." In *Computational Models of Scientific Discovery and Theory Formation*, Ed. J. Shrager and P. Langley (Morgan Kaufmann, San Francisco, 1990).