AN INTEGRATED SYSTEM FOR SYNTHESIS AND EVALUATION OF BIOLOGICAL MODELS

Andrew Pohorille
NASA-Ames Research Center
Moffett Field, CA 94025

Jeff Shrager, Stephen Racunas
Institute for the Study of Learning and Expertise
2164 Staunton Court
Palo Alto, CA 94306

Karl Schweighofer
SETI Institute
515 N. Whisman Road
Mountain View, CA 94043

Abstract - **We have developed software that enables biologists to specify, evaluate, refine and share biological models and hypotheses. Using this software, astrobiologists, earth and biomedical scientists can bring data and background knowledge to bear in evaluating the consistency and correctness of their models of biological and ecological processes, and in refining them whenever the models are incomplete or do not agree with data. Our core technologies include a graphical user interface that enables scientists to express and analyze models, tools for extracting constraints on models from experiments, databases and background knowledge, and computational methods for the evaluation and refinement of models, subject to these constraints. The software also includes capabilities that foster scientific collaborations between researchers in different subfields and geographical locations. Our system has been applied to problems aimed at discovering the molecular, genetic, and biochemical mechanisms involved in controlling evolution, metabolic diversity, and the ability of life to survive in space.**

I. INTRODUCTION

Recent revolutionary advances in molecular, cellular and systems biology open remarkable new opportunities not only for medicine, pharmacology and biotechnology, but also for NASA-related research. Biological techniques that were not available only a few years ago can now be used to gain new insights into the evolution of simple forms of life, its interactions with the biosphere under different conditions, and its capability to survive in different environments on Earth and in space. These techniques involve high-throughput genome sequencing and measurements of levels of mRNA, proteins and metabolites in cells. All these methods are data-rich but relatively knowledge-poor – they provide large amounts of data, but usually require sophisticated analysis and interpretation to yield answers to questions asked by biologists. Examples of such questions are: How is the evolution of a genome reflected in metabolism and phenotype? How do members of a microbial community interact and respond to environmental change? Which genomic and metabolic characteristics determine the capability of microorganisms to survive long-term and possibly interplanetary space travel? These questions can be fully answered only by building data based, causal models that link environmental conditions to gene expression and, ultimately, to cell behavior modulated by gene regulation.

Currently, the majority of bioinformatics tools deals only with one type of data (e.g. gene sequences or gene expression measurements) and do not utilize other sources of information. Even though thousands of measurements from high-throughput

experiments are often available, they all come from only a few biologically independent samples, forcing biologists to reason from statistically sparse data that usually have low signal/noise ratio. These limitations are particularly relevant to astrobiology because collecting many independent samples in field and space experiments might be expensive, technically difficult and, in some cases, impossible. This means that models of cellular or ecological processes entirely based on a single experiment or a small number of experiments, no matter how well designed, are unlikely to be sufficiently accurate to have high explanatory and predictive power. Including many sources of information can greatly improve the reliability of computational models. Also, the availability of easy to use computational tools for complex database searches and knowledge manipulation eliminates the need for computer programmers to mediate between biologists and their data. This, in turn, increases productivity of experimentalists involved in building and testing biological models.

To meet the bioinformatics needs of NASA, we have been developing computational tools that enable biologists to specify, evaluate, refine and share biological models and hypotheses. Astrobiologists, earth and biomedical scientists can bring data and background knowledge to bear to evaluate the consistency and correctness of their models of biological and ecological processes, and to aid in refining them whenever the models are incomplete or do not agree with data. Our core technologies include a graphical user interface that enables scientists to express and analyze models, tools for extracting constraints on models from experiments, databases and background knowledge, and computational methods for the evaluation and refinement of models subject to these constraints. These methods operate on genomic, proteomic, and metabolic data, higher-level knowledge and representations of metabolic and regulatory networks. In addition, our software includes capabilities that foster scientific collaborations by allowing researchers in different subfields and geographical locations to share their knowledge and discoveries, and work together towards achieving common scientific goals.

The reminder of the paper consists of four sections. In the next section we provide a brief background for non-biologists on the processes that need to be modeled or yield background knowledge. This is followed by an overview of our knowledge discovery system and a brief description of its main component. In Section IV we compare our approach to other, related work. Further, we illustrate in a few examples how important research areas in astrobiology, can benefit from using our system. We close with conclusions and a brief outline of the extension of our system that we plan to implement.

II. BIOLOGICAL BACKGROUND

Genetic information about a cell is encoded in DNA forming the genome. At present, sequencing complete genomes of microorganisms is routine. Once the sequence is available, individual genes are identified and proteins coded by these genes are assigned putative general functions. Usingn this computational procedure, called gene annotation, approximately 70% of all genes in the genome can be identified.

More recently, applications of genomic sequencing have been extended to microbial ecology as a strategy for assessing the genetic and functional diversity of uncultured organisms sharing the same environment [1]. In this case, what is being sequenced is a metagenome - the collective genomes of the microorganisms recovered from a sample in a given environment. The goal is to identify the members of a microbial community and to understand how they interact with each other. This is of great interest to astrobiology in relation to understanding the historical record of how microorganisms shaped the planetary environment and how microbial communities can survive in space.

The genomic sequence is only the starting point for understanding cell behavior in response to signals (perturbations) that might be physical (light, temperature, radiation, gravity) or chemical (levels of chemicals supplied from the environment or produced inside the cell). Cellular responses are

precisely regulated: through signal transduction pathways, these perturbations affect the levels of regulatory proteins, which activate the production of RNA transcribed by specific genes. Most RNA produced during transcription is translated into proteins, and the level of each protein so generated is considered proportional to the amount of its RNA transcript. The combined processes of translation and transcription are called gene expression. A large fraction of the proteins produced in this process catalyze chemical reactions, mediate transport or regulate future expression of different genes, often forming a complex feedback network. This network of cellular reactions supporting life, called a metabolic network, defines a cell's ability to self-maintain, grow, utilize nutrients and secrete products. Taken together it describes much of cell behavior. The complete metabolic network is never active at once; environmental signals (e.g. light or supply of different nutrients) influence, largely *via* regulatory proteins, which parts of the metabolic network are active under particular conditions. From this knowledge it is possible to calculate the fluxes of metabolites, and in particular the rates of production of compounds that are released to biosphere (e.g. oxygen) or used as nutrients by other microorganisms (e.g. in microbial mats).

For microorganisms, large parts of metabolic networks are inferred from the annotated genome exploiting similarities in metabolic pathways between related organisms. Gene regulatory logic, which governs the activity of the metabolic network, is in general not known even for simple organisms and needs to be reconstructed from the data and background knowledge. The primary sources of data for these tasks are high-throughput measurements of gene expression in a cell, defined by the levels of its RNA transcripts. This is done using DNA microarray technology or other, similar biological techniques. Microarray data, however, are inherently noisy, reflecting imperfections in the experimental technique, the natural heterogeneity of biological samples and the stochastic character of cellular processes. Thus, reliable predictions based entirely on these experiments are difficult, especially because the relations between the levels of RNA transcripts and active proteins are not always as simple as postulated. A more direct approach is to measure the levels of proteins in a cell (proteomics) [2,3]. In this approach, however, there are still unresolved issues with rapid identification of a large number of proteins. The rates of production of metabolites can also be measured using the recently developed techniques of metabolomics [4]. However, large-scale identification of metabolites still remains a major challenge.

High-throughput studies can be augmented by other experiments. For example, genetic engineering can be used to investigate how cellular behavior is affected after removing, adding or replacing specific genes in the genome. When successful, these experiments are highly informative, but many of them yield modifications that are either lethal or silent.

III. MODEL DISCOVERY SYSTEM

A. Overview

The brief biological background illustrates the conceptual foundations of our system – a variety of experimental techniques provide information relevant to understanding complex biological processes. Each technique has its strengths and weaknesses but none is sufficient to describe the process completely. To do so, one needs to combine results from different experiments, and analyze them using appropriate computational tools.

The general problem to be solved can be stated as follows: given (1) experimental data collected by the researcher, such as gene sequences, gene expression or protein levels under different conditions, (2) relevant background information, and (3) a hypothesis or a model, evaluate support for this model or hypothesis and, if required, modify it to obtain a better agreement with the data and background knowledge. A hypothesis is understood to be a statement about a trait of the organism that can be directly linked to its genomic or metabolic characteristics. For example, the biologist might hypothesize that under specific but different conditions a given organism can grow either aerobically or anerobically or use different

nutrients. In this context, a model refers to a complete, partial or approximate description of regulatory logic and metabolic fluxes in a cell under given conditions, and can be considered as a special case of a hypothesis.

As shown in Fig. 1, the model discovery process consists of four phases. In the Hypothesis Formulation Phase (I), a hypothesis is expressed in formal language using an intuitive, biologist-friendly interface. If the hypothesis cannot be summarily refuted or confirmed using the results of past analyses, the Constraint Formation Phase (II) is carried out. In this phase, direct experimental data, ancillary data, and background knowledge that originally take many different forms are converted into a set of uniformly represented constraints, each of which is assigned a range of certainty. This is followed by the Evaluation Phase (III), in which the hypothesis is evaluated according to its agreement with data and constraints. During this process, variations of the original hypothesis that fit data and constraints better may be discovered. Next, in the Visualization Phase (IV), the hypothesis, its variants, and the appropriate supporting and contradictory evidence are presented to the user in an easy to understand summary.

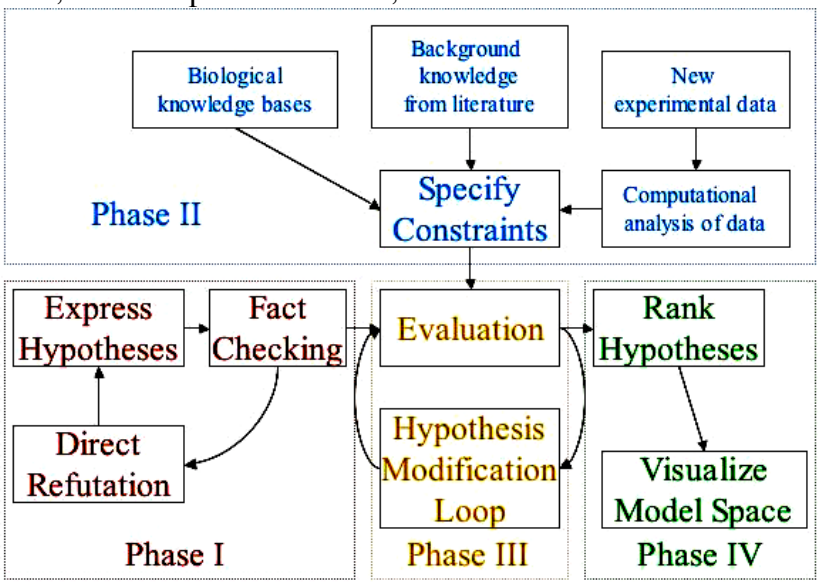

Fig. 1. The four phases of hypotheses design and validation.

B. Hypothesis Formulation Phase (I)

To effectively aid biologists in their work, we must first speak their language. The fundamental unit of biological reasoning is the formal hypothesis. Thus, we must develop an interface for expressing hypotheses using conventions familiar to, and comfortable for, biologists. This means that the interface must accept hypotheses expressed either as reaction pathway diagrams, textual descriptions, or a combination of both. To support hypothesis evaluation, it is helpful to have a web-enabled interface that guides users in building syntactically correct and semantically accurate hypotheses. This will be done through HyBrow system that we have just developed.

The HyBrow prototype (www.hybrow.org) allows users to compose sentences by specifying entities and the relationships between them from a structured hypothesis ontology. A researcher can also sketch a diagram representing interactions and spatial relationships among cellular components using a separate graphical interface. We will leverage this work to create a web-based platform specifically designed to handle the hypotheses of astrobiologists. A demo screenshot showing parallel editing of both graphical and textual representations of a hypothesis about the regulation of metabolism in the presence of alternate energy sources appears in Fig. 2.

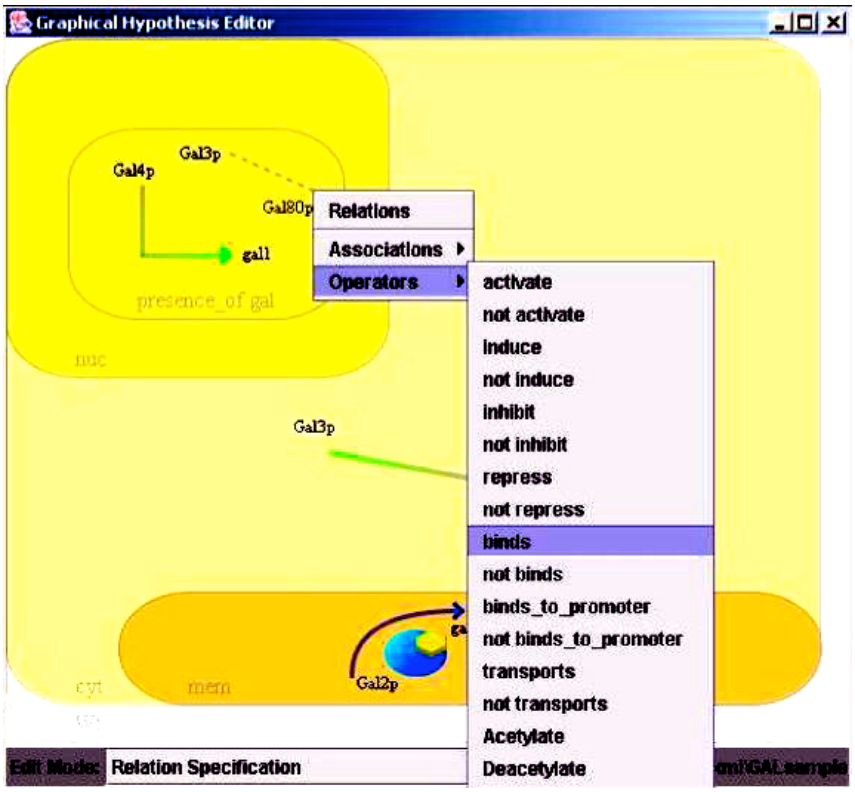

Fig. 2. Graphical user interface in HyBrow allows users to compose their hypotheses graphically or as a structured text.

C. Constraint Formation Phase (II)

Model constraints provide the basis for establishing evidence for partial orderings over models and hypotheses based on support that they received from current data and previous information. The user can supply constraints explicitly, but usually they are extracted from experimental data obtained by the biologist, literature and databases or knowledge bases using a suite of specialized computational tools.

Constraints arising from high-throughput data, such as gene expression, gene sequencing and large-scale homology comparisons are determined from analyses that utilize bioinformatics tools available through web-based interfaces. Constraints from auxiliary knowledge are harder to extract because they cover many types of knowledge, with myriad sources of uncertainty. Therefore, an efficient procedure for obtaining constraints and managing their uncertainties has to be based on an interactive environment, in which the knowledge and data are integrated, and the suite of computational tools necessary for their analysis is already in hand. For this purpose, we developed BioBike, a biology-based programming environment and a biological data repository [5]. BioBike provides integrated access to a number of data sources, including the Gene Ontology, the Kegg knowledge base, the BioCyc database for metabolic pathways, biological literature, as well as basic bioinformatics tools. BioBike embeds the biological knowledge and data in a frame-based, web-accessible programmable knowledge environment. This formalism allows for formulating and executing complicated queries to data and knowledge bases in a simple, natural way instead of carrying out painstaking searches that usually require programmer's assistance (for examples see [5]. For BioBike server see http://nostoc.stanford.edu.

The significance of constraints from background knowledge can be illustrated in an example from

our work [6]. To test a hypothesis about the existence of a direct causal link between two genes in a signaling pathway involved in cancer and apoptosis, biologists performed gene expression measurements. Our subsequent analysis revealed that several models of the signaling pathway that both involved and did not involve the link of interest described the results equally well. However, once results of previous, relevant experiments were taken into account, the hypothesis could be rejected with high confidence on the basis of the estimated Bayesian equivalent of the p-value, well known in classical statistics. This was because models that involved the direct link were inconsistent either with gene expression data or with background knowledge.

D. Evaluation Phase (III)

To evaluate biological hypotheses or models we currently use a deduction-based approach to biocomputation that semi-automatically combines knowledge, software, and data to satisfy biologists' goals expressed in a high-level logical language. The approach is implemented in a system called BioDeducta, which combines SNARK theorem prover with the BioBike integrated knowledge base and biocomputing platform [7]. The user expresses a high-level conjecture, representing a biocomputational goal query, without indicating how this goal is to be achieved. A subject domain theory, represented in SNARK's logical language, expresses the meaning of the terms in the conjecture in terms of the capabilities of the available resources and of the background knowledge necessary to link them together. If the subject domain theory enables SNARK to prove the conjecture—that is, to find paths between the goal and BioBike resources—then the resulting proofs record various solutions to the conjecture/query. The proofs also provide specific provenance for each result, indicating in detail how they were computed.

IV. RELATION TO OTHER WORK

Although different aspects of an interactive approach to discovering biological models, in which constraints derived from multiple sources of information are incorporated, have been examined previously, no existing framework allows for taking advantage of all of them simultaneously. Below we briefly review previous, representative efforts.

Methods for discovering causal models generally fall into two broad categories: score based and constraint based. In score based methods, one tries to infer models that best reproduce the observed data according to a score function such as posterior probability, likelihood, or mean squared error. For example, Friedman [8,9] and Hartemink et al. [10] attempted to learn Bayesian network models of gene regulation by focusing on models with the highest posterior probability given the observed data. Other researchers have proposed finding a system of linear differential or difference equations that best fit the observed data [11-17]. The main drawback of score-based approaches is that they can be highly over-parameterized as the models typically have many more parameters than the number of data samples from a biological experiment. A linear system of equations will typically have $O(N^2)$ parameters, where N is the number of variables, whereas the number of independent samples in gene expression studies is usually much smaller than N. Empirical studies show that when the number of samples is small, these methods usually do little or no better than random guessing [15,18].

Constraint-based methods seek models that are consistent with a set of constraints derived from the observed data. For example, Saavedra et al. [19] applied the Tetrad framework [20], which is based on matching conditional independence relations observed in the expression data to those entailed by the model. In our previous work [21], we used conditional independence relationship implied by a linear model to revise an existing model of gene regulation in response to new observational data. In GenePath, Zupan et al. [22] have attempted to develop causal models that are consistent with constraints derived from a set of interventions. Specifically, they use information on effects of deleting one or two genes from the genome to determine causal structure. Although this is a

powerful approach, it relies on the availability of deleting data.

Recently, researchers have begun to recognize the need for incorporating multiple sources of information to increase statistical power of their methods. For example, Holmes & Bruno [23] and Segal et al. [24] proposed to model both expression data and the nucleotide sequences of promoter regions. They developed a generative probabilistic model that explains both data sources as a function of an unobserved cluster variable. The main difficulty of this approach is to create a realistic probabilistic model capable of explaining all of the data types that one wishes to incorporate. Thus, adding a new type of data requires developing a new probabilistic model.

V. SELECTED APPLICATIONS

A. Spaceflight-induced Gene Expression Changes in Mice

In order to understand genomically induced physiological changes in higher organisms in response to space flight, we analyzed results from the first gene expression experiment on space-flown mice. By combining gene expression data for liver and kidney tissue with earlier physiological data on other space-flown organisms and ground-based information on mice, we found support for the hypothesis that pharmacokinetics in space is altered due to perturbed excretion of drugs rather than changes in their metabolism. We also found support for the hypotheses that stress response to space flight is increased, but immune response is reduced. In contrast, we found no support for a popular but unproven hypothesis that there is a universal set of gravity-activated genes. In general, our results demonstrate that, similar to bone, muscle, and immune function, alterations in liver parallel those seen in other mammals. This indicates that the mouse adequately models the spaceflight-induced physiological changes that are of concern in space medicine.

B. Light Adaptation of Cyanobacteria

A problem of considerable interest to biologists is how organisms adapt to their environmental niches. Following the work of Bahya et al. [25], who studied the genomic differences among the many strains of the cyanobacteria to understand their adaptation to niches of differing levels of light and nutrients, we addressed the same issue using BioDeducta. Among the cyanobacterium subspecies procholorococcus, one strain, ProMed4, is adapted to high light, living in the upper part of the ocean, whereas another strain, Pro9313, is adapted to lower light, living in somewhat deeper waters. Bhaya et al. were interested in identifying proteins (and genes that code for them) that are involved in this adaptation. One way to address this biological question is to ask which proteins in ProMed4 have no ortholog—that is, no gene of similar apparent function (based upon sequence similarity)—in Pro9313. One can get an even finer bead on this question by examining microarray expression results for the genes that produce those proteins, asking which of those genes unique to ProMed4 demonstrate a significant light response, and therefore might be called the high light adaptive genes. Unfortunately, microarrays for the prochlorococci have only recently been developed, so no such experimental work exists. However, there are a number of studies on the related freshwater cyanobacterium, synechocystis s6803. Indeed, research specific to light acclimation has been conducted in s6803 [26]. Going one step farther, one may focus specifically on the genes that are annotated as photosynthesis-related according to some formalization of gene function, such as the Gene Ontology. Armed with an appropriate subject domain theory, BioDeducta was able to solve the problem of identifying high light adaptive genes rapidly, and the solution agreed with that published Bahya et al. after extensive efforts.

C. Metabolic Diversity of Aquificales

To understand metabolic diversity of closely related organisms living in different environmental conditions, Awe carried out comparative analysis of four recently sequenced strains of microorganisms, called Aquificales. To determine similarities and differences in metabolisms of these organisms, we used the capabilities of BioBike to combine genomic sequences with Gene Ontology, metabolic pathway databases and prior knowledge about

several, related organisms. We discovered that, contrary to the expectation that Aquificales are strictly chemolithoautotrophic, some strains could grow heterotrophically and fix nitrogen. Since Aquificales are deeply rooted in the tree of life, understanding their metabolic capabilities provides important leads to how life evolved and adapted.

VI. CONCLUSIONS AND OUTLOOK

To advance a majority of NASA's biology related goals, such as reconstruction of the history of life on Earth and its interactions with the environment, understanding how life adapts to conditions in space and using living organisms for *in situ* resource utilization, it is required to combine genomic and physiological studies with sophisticated modeling of biological and ecological processes. We have developed computational tools for such modeling that can take advantage of many different data sources, and by doing so increase their predictive and explanatory power. We also demonstrated the utility of our system in several problems of interest to NASA.

Currently, we work on significant extensions of our system. One promising direction is to increase the capabilities of evaluating the reliability of models, presently handled by BioDeducta, by incorporating complete metabolic and regulatory models using flux balance analysis and Mixed Integer/Linear Programming techniques. Searching the space of models will be handled using novel Monte Carlo techniques. We also work on considerable extensions of the HyBrow system, content unification and consistency of databases, and greatly improved collaboration and visualization tools. We expect that, once these extensions are completed, the system will be unique in its expressiveness and evaluation power, and will become a standard tool for researchers involved in NASA-sponsored biological research.

ACKNOWLEDGMENT

The authors thank the NASA Advanced Information Science Research Program for supporting this work.

REFERENCES

[1] Rusch D.B. et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5(3): e77.

[2] Lipton, M.S. et al. (2002). Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags, *Proc. Natl. Acad. Sci. U.S.A.,* **99**, 11049-11054.

[3] Ram, R.J. et al. (2005). Community Proteomics of a Natural Microbial Biofilm, *Science*, **308**, 1915-1920.

[4] Phelps, T.J., Palumbo, A.V. and Beliaev, A.S. (2002). Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Curr. Opinion Biotechnol*., **13**, 20-24.

[5] Massar, J.P., Travers, M., Elhai. J. and Shrager, J. (2005). BioLingua: a programmable knowledge environment for biologists. *Bioinformatics*, **21**, 199-207.

[6] Chrisman, L, Langley, P., Bay, S. and Pohorille, A. (2003). Incorporating biological knowledge into evaluation of causal regulatory hypotheses. *Proceedings of the Pacific Symposium on Biocomputing*, 128-139.

[7] Shrager J, Waldinger R, Stickel M, Massar J (2007) Deductive Biocomputing. PLoS ONE 2(4): e339.

[8] Friedman, N. (2004). Inferring cellulat networks using probabilistic graphical models, *Science*, **303**, 799-805.

[9] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol*., **7**, 601-620.

[10] Hartemink, A,, Gifford, D., Jaakkola, T., and Young, R. (2002). Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Models. *Pacific Symposium on Biocomputing*, 437-449.

[11] D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1999). Linear modeling of mRNA expression levels durning CNS development and injury. *Pacific Symposium on Biocomputing*. 41-52.

[12] Weaver, D., Workman, C. and Stormo, G. (1999). Modeling Regulatory Networks with Weight Matrices. *Pacific Symposium on Biocomputing*, 112-123.

[13] Mjolsness, E., Mann, T., Castano, R. and Wold, B. (2000). From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data. *Neural Information Processing Systems*, **12**, 928-934.

[14] D'Haeseleer, P. Liang S. and Somogyi, R. (2000). Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707-726.

[15] van Someren, E., Wessels, L.F.A. and Reinders, M.J.T. (2001). Genetic Network Models: A Comparative Study. Proceedings of SPIE, Micro-arrays: Optical Technologies and Informatics (BIOS01), 4266, 236-247.

[16] Wahde. M. and Hertz, J. (2001). Modeling Genetic Regulatory Dynamics in Neural Development. *J. Comput. Biol*., **8**, 429-44.

{17] de Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2003). Inferring Gene Regulatory Networks From Time-Ordered Gene Expression Data of Bacillus Subtilis Using Differential Equations. *Proceedings of the Pacific Symposium on Biocomputing*. 17-28.

[18] Wimberly, F., Heiman, T., Ramsey, J. and Glymour, C. (2003). Experiments on the Accuracy of Algorithms for Inferring the Structure of Genetic Regulatory Networks from Microarray Expression Levels. IJCAI-2003 Workshop on Learning Graphical Models for Computational Genomics.

[19]Saavedra, R., Spirtes, P., Ramsey, R. and Glymour, C. (2001). Issues in learning gene regulation from microarray databases. Technical Report IHMC-TR-030101-01, Institute for Human and Machine Cognition.

[20] Scheines, R., Spirtes, P., Glymour, C., Meek, C. and Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioural Research*, **33**, 65-118.

[21] Bay, S.D., Shrager, J., Pohorille, A. and Langley P. (2002). Revising regulatory networks: From expression data to linear causal models, *J. Biomed. Informatics*, **35**, 289-297.

[22] Zupan, B., Bratko, I., Demsar, J., Juvan, P., Curk, T., Borstnik, U., Beck, J.R., Halter, J., Kuspa, A. and Shaulsky, G. (2003). GenePath: a system for inference of genetic networks and proposal of genetic experiments. *A.I. in Medicine* **29**, 107-130.

[23] Holmes, I. and Bruno, W.J. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. 8th Int. Conference on Intelligent Systems for Molecular Biology.*

[24] Segal, E., Yelensky, R. abd Koller, D. (2000). Genome-wide Discovery of Transcriptional Modules from {DNA} Sequence and Gene Expression. *Bioinformatics*, **19**, 1273-1282.

[25] Bhaya, D, Dufresne, A, Vaulot, D, Grossman, A: Analysis of the hli gene family in marine and freshwater cyanobacteria. FEMS Microbiology Letters, 2002, 215.

[26] Hihara, Y, Kamei, A, Kanehisa, M, Kaplan, A, Ikeuchi, M: DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. Plant Cell, 2001, 13(4):793-806.