Title:
Elucidating Genetic Regulatory Networks Using Graphical Models and Genomic Expression Data

Authors:
Alexander J. Hartemink
MIT Laboratory for Computer Science
200 Technology Square
Cambridge, MA 02139
USA
amink@alum.mit.edu

David K. Gifford
MIT Laboratory for Computer Science
200 Technology Square
Cambridge, MA 02139
USA
gifford@lcs.mit.edu

Tommi S. Jaakkola
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
USA
tommi@ai.mit.edu

Richard A. Young
Whitehead Institute for Biomedical Research
Nine Cambridge Center
Cambridge, MA 02142
USA
young@wi.mit.edu

Abstract:
We demonstrate how graphical models, and Bayesian networks in particular, can be used to model genetic regulatory networks. These models can be scored in a principled manner in the presence of genomic expression data. These methods are well-suited to this problem owing to their ability to model more than pair-wise relationships between variables, their ability to guard against over-fitting, and their robustness in the face of noisy data. We develop methods for extending the semantics of these Bayesian networks to include edge annotations that allow us to model statistical dependencies between biological factors with greater refinement. We derive principled methods for scoring these annotated Bayesian networks. We apply our scoring framework to validate models of regulatory networks in comparison with one another. To demonstrate the utility of this framework for the elucidation of genetic regulatory networks, we apply these methods in the context of the *Saccharomyces cerevisiae* galactose regulatory system.

Keywords:
genomics, functional genomics, computational functional genomics, genetic regulatory network, genomic expression data, expression array, Affymetrix GeneChip, graphical model, Bayesian network, annotated graphical model, annotated Bayesian network, annotation, validation

Journal Name:
IEEE Intelligent Systems, Special Issue on Intelligent Systems in Biology

ELUCIDATING GENETIC REGULATORY NETWORKS USING GRAPHICAL MODELS AND GENOMIC EXPRESSION DATA

ALEXANDER J. HARTEMINK
MIT Laboratory for Computer Science
200 Technology Square, Cambridge, MA 02139

DAVID K. GIFFORD, TOMMI S. JAAKKOLA
MIT Artificial Intelligence Laboratory
200 Technology Square, Cambridge, MA 02139

RICHARD A. YOUNG
Whitehead Institute for Biomedical Research
Nine Cambridge Center, Cambridge, MA 02142

1 Introduction

Genes are blueprints for proteins, the molecular workhorses with roles in cellular structure, motility, metabolism, homeostasis, signaling, signal transduction, reproduction, and repair. One of the most intriguing roles for proteins, however, is that of genetic regulation: control of precisely which genes are translated into proteins at any given time in the cell. Through these genetic regulatory mechanisms, proteins are responsible for controlling their own existence, and yet very little is known about the sets of signals and controls that activate and repress the expression of specific genes. In this paper, we present a principled, hypothesis-driven method for elucidating these genetic regulatory networks using graphical models and genomic expression data. We discuss the suitability of this approach, as well as its limitations, and demonstrate its application in the context of the galactose system in yeast.

2 Background

While the reductionist approach to biology has proven immensely effective over the course of the last century, and the latter half of the last century in particular, our efforts are increasingly focusing on more integrated approaches to understanding complex biological systems. The success of high-throughput genome sequencing efforts (most notably the Human Genome Project) and an exponentially-expanding quantity of genomic expression data present a

significant opportunity to use integrated computational methods to transform our understanding of the cellular processes governing life. Our ability to observe and measure the responses of different cells to diverse treatments will have a profound impact on the understanding of cell biology, the diagnosis and treatment of disease, and the efficacy of designing and delivering targeted therapeutics. The long term promise is that some day we might be able to find a cure for diseases like Alzheimer's, cystic fibrosis, or cancer.

2.1 Gathering genomic expression data

A number of technologies exist for gathering data characterizing the levels of gene expression of cells on a genome-wide scale. The most prominent such methods are array-based, but other methods such as SAGE (Serial Analysis of Gene Expression) and RT-PCR (Reverse Transcriptase Polymerase Chain Reaction) are used in certain contexts. In this paper, we consider genomic expression data gathered using Affymetrix GeneChips, which are high-density oligonucleotide arrays printed using a lithographic masking process. GeneChip arrays consist of tens of thousands of *features*, each containing a unique set of short DNA strands that act as probes for binding specific target nucleic acid molecules. To quantify the genome-wide levels of gene expression for a population of cells, mRNA transcripts[a] are extracted from the cells, labeled with fluorescent tags, and hybridized to arrays containing features designed to collectively probe for all the various genes in the genome.

2.2 Data-driven analysis of genomic expression data

The first forays into analysis of genomic expression data can be characterized as primarily data-driven, in the sense that they have focused on the discovery of patterns within the observed data itself. Within this analysis paradigm, data gathered from expression arrays is first preprocessed to make it comparable with other such data, and then the resultant data matrix (genes $\times$ experiments) is mined for interesting patterns. Common methodologies for data analysis include smoothing data, extracting trends from data, correlating data vectors, clustering data, ordering clustered data, labeling clustered data, categorizing data, and representing suitably analyzed data in suggestive visual forms. Extensions to this basic idea include identifying common *cis*-acting sequence motifs within clusters and correlating lagged data vectors from time-series data.

[a] mRNA (messenger RNA) transcript molecules serve as physical intermediaries between genes and proteins.

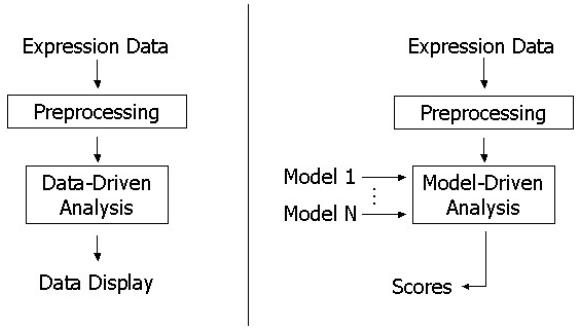

Figure 1. *Comparison of two different paradigms for the analysis of genomic expression data. The essential difference is that in the prevailing data-driven analysis paradigm depicted on the left, the results are the data (clustered, ordered, summarized, and visualized for the user in suggestive ways). In the proposed model-driven analysis paradigm depicted on the right, the results are numeric scores that provide a direct measure for comparing the posterior likelihood of the models in the presence of the observed data.*

This paradigm has proven quite successful in identifying a number of striking patterns within gene expression data. For example, various genes of similar function often cluster together, especially when the topological clusters are optimally ordered,[1] certain genes from cell-cycle synchronized cells are noted to behave cyclically with periodicity related to the underlying period of the cell-cycle, and various genes have been identified that seem to offer predictive power in terms of categorizing types of cancers, and even subtypes of cancers based not only on morphology but on other phenotypic variables like mortality or response to treatment.[2]

Unfortunately, these data-driven techniques for analyzing genomic expression data generally do not permit the rigorous statistical testing of hypotheses about the structure of the complex regulatory networks responsible for transcriptional control. Moreover, although we know that cells regulate transcription through combinatorial multi-variate control processes, most of these methods rely on pair-wise measures such as correlation, Euclidean distance, or (pair-wise) mutual information to calculate gene expression similarity.

2.3 Model-driven analysis of genomic expression data

These methodologies have been useful in uncovering interesting patterns or regularities in the expression data that need to be explained. To explain these patterns, we would like to be able to postulate models describing the

underlying biological mechanisms that give rise to them and then score these models in order to determine which are most consistent with the observed data. We therefore propose a model-driven framework for the analysis of gene expression data in which we represent hypotheses about the function of underlying genetic regulatory networks in a compact probabilistic form and develop principled methods for scoring these hypotheses in comparison with one another in terms of their relative ability to explain noisy expression data.

There are a number of possible modeling frameworks that we might consider. At one end of the spectrum are highly specified models such as those based on differential equations or stochastic Petri nets. These kinds of models seek to explain observed expression levels by capturing the very small-scale molecular dynamics of fundamental reactions taking place within the cell, in some cases simulating not only the temporal evolution of molecular concentrations and reactions, but also the spatial aspect of these phenomena as well. The difficulty with using such a model for this task is that it is usually so highly specified that it requires not only an exact knowledge of which factors interact with which other ones, which is precisely what we do not know, but also the reaction rates associated with such interactions, in terms of binding affinities, free energies, dissociation rates, equilibrium constants, and the like. While models at this level of specification represent the Holy Grail of our ability to understand what is happening in the cell, with the exception of certain special cases, they are unattainable at this juncture because so little is currently known about which factors in the cell interact with which other factors, let alone the frequencies and rates at which such interactions occur. In contrast, we need models that are more abstracted than these, capable of capturing the kernel phenomena without requiring a burdensome level of specification.

At the other end of the modeling spectrum are highly abstracted models. If the semantics associated with these models are so highly abstracted that they lose the ability to represent core regulatory phenomena, however, then we can make similarly little progress. One example of such an overly abstracted model might be a Boolean network model. In a Boolean network model, all factors in the genetic regulatory network are represented by Boolean variables, which can only take on two possible values. Moreover, in a Boolean network all relationships between variables are required to be logical, which allows little room for explaining levels of gene expression that have become corrupted by noise during the measurement process or are not the result of clean and logical regulatory processes but rather ones that are inherently stochastic.

Between these two extreme ends of the modeling spectrum lie a family of models known as *graphical models*, a family of flexible and interpretable

models for compactly representing probabilistic relationships among variables of interest in the form of a graph. While this family of models is fairly large and includes a number of possibly relevant classes of models, we concentrate here on a particular class of models known as *Bayesian networks*. In our modeling framework, Bayesian networks are used to describe relationships between variables in a genetic regulatory network.

Bayesian networks can describe arbitrary combinatorial control of gene expression and thus are not limited to pair-wise or linear interactions between genes. Due to their probabilistic nature, Bayesian networks are robust in the face of both noisy expression data and imperfectly specified hypotheses about the function of genetic regulatory networks. Moreover, Bayesian networks cleanly handle missing data and permit latent variables to represent unobserved factors, and when we extend the semantics of Bayesian networks to allow edge annotations (as described in Section 6), Bayesian networks can specify relationships between variables at varying levels of refinement. Most importantly, models of genetic regulatory networks that are based on Bayesian networks are biologically interpretable and can be scored rigorously against observed genomic expression data.

In contrast to models employing differential equations to simulate the molecular dynamics of interactions between factors in the cell, determining the precise dynamics of genetic regulation is outside the scope of the Bayesian network techniques we present here. Rather, we seek to develop comprehensive high-level models that are able to suggest which factors in the cell are interacting with which others. This information could be used as the basis for constructing more highly specified, low-level models based on differential equations in the future.

We should mention that our work on Bayesian networks for modeling genetic regulatory networks was developed concurrently with similar work by Friedman, *et al.*[3] and Murphy and Mian.[4] While their research concentrates on different aspects of this domain, all three bodies of work taken together represent a fairly comprehensive treatment of this topic in published literature to date.

3 Using Bayesian networks to model genetic regulatory networks

Variables in a Bayesian network can be either discrete or continuous, and can represent mRNA concentrations, protein concentrations, protein modifications or complexes, metabolites or other small molecules, experimental conditions, genotypic information, or conclusions such as diagnosis or prog-

nosis. A variable that describes an observed value is called an *information variable*, while a variable that describes an unobserved value is called a *latent variable*.

A Bayesian network describes the relationships between variables at both a qualitative and a quantitative level. At a qualitative level, the relationships between variables are simply dependence and conditional independence. These relationships are encoded in the structure of a directed graph, S, to achieve a compact and interpretable representation. Vertices of the graph correspond to variables, and directed edges between vertices represent dependencies between variables. The fewer edges a model has, the more constrained is the model since it makes more independence assertions. In practice, we seek sparse models because they are able to explain away certain "indirect" dependencies through more "direct" dependencies mediated by other variables. Formally, if vertices X and Y are d-separated by a set of vertices Z, then X and Y are conditionally independent given Z. In particular, if there is a directed edge from X to Y, then Y is dependent on X. Since Y can have multiple incoming directed edges, it can depend combinatorially on multiple variables. We call variables that have a directed edge to Y the *parents* of Y, denoted $\mathrm{Pa}(Y)$.

At a quantitative level, relationships between variables are described by a family of joint probability distributions that are consistent with the independence assertions embedded in the graph. Each member of this family is described by the vector, $\boldsymbol{\theta}$, of parameters that characterize it. As this method is Bayesian in nature, we do not consider only a single value for $\boldsymbol{\theta}$, but rather a distribution over all possible values of $\boldsymbol{\theta}$ that are consistent with the structure of the graph, S. This distribution over distributions enables these models to avoid over-fitting, a common problem when parameters are restricted to a single value in the context of small quantities of data.

In a Bayesian network, each joint probability distribution over the space of variables can be factored into a product over the variables, where each term is simply the probability distribution for that variable conditioned on the set of parent variables:

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathrm{P}(X_i \mid \mathrm{Pa}(X_i)) \tag{1}$$

This follows from the conditional Markov assumption which states that each variable is independent of its non-descendants when conditioned on its parents. The parameters that characterize the conditional probability distributions on the right hand side of Equation 1 therefore comprise the parameter vector, $\boldsymbol{\theta}$.

Although we only discuss static models of regulatory networks in this

paper, Bayesian networks can also be used to model dynamic processes such as feedback. This is accomplished by "unrolling" a static model, creating a series of connected models that contain dependencies spanning across time steps.[5] In a modeling context, dynamic Bayesian networks smoothly interpolate between static graphical models and differential equation models.

While continuous variables are permitted, for the remainder of this paper we consider only discrete variables to simplify the exposition. Each variable is thus in one of a set of states, and the number of states used to model a variable represents a trade-off between precision, the ability to intuit what the state of the variable means, and the computational complexity of evaluating a model with a given number of states.

4 Scoring network models with the Bayesian scoring metric

When scoring Bayesian networks against observational data, we employ the Bayesian scoring metric, a principled statistical scoring metric that allows us to directly compare the merits of alternative models of genetic regulatory networks.[b] The model scores produced by the Bayesian scoring metric permit us to rank alternative models based on their ability to explain observed data economically. Moreover, the difference between the scores for any two models leads to a direct significance measure for determining how strongly one should be preferred over the other.

According to the Bayesian scoring metric, the score of a model is defined as the logarithm of the probability of the model being correct given the observed data. Formally,

$$\begin{aligned} \mathsf{BayesianScore}(S) &= \log \mathrm{p}(S \mid D) && (2) \\ &= \log \mathrm{p}(S) + \log \mathrm{p}(D \mid S) + c && (3) \end{aligned}$$

[b] Due to space limitations, we present here only the basic intuition behind the Bayesian scoring metric; more detailed quantitative treatments are available elsewhere.[6] We note that the entire discussion is equally valid in the case of dynamic Bayesian networks.

where the first term is the log *prior* distribution of S, the second term is the log *likelihood* of the observed data D given S, and c is a constant that does not depend on S. The likelihood term can be expanded as follows:

$$\mathrm{p}(D \mid S) = \int_{\boldsymbol{\theta}} \cdots \int \rho(D, \boldsymbol{\theta} \mid S)\, \mathrm{d}\boldsymbol{\theta} \tag{4}$$

$$= \int_{\boldsymbol{\theta}} \cdots \int \mathrm{p}(D \mid \boldsymbol{\theta}, S)\rho(\boldsymbol{\theta} \mid S)\, \mathrm{d}\boldsymbol{\theta} \tag{5}$$

From this last expression, we see that the likelihood component of a model's score can be viewed as the average probability of generating the observed data over all possible values of the parameter vector, $\boldsymbol{\theta}$.

Because the Bayesian scoring metric includes an average over a family of probability distributions, it is well suited to our context for a number of reasons. First, it includes an inherent penalty for model complexity, thereby balancing a model's ability to explain observed data with its ability to do so economically. Consequently, it guards against over-fitting models to data. Second, regulatory network models are permitted to be incomplete. An incomplete model contains additional degrees of freedom pertaining to the possible ways of completing the model, and is thus penalized by the scoring metric for these additional degrees of freedom. Scores improve as a model converges to properly depict underlying regulatory mechanisms without extraneous degrees of freedom, thereby allowing network elucidation to proceed incrementally. Third, it allows us to represent uncertainty about the precise dependencies between variables since we need not select a single value for $\boldsymbol{\theta}$, but rather can permit all feasible values to exist in the distribution over $\boldsymbol{\theta}$.

We have mentioned the ability to represent as Bayesian networks models that contain variables that are unobserved or for whom data is occasionally missing. The difficulty with these latent variable models is that the integrals computed as part of the Bayesian scoring metric can no longer be solved exactly once we are faced with incomplete data. One way to score models with latent variables is to instantiate the latent variables by sampling from the distribution of possible values for each such variable (*e.g.*, MCMC methods). Though this is feasible for small networks, it becomes computationally prohibitive as networks become very large. In such settings, variational approximation methods[7] can be used, either on their own or in conjunction with sampling. In addition, variational methods can also yield upper and lower bounds on the score, often enabling the highest scoring graph to be identified without resorting to sampling. For reasons of computational simplicity, we consider in this paper only models with variables for which we have complete

data. The extension to the context of incomplete data, while computationally burdensome, is fairly straightforward.

4.1 Prior establishment

In a Bayesian setting, we need to establish prior distributions both over the set of parameter vectors, $\boldsymbol{\theta}$, that describe the joint probability distribution, and over the set of model structures, S. In a discrete Bayesian network satisfying the reasonable assumptions of parameter modularity, parameter independence, and likelihood equivalence, Heckerman, *et al.*[6] have shown that the parameters of a discrete Bayesian network are distributed according to a Dirichlet distribution. If there is prior information about parameters, this information can be captured in the form of an equivalent prior network with Dirichlet distributed parameters.[6] However, if there is no prior information about parameters, an uninformative prior is frequently employed. In both cases, an *equivalent sample size* needs to be specified. This value is a measure of how confident we are in the prior relative to the quantity of data.

With respect to the prior distribution over graph structures, this is usually done uniformly over structures for computational convenience, though other alternatives can be considered[c].

5 Example: scoring models of the galactose system

As a demonstration of the utility of Bayesian networks for modeling genetic regulatory networks, we analyze and score models of the regulatory network responsible for the control of genes necessary for galactose metabolism in *S. cerevisiae*. As this is a fairly well-understood model system in yeast, it affords us the opportunity to evaluate our methodology in a setting where we can rely on accepted fact. We are also utilizing our Bayesian network methodology to explore other systems that are less well-understood like pheromone response and cell-cycle control in yeast, but do not present those results here.

5.1 Data preparation

A set of 52 samples of unsynchronized *Saccharomyces cerevisiae* populations were observed under a diversity of experimental conditions. The set of sam-

[c]For example, we might consider simple nonuniform priors over structures based on either the number of edges present or their degree of divergence from some pre-specified prior structure. Nonuniform priors over structures can also arise from choices based on computational convenience if we are examining the space of model equivalence classes (PDAGs) or vertex orderings rather than the space of model structures.

ples ranges widely but consists primarily of observations of various wild-type and mutant *S. cerevisiae* strains made under a variety of environmental conditions including exposure to different nutritive media as well as exposure to stresses like oxidative species, excessive acidity, and excessive alkalinity. Whole-genome expression data for each of these 52 observations was collected using Affymetrix Ye6100 GeneChips. These GeneChips are manufactured using a 50-micron process and require four chips to measure the expression of all 6135 genes in the *S. cerevisiae* genome.

5.2 Data normalization and discretization

The reported *average difference* values from these 208 Affymetrix GeneChips were normalized using MAP spiked-control normalization methods.[8] The output of this process was a 6135×52 matrix of normalized log expression values, one row for each gene in the yeast genome and one column for each experimental observation. From this matrix, we extracted rows for each of the genes of interest (described below) and performed binary discretization independently for each gene using a maximum-likelihood separation technique. Other sensible discretization methods could also have been employed; for the particular data set and models in our example, results do not depend on the discretization method and are robust among various different sensible methods. In general, however, the discretization method employed will affect reported scores, and we continue to develop discretization methods that are well suited for expression array data.[d]

5.3 Model preparation

Examples of genetic regulatory networks represented as Bayesian networks are shown in Figure 2. Boxed variables suffixed with "m" describe mRNA levels that can be determined from expression array data. Unboxed variables suffixed with "p" describe protein levels; in this model they would be latent variables whose values cannot be measured directly. The two networks in the figure represent two competing models of a portion of the galactose system in yeast, and differ in terms of the dependence relationships they assert hold between the variables Gal80p, Gal4m, and Gal4p. To quote from Johnston, "it was originally proposed that Gal80 protein is a repressor of *GAL4* transcription. It is now clear that *GAL4* is expressed constitutively and that its

[d] Bayesian networks are capable of modeling continuous variables using parametric or semiparametric density estimation, but discretization is more robust in a setting such as this one where only a small number of observations is available.

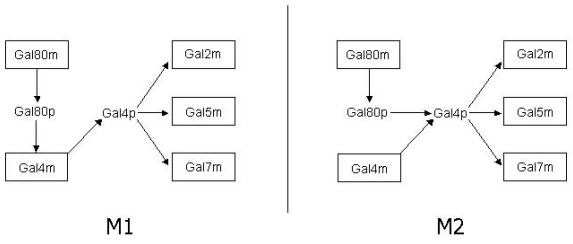

Figure 2. *Representative Bayesian networks for describing a portion of the galactose system in yeast. The model M1 on the left represents the claim that Gal80p represses the transcription of Gal4m, while the model M2 on the right represents the claim that Gal80p inhibits Gal4p activity posttranslationally.*

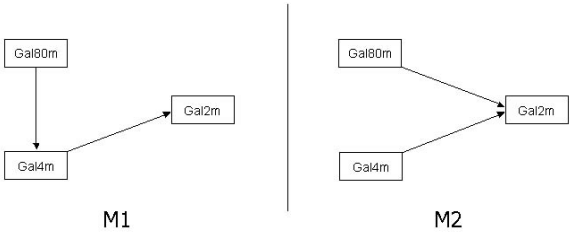

Figure 3. *Simplified Bayesian networks for describing a portion of the galactose system in yeast. These simplified versions of M1 and M2 capture the kernel of the conditional independence assertions of the more complex models of Figure 2. As above, in M1, Gal2m is independent of Gal80m when conditioned on Gal4m, and in M2, Gal4m is marginally independent of Gal80m.*

activity is inhibited by Gal80 protein posttranslationally." [9] The network on the left (M1) represents the original proposition, while the network on the right (M2) represents the new assertion. The models in Figure 3 represent the same conditional independence assertions of the models in Figure 2, but are simplified to reveal the kernel of the distinction between the two hypotheses in terms of the effects on the observed transcript levels, namely that in M1, Gal2m is independent of Gal80m when conditioned on Gal4m, while in M2, Gal4m is marginally independent of Gal80m.

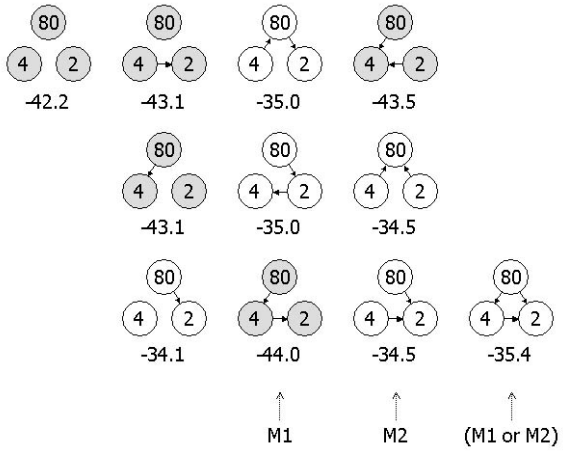

Figure 4. *Scores for all model equivalence classes of the three variable galactose system. The classes of models that score poorly are shown shaded. The previously considered models M1, M2, and (M1 or M2) are indicated.*

5.4 Model validation and comparison

Using the Bayesian scoring metric, we are able to compare the two models shown in Figure 3 in terms of their relative likelihood of explaining the observed (now discretized) data. The model M1 received a score of -44.0, while the model M2 received a score of -34.5. This score difference translates to the data being over 13,000 times more likely to be observed under M2, the currently accepted model. For extra measure, we also scored a more complex model (M1 or M2) that would admit either of the two models as special cases. The data do not persuade us to accept such a model since the score (-35.4) is lower than that of the currently accepted model.

We then broadened our scope to consider not only these three models, but all possible models among these three variables.[e] Results of this analysis are shown in Figure 4. As is evident from the figure, the models fall into two primary groupings based on their score: those scoring between -34.1 and -35.4 (unshaded) which all include an edge between Gal80m and Gal2m, and those scoring between -42.2 and -44.0 (shaded) which all do not include an edge

[e]Note that some model possibilities are equivalent to others in that they describe the same set of conditional independencies; more accurately then, we consider all possible model equivalence classes.

between Gal80m and Gal2m. This lends support to the claim that Gal80m and Gal2m are very unlikely to be conditionally independent given Gal4m, again consistent with the currently accepted hypothesis.

It is interesting to note that the best scoring model in Figure 4 actually has no edge from Gal4m to Gal2m, indicating that there is little evidence in the data set for requiring this edge to be present. This is consistent with the fact that under normal conditions, Gal4m is constitutively expressed and its influence on Gal2m is usually regulated by the action of Gal80 protein, as hypothesis M2 indicates. If the data set instead contained experiments with *GAL4* deletion mutants in which the absence of Gal4m resulted in a loss of Gal2m expression, there would be strong support for the inclusion of this edge. We discuss these subtleties at the end of this paper.

6 Representing and scoring network models with annotated edges

Biologists use many specialized terms to describe the actions and interactions of factors within the cell. Each of these terms implies a specific kind of relationship between the factors involved in the interaction, a relationship that is specified at a finer degree of granularity than the generic statement about conditional dependence that is implied by the existence of an edge connecting the corresponding vertices in a Bayesian network. How can we leverage refined knowledge about the form of the relationship between factors in the cell, and how can we discover such refined knowledge from data?

We propose a general method for providing increased refinement of knowledge in graphical models by introducing a *constraint* framework. In this framework, knowledge about the structure of the relationship between two variables is represented in the form of a constraint that the relationship must satisfy. The data is not forced to obey these constraints (after all, the data is noisy) but the parameters that characterize the distributions used to model the data are forced to obey these constraints. Before describing the exact method for exploiting such constraints, we first discuss how the presence of constraints can be beneficial in scoring and discovering graphical models in the presence of observed data.

As mentioned in Section 4, the likelihood component of a model's score can be viewed as the average probability of generating the observed data over all possible values of the parameter vector, $\boldsymbol{\theta}$. From a sampling perspective, the contribution of the likelihood term to the score can be viewed as a two-level data generation process whereby a realization of the parameter vector, $\boldsymbol{\theta}$, is selected at random from its prior distribution, and then the probability

of generating the observed data is calculated using this realization of $\boldsymbol{\theta}$. The probability of generating the data is then averaged over repeated samplings. This interpretation reveals that a model will score poorly if there is not a sufficiently large mass of realizations in the complete distribution of $\boldsymbol{\theta}$ that are capable of generating the data with sufficiently high probability.

On the other hand, if the model is constrained to the extent that the distribution of $\boldsymbol{\theta}$ has a great deal of its mass concentrated on realizations that are capable of generating the data with sufficiently high probability, then the constrained model will score better under the Bayesian scoring metric. Note that one constraint on the type of relationship between variables is conditional independence (edge absence), which is merely a special case in this framework. Whether the relationship is constrained as independence, or in some other fashion, if the constraint permits the model to avoid unneeded complexity, then the model's score will increase under the Bayesian scoring metric.

6.1 Annotation semantics

We now extend Bayesian network models by adding the ability to annotate edges, permitting us to represent additional information about the type of dependence relationship between variables. Although many such annotations are possible, because monotonic relationships are especially useful in a biological setting and the most straightforward to characterize semantically, we consider here only the following four types of edges:

- An *unannotated* edge from X to Y represents a dependence that is unconstrained (the default case). In the presence of unannotated edges from all parents of Y, we can represent arbitrary combinatorial control of Y.
- A *positive ("+")* edge from X to Y indicates that higher values of X are constrained to bias the distribution of Y higher. This monotonic influence of X on Y holds for all possible values of the other parents of Y, though the strength of the influence can vary with the setting of the other parents. Formally, for all values y of Y, for all values $x_i < x_j$ of X, and for all instantiations $\mathcal{I}$ of the variables in $\mathrm{Pa}(Y)/X$, we require $\mathrm{P}(Y > y \mid X = x_i, \mathcal{I}) \leq \mathrm{P}(Y > y \mid X = x_j, \mathcal{I})$.
- A *negative ("−")* edge from X to Y indicates that higher values of X are constrained to bias the distribution of Y lower. This monotonic influence of X on Y holds for all possible values of the other parents of Y, again with possibly varying strength. Formally, for all values y of Y, for all values $x_i < x_j$ of X, and for all instantiations $\mathcal{I}$ of the variables in $\mathrm{Pa}(Y)/X$, we require $\mathrm{P}(Y > y \mid X = x_i, \mathcal{I}) \geq \mathrm{P}(Y > y \mid X = x_j, \mathcal{I})$.

- A *positive/negative ("+/−")* edge from X to Y indicates that Y's dependence on X is either positive or negative but the true relationship is not known. This monotonic influence of X on Y holds for all possible values of the other parents of Y, again with possibly varying strength.

Because edge annotations describe the relationship between a variable and a single parent while Bayesian networks describe the relationship between a variable and all its parents, we have chosen to specify the semantics of annotations by requiring that the implied constraints hold for all possible values of the other parents.

A given Bayesian network can have any combination of edge annotations. This enables us to represent finer degrees of refinement regarding the types of relationships between variables when we desire, but does not force us to do so as unannotated edges are always permitted. It also allows a model to evolve as more knowledge is gained about the types of influences that are present in the biological system under study. For example, all edges can be initially unannotated, with +/− and then + and − annotations being added incrementally as activators and repressors are later identified.

The implied constraints on the form of the dependence between variables permit us to score annotated models much as we score unannotated models. We simply modify the scoring metric so that the likelihood term is now the average probability of generating the observed data over all possible values of the parameter vector $\boldsymbol{\theta}$ that satisfy the constraints implied by the annotations.

7 Example: scoring annotated models of the galactose system

When we expand the semantics of Bayesian networks to include annotated edges, we are able to score models that describe more fine-grained relationships between variables. For example, when we consider again the two models M1 and M2, and allow the edges in each model to take on all possible combinations of annotations (−, +/−, or +), we are able to score the models as shown in Table 1. In model M1, adding different kinds of annotations fails to change the score significantly, as the structure of the graph is fundamentally limited in explaining the observed expression data. The same effect is observed when the edge between Gal4m and Gal2m is considered in model M2, which is consistent with the results of Figure 4 indicating that the coupling between Gal4m and Gal2m is indeed quite weak. In contrast, adding a + annotation to the edge between Gal80m and Gal2m results in a score comparable with previously achieved scores, but adding a − annotation to the same edge worsens the score dramatically. Such an asymmetric response is

Table 1. *Scores for models M1 and M2 under all possible configurations of annotated edges.*

		Gal4m ⟶ Gal2m		
		−	+/−	+
Gal80m	−	-45.3	-44.6	-44.2
↓	+/−	-44.6	-43.8	-43.4
Gal4m	+	-44.2	-43.4	-43.0

M1

		Gal4m ⟶ Gal2m		
		−	+/−	+
Gal80m	−	-48.9	-47.3	-46.7
↓	+/−	-35.5	-35.4	-35.4
Gal2m	+	-34.8	-34.8	-34.7

M2

to be expected as failure to explain the observed data is more revealing than success. This example illustrates that when the constraints implied by edge annotations cannot be satisfied by the data, scores result that are as poor as when the underlying structure is incorrect. For this reason, annotations serve as a useful discriminator of the kinds of relationships present in the data.

The lowest score (-33.6) is achieved by model M2 when the edge from Gal4m to Gal2m is unannotated and the edge from Gal80m to Gal2m is labeled +. Although Gal80 is known to act in a repressive role in the cell, its level increases as galactose becomes available for metabolism. This increase, however, is more than offset by a rise in the level of a factor that counteracts the effect of Gal80. The identity of this factor is currently unknown and thus remains unmodeled here, but it is believed to be a byproduct of the metabolism of galactose.[9] A complete model would include the effect of this latent (unmeasured) variable, and in such a model, it would be expected that with sufficient data, the edge between Gal80 and Gal2 would be labeled −, corresponding to the direct repressive role of Gal80. Nevertheless, in the limited model considered here, a + annotation for the edge is indeed correct as the level of Gal80 rises concomitantly with the level of Gal2 in our experimental data.

8 Discussion

There are certain limitations when using Bayesian networks for modeling genetic regulatory networks. The most important of these is the caution with which models must be interpreted. While graphs are highly interpretable structures for representing statistical dependencies, they have the potential to be misleading if interpreted incorrectly. It is important to distinguish between statistical interaction and physical interaction.

For example, if the data strongly supports the inclusion of an edge be-

tween two variables X and Y, that may indicate a physical interaction between these two factors in the cell. Alternatively, it is possible that an unmodeled variable Z actually intermediates between X and Y, such that X and Y exhibit statistical dependence but no physical interaction. As in the example in Section 7, caution must be used when interpreting models that may be missing critical explanatory variables. In contrast, if the data strongly supports the exclusion of an edge between two variables X and Y, that may indicate there is no physical interaction between these two factors in the cell. Alternatively, we may not have observed the cell under an appropriate set of conditions where this interaction could have been observed. This was the case in Section 5 when there was not strong support for including an edge between Gal4m and Gal2m, though the two factors are known to interact in the cell.

In general, multiple biological mechanisms may map to the same set of statistical dependencies and thus be hard to distinguish on the basis of statistical tests alone. Moreover, if there is not sufficient data to observe a system in a number of different configurations, we may not be able to uncover certain dependencies. These two limitations are not specific to this methodology, however, but rather are true for scientific inquiry in general.

As for the cost associated with scoring large models, it should be noted that this cost is to a large extent based on the in-degree (number of parents) of the variables in the models. As we scale up to larger models, the in-degree is likely to remain fairly small whereas the out-degree might be very large, which is fine for our Bayesian network approach.

One limitation of comparing regulatory network models is that human effort is needed to formulate the models being compared. However, with a principled scoring metric, automatic *model induction* becomes possible. Although in this paper we only present examples in the context of model validation and comparison, we have also successfully used our modeling framework in the context of model induction (work still in progress).

9 Directions for future work

9.1 Beyond genomic expression data: information fusion

In this paper, we have discussed how genomic expression data can be used to elucidate genetic regulatory networks but there are many other sources of data that we can exploit for this task. We distinguish between two classes of additional data. The first consists of data that can be measured simultaneously with gene expression. Examples include protein expression, protein modification, levels of metabolites or other small molecules, or even cell mor-

phology. As long as these are observed in tandem with the levels of gene expression, they can be modeled simply by adding additional variables to the graph.

The second class consists of data that cannot be gathered at the same time as the levels of gene expression are measured. Examples include learning from a two-hybrid screen that two proteins interact, learning from location analysis that a transcription factor binds to the upstream sequences of certain genes, or learning from sequence analysis that two genes share a common promoter motif. In theory, the Bayesian methodology provides principled ways for incorporating this additional information as it has a natural provision for incorporating prior information into its scoring metric; in practice, giving appropriate weight to each of these sources of information poses a significant challenge.

9.2 Guided discovery of network models

Rather than developing a single monolithic algorithm that unearths major biological insights automatically from large mounds of data, we can consider algorithms that work to develop these insights by combining the user's deep intuition about the operation of biological systems with the computational learning that is possible with vast quantities of data. Thus, an important goal will be to develop algorithms and tools that are capable of augmenting the user's intuition. Bayesian frameworks are ideally suited to combining the prior information of users with the information embedded in repeated observation of the system in question, but we will likely need to consider tools that do not gather all the prior information in advance and then discard the user, but rather tools that extract helpful information from the user as it is needed in the learning process. This process will likely be interactive and online, not limited simply to batch learning from data. Ideally, these tools will be able to suggest new experiments to be conducted and the interactive process will proceed as new data continue to be generated.

9.3 Experimental suggestion

The necessity of observing a system in a number of configurations in order to best elucidate its structure suggests the possibility of performing *experimental suggestion* in the future. In such a context, existing models and data could be used to generate suggestions for new experiments, yielding data that would optimally elucidate a given regulatory network.

As discussed above, elucidation of genetic regulatory networks will not simply be a batch learning process. The space of possible models to consider

is so large that we cannot even begin to imagine gathering sufficient data to allow an algorithm to simply churn away and produce a correct model without any intervention. Rather, we will need to consider learning that is incremental and learning algorithms that are online.

In particular, rather than gathering data sampled from the joint probability space over all relevant variables in cellular regulatory networks, it will be important to carefully design experiments to learn information about the specific portions of these networks that remain ambiguous. Being able to suggest the next series of experiments to conduct is especially valuable in this context of learning from genomic expression data because the data is costly to gather, in terms of both laboratory time and money. It would be quite useful to know in advance which are likely to be the most informative experiments to conduct for elucidating biological mechanisms of interest.

This field is known as active learning and there is an existing literature that can be applied and extended in this domain. Of special interest is the ability to suggest experiments for collecting not only observational data but also interventional data. In the context of genetic regulatory networks, this can be implemented by deleting a gene so that it cannot be expressed or by constitutively over-expressing a gene. Interventional data needs to be treated differently from observational data in the context of learning, but the framework easily extends to handle interventional data.

9.4 Other extensions

Other directions for future work that were mentioned earlier include using graphical models such as dynamic Bayesian networks to model the simple dynamics of genetic regulatory networks, using variational methods to produce efficient tools for scoring regulatory networks with latent variables, increasing the variety of edge annotations permitted in models, and fast algorithms to search for appropriate annotations during the model induction process.

References

1. Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," in *9th International Conference on Intelligent Systems for Molecular Biology (ISMB 2001)*, ISCB, July 2001.
2. A. A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.
3. N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian net-

works to analyze expression data," in *4th Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, ACM-SIGACT, April 2000.
4. K. Murphy and S. Mian, "Modelling gene expression data using dynamic Bayesian networks," tech. report, University of California at Berkeley, 1999.
5. X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 33–42, Morgan Kaufmann Publishers, 1998.
6. D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.
7. T. S. Jaakkola and M. I. Jordan, "Variational probabilistic inference and the QMR-DT database," *Journal of Artificial Intelligence Research*, vol. 10, pp. 291–322, 1999.
8. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Maximum likelihood estimation of optimal scaling factors for expression array normalization," in *International Symposium on Biomedical Optics (BiOS 2001)*, SPIE, January 2001.
9. M. Johnston, "A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae," *Microbiological Reviews*, vol. 51, no. 4, pp. 458–476, 1987.