USING GRAPHICAL MODELS AND GENOMIC EXPRESSION DATA TO STATISTICALLY VALIDATE MODELS OF GENETIC REGULATORY NETWORKS

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

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

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

We propose a model-driven approach for analyzing genomic expression data that permits genetic regulatory networks to be represented in a biologically interpretable computational form. Our models permit latent variables capturing unobserved factors, describe arbitrarily complex (more than pair-wise) relationships at varying levels of refinement, and can be scored rigorously against observational data. The models that we use are based on Bayesian networks and their extensions. As a demonstration of this approach, we utilize 52 genomes worth of Affymetrix GeneChip expression data to correctly differentiate between alternative hypotheses of the galactose regulatory network in *S. cerevisiae*. When we extend the graph semantics to permit annotated edges, we are able to score models describing relationships at a finer degree of specification.

1 Introduction

The vast quantity of data generated by genomic expression arrays affords researchers a significant opportunity to transform biology, medicine, and pharmacology using systematic computational methods. The availability of genomic (and eventually proteomic) expression data promises to have a profound impact on the understanding of basic cellular processes, the diagnosis and treatment of disease, and the efficacy of designing and delivering targeted therapeutics. Particularly relevant to these objectives is the development of a deeper understanding of the various mechanisms by which cells control and regulate the transcription of their genes. In this paper, we present a principled method for using genomic expression data to elucidate these genetic regulatory networks.

While the potential utility of expression data is immense, some obstacles will need to be overcome before significant progress can be realized. First, data from expression arrays is inherently noisy. Despite this, expression analysis results to date have generally been reported without measures of statistical significance. Second, our knowledge regarding genetic regulatory networks is extremely limited. Consequently, hypotheses about their structure or function may be incomplete or include knowledge at varying levels of refinement. Third, gene expression is regulated in a complex and seemingly combinatorial manner.[1] Nevertheless, most analysis of expression array data utilizes only pair-wise measures to compare expression profiles.

Existing techniques for analyzing genomic expression data do not permit the statistical testing of hypotheses about the form or functioning of complex multi-variate regulatory networks responsible for transcriptional control. Typically, analysis is performed by clustering the expression profiles of a collection of genes using pair-wise measures such as correlation,[2,3,4,5] Euclidean distance,[6,7,8] or mutual information.[9,10] Results are visualized graphically and used to demonstrate coordinated patterns of expression.[2,5,7] Extensions to this basic idea include identifying clusters with common *cis*-acting sequence motifs[8,11] and computing regulatory dependencies by correlating lagged time-series data.[12] As noise in expression array data is typically not analyzed in detail, the significance of alternative conclusions from these studies cannot be quantitatively compared. Finally, a single framework currently does not exist that permits models to describe latent variables (such as protein levels) and make predictions that can be verified later as data becomes available.

Previous efforts at modeling genetic regulatory networks have generally fallen into one of two classes, either employing Boolean models,[13,14,15] which are restricted to logical (Boolean) relationships between variables, or using systems of differential equations to model the continuous dynamics of coupled biological reactions.[16,17,18] While low-level dynamics are critical to a complete understanding of regulatory networks, they require a great deal of specification, not only in terms of reaction rates and diffusion constants but also in the precise structure of the relationships between variables.

The work of Friedman, *et al.*,[19] represents a single point of departure by using Bayesian networks to analyze expression data. Their work focuses on the discovery of Bayesian networks, however, and does not consider either the addition of latent variables for capturing the influence of currently unobserved variables or the annotation of edges for representing biological knowledge at different levels of refinement.

We propose a technique for scoring models of genetic regulatory networks based on Bayesian networks and their extensions. In our approach, Bayesian networks are used to describe relationships between variables in a genetic regulatory network. Unlike other approaches such as clustering, a Bayesian network can describe arbitrary combinatorial control of gene expression and thus it is not limited to pair-wise interactions between genes. Due to their probabilistic nature, Bayesian networks are robust in the face of both imperfect data and imperfect models. Moreover, Bayesian networks permit latent variables capturing unobserved factors and allow relationships at varying levels of refinement to be specified. Most importantly, the models are biologically interpretable and can be scored rigorously against observational data.

Sections 2 and 3 of this paper provide an introduction to Bayesian networks and the Bayesian scoring metric, a principled method for scoring these networks. In Section 4, we use this metric to score models of the galactose regulatory network in yeast. In Section 5, we extend Bayesian networks by adding the ability to represent and score models with annotated edges, and in Section 6, we score annotated models of the galactose regulatory network. We discuss limitations and further extensions in Section 7.

2 Modeling regulatory networks with Bayesian networks

Bayesian networks[20,21,22] are a member of the family of graphical models, a class of flexible and interpretable models for representing probabilistic relationships among variables of interest. 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 prognosis. 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. Formally, if 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 of parameters, $\boldsymbol{\theta}$, 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. 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:

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

The parameters that characterize the conditional probability distributions on the right hand side of Equation 1 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.[23,24,25] This is accomplished by "unrolling" a static model, creating a series of connected models that contain dependencies spanning across time steps. In a modeling context, dynamic Bayesian networks smoothly interpolate between static graphical models and differential equation models.

3 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.[a] 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,

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

[a] Due to space limitations, we present here only the basic intuition behind the Bayesian scoring metric; more detailed quantitative treatments are available elsewhere.[26,27] 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:

$$\begin{aligned} \mathrm{p}(D|S) &= \int_{\boldsymbol{\theta}} \cdots \int \rho(D, \boldsymbol{\theta}|S)\, \mathrm{d}\boldsymbol{\theta} && (4) \\ &= \int_{\boldsymbol{\theta}} \cdots \int \mathrm{p}(D|\boldsymbol{\theta}, S)\rho(\boldsymbol{\theta}|S)\, \mathrm{d}\boldsymbol{\theta} && (5) \end{aligned}$$

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 $\boldsymbol{\theta}$.

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[28,29] can be used, either on their own or in conjunction with sampling. Moreover, variational methods also yield upper and lower bounds on the score, enabling the highest scoring graph to often be identified without resorting to sampling.

4 Example: scoring models of the galactose system

As an initial demonstration of the utility of Bayesian networks, we have chosen to analyze and score models of the genetic 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 utilizing our methodology to explore other systems that are less well-understood, but do not present those results here.

Examples of genetic regulatory networks represented as Bayesian networks are shown in Figure 1. Boxed variables describe mRNA levels that can be determined from expression array data. Unboxed variables describe protein levels; in this model we treat them as 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 activity is inhibited by GAL80 protein posttranslationally."[30] 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 2 represent the same conditional independence assertions of the models in Figure 1, but are simplified to reveal the kernel of the distinction between the two hypotheses.

Expression data for this analysis consisted of 52 genomes worth of Affymetrix *S. cerevisiae* GeneChip data. To score these two competing hypotheses' ability to explain the observed data, we used the Bayesian scoring metric, as described in the previous section. We performed binary quantization independently for each gene using a maximum-likelihood separation technique. Other sensible quantization methods could also have been employed; for the particular data set and models in our example, the results do not depend on the quantization method and are robust among various different sensible methods. In general, however, the quantization method employed will affect reported scores, and we are developing quantization methods that are suited for expression array data.[b]

Using the Bayesian scoring metric, we are able to compare the two models shown in Figure 2 in terms of their relative likelihood of explaining the observed (now quantized) data. The model M1, in which Gal80p represses transcription of Gal4m, received a score of -44.0, while the model M2, in which Gal80p inhibits Gal4p activity, 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

[b] Bayesian networks are capable of modeling continuous variables using parametric or semi-parametric density estimation, but quantization is more robust in a setting such as this one where only a small number of datasets is available.

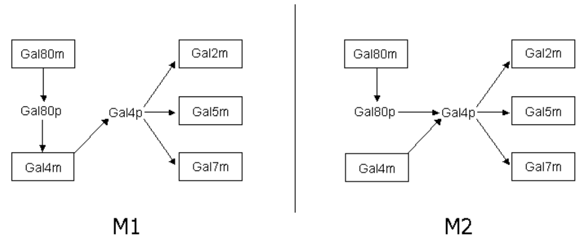

Figure 1. *Representative Bayesian networks for describing a portion of the galactose system in yeast. The model M1 on the left represents the claim that Gal80p represses the transcription of Gal4m, while the model M2 on the right represents the claim that Gal80p inhibits Gal4p activity posttranslationally. In M1, Gal2m is independent of Gal80m when conditioned on Gal4m, and in M2, Gal4m is marginally independent of Gal80m.*

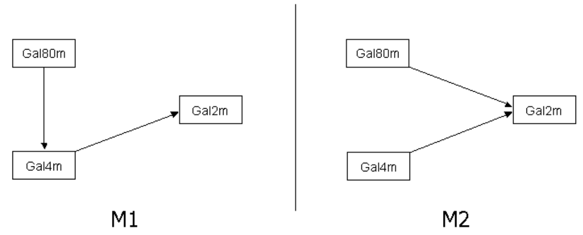

Figure 2. *Simplified Bayesian networks for describing a portion of the galactose system in yeast. These simplified versions of M1 and M2 capture the kernel of the conditional independence assertions of the more complex models of Figure 1. As above, in M1, Gal2m is independent of Gal80m when conditioned on Gal4m, and in M2, Gal4m is marginally independent of Gal80m.*

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.[c] Results of this analysis are shown in Figure 3. As is evident from the figure, the models fall into two primary groupings based on their score: those that include an edge between Gal80 and Gal2 (unshaded) which score between -34.1 and -35.4, and those

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

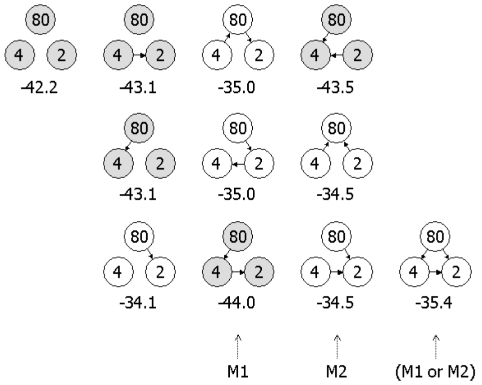

Figure 3. *Scores for all model equivalence classes of the three variable galactose system. The classes of models that score poorly are shown shaded. The previously considered models M1, M2, and (M1 or M2) are indicated.*

that do not include an edge between Gal80 and Gal2 (shaded) which score between -42.2 and -44.0. This lends support to the claim that Gal80 and Gal2 are very unlikely to be conditionally independent given Gal4, again consistent with the currently accepted hypothesis.

5 Representing and scoring models with annotated edges

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, we consider here only four types in the context of binary variables:

- An *unannotated* edge from X to Y represents a dependence that can be arbitrary (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 will 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 instantiations $\mathcal{I}$ of the variables in $\mathrm{Pa}(Y)/X$, we require $\mathrm{P}(Y=1|X=1,\mathcal{I}) > \mathrm{P}(Y=1|X=0,\mathcal{I})$.

- A *negative ("−")* edge from X to Y indicates that higher values of X will 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 instantiations $\mathcal{I}$ of the variables in $\mathrm{Pa}(Y)/X$, we require $\mathrm{P}(Y = 1|X = 0, \mathcal{I}) > \mathrm{P}(Y = 1|X = 1, \mathcal{I})$.

- 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 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 allows 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 since unannotated edges are always permitted. It also permits 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 − annotations being added incrementally as activators and repressors are 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.

6 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 Gal4 and Gal2 is considered in model M2,

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

		Gal4 → Gal2		
		−	+/−	+
Gal80	−	-45.33	-44.58	-44.16
↓	+/−	-44.59	-43.83	-43.41
Gal4	+	-44.17	-43.41	-42.98

M1

		Gal4 → Gal2		
		−	+/−	+
Gal80	−	-48.89	-47.27	-46.68
↓	+/−	-35.53	-35.44	-35.36
Gal2	+	-34.83	-34.75	-34.66

M2

which is consistent with the results of Figure 3 indicating that the coupling between Gal4 and Gal2 is indeed quite weak. In contrast, adding a + annotation to the edge between Gal80 and Gal2 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 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.

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.[30] 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. This example reveals that caution must be used when interpreting results from models that are incomplete.

7 Discussion/Conclusion

The galactose example is intended to illustrate that expression array data can be quite useful in elucidating regulatory networks. While nine of the 52 experiments were carbon source time-series experiments, it should be noted that none of the 52 was performed with the goal of distinguishing between these two models. Nevertheless, they were successfully exploited to select the currently accepted model over the one that had previously been postulated to be true, as well as clarifying the degree and sign of the dependencies between the variables in these data sets. As more experiments become available and more complex models are formulated, these methods will be able to distinguish between subtle differences in proposed models in ways that are not possible without computational assistance.

As previously discussed, model scores depend on the available data, which has two implications. First, while Bayesian networks are well-suited to dealing robustly with noisy data, as noise increases, the score difference between correct and incorrect models (and thus the significance) goes down. In the limit of uninformative data, correct models will score as poorly as incorrect ones, which is to be expected. Second, the ability of particular data to enhance score difference between models 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.

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. We are currently working to develop model induction methods, especially ones that are feasible in models with latent variables.

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.

Acknowledgments

The authors wish to thank Tarjei Mikkelsen for his significant contributions to this project and anonymous reviewers for their helpful comments. In addition, Hartemink gratefully acknowledges support from the NIH and Merck.

References

1. F. C. Holstege, *et al. Cell*, 95:717–728, 1998.
2. M. B. Eisen, *et al. PNAS*, 95:14863–14868, 1998.
3. P. T. Spellman, *et al. Mol. Biol. Cell*, 9:3273–3297, 1998.

4. V. R. Iyer, *et al.* *Science*, 283:83–87, 1999.
5. A. A. Alizadeh, *et al.* *Nature*, 403:503–511, 2000.
6. X. Wen, *et al.* *PNAS*, 95:334–339, 1998.
7. P. Tamayo, *et al.* *PNAS*, 96:2907–2912, March 1999.
8. S. Tavazoie, *et al.* *Nat. Genet.*, 22:281–285, 1999.
9. P. D'haeseleer, *et al.* In R. C. Paton and M. Holcombe, editors, *Information Processing in Cells and Tissues*, 203–212. Plenum Publishing, 1998.
10. A. J. Butte and I. S. Kohane. In *Pac. Symp. Biocomp.*, 5:415–426, 2000.
11. F. P. Roth, *et al.* *Nat. Biotech.*, 16:939–945, 1998.
12. T. Chen, *et al.* In *RECOMB 1999.* ACM-SIGACT, April 1999.
13. R. Thomas. *J. Theor. Biol.*, 42:563–585, 1973.
14. S. Liang, *et al.* In *Pac. Symp. Biocomp.*, 3:18–29, 1998.
15. T. Akutsu, *et al.* In *Pac. Symp. Biocomp.*, 4:17–28, 1999.
16. L. Glass and S. A. Kauffman. *J. Theor. Biol.*, 39:103–129, 1973.
17. H. H. McAdams and A. Arkin. *Annu. Rev. Biophys. Biomol. Struct.*, 27:199–224, 1998.
18. A. Arkin, *et al.* *Genetics*, 149:1633–1648, 1998.
19. N. Friedman, *et al.* In *RECOMB 2000.* ACM-SIGACT, April 2000.
20. J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.
21. S. Lauritzen and D. Spiegelhalter. *J. Roy. Stat. Soc. B*, 50:154–227, 1988.
22. F. Jensen. *Introduction to Bayesian networks.* Springer-Verlag, 1996.
23. T. Dean and K. Kanasawa. *Computational Intelligence*, 5(3):142–150, 1989.
24. U. Kjaerulff. In *Proc. 8th Ann. Conf. on Uncertainty in Artif. Intell.*, 121–129. Morgan Kaufmann, 1992.
25. X. Boyen and D. Koller. In *Proc. 14th Ann. Conf. on Uncertainty in Artif. Intell.*, 33–42. Morgan Kaufmann, 1998.
26. G. Cooper and E. Herskovitz. *Mach. Learn.*, 9:309–347, 1992.
27. D. Heckerman, *et al.* *Mach. Learn.*, 20(3):197–243, 1995.
28. T. Jaakkola and M. Jordan. *J. of Artif. Intell. Res.*, 10:291–322, 1999.
29. H. Attias. In *Proc. 15th Ann. Conf. on Uncertainty in Artif. Intell.*, 21–30. Morgan Kaufmann, 1999.
30. M. Johnston. *Microbiological Rev.*, 51(4):458–476, 1987.