21

Bayesian Networks and Informative Priors: Transcriptional Regulatory Network Models

ALEXANDER J. HARTEMINK
Duke University

Abstract

We discuss the use of Bayesian networks as robust probabilistic models of the multivariate statistical dependencies among interacting variables in transcriptional regulatory networks. We explain how principled scores can be computed to compare network models with one another in terms of their ability to explain observed data simply. With principled scores, we can automatically learn static or dynamic network models that provide simple explanations for a variety of high-throughput data. We make a case for, and demonstrate the utility of, informative priors over network structures and parameters: informative priors can be used to incorporate different kinds of data into the learning process, and also to guide the learning process toward network models that exhibit greater biological plausibility. Results from both simulated and experimental data illustrate the benefits of this modeling framework.

21.1 Introduction

Proteins are the primary molecular workhorses of the cell, playing significant roles in metabolism, biosynthesis and degradation, transport, homeostasis, structure and scaffolding, motility, sensing, signaling and signal transduction, replication, and repair. However, one of the most intriguing roles for proteins is that of transcriptional regulation: control of precisely which genes are being transcribed into RNA at any given time. Since ribosomes subsequently translate most of this RNA into protein, proteins are in large part responsible for regulating their own existence. Although much has been learned about the large network of molecular interactions that regulate transcription, it would probably be fair to say that far more still remains to be learned.

Discovering and understanding the operation of large transcriptional regulatory networks is clearly an important problem in both molecular and synthetic

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

biology. Progress has been accelerated by the advent of high-density DNA microarrays, which can be used to profile levels of RNA expression across the entire genome. When this expression data started becoming available after 1996, the first generation of methods for analyzing it were data-driven, initially unsupervised (e.g., clustering, correlation, and visualization) and later also supervised (e.g., classification). While these kinds of data-driven methods are useful in uncovering interesting patterns in the data, they typically provide little traction for explaining the biological mechanisms that give rise to these patterns. To overcome this limitation, model-driven methods for analyzing RNA expression data began to be developed.

The suggestion that probabilistic graphical models – and Bayesian networks in particular – might serve as an appropriate framework for representing transcriptional regulatory networks and learning models of their structure in the presence of noisy high-throughput RNA expression data was made independently in 1999 at least twice (and, in all likelihood, more than twice): by Murphy and Mian in an unpublished technical report [23], and by Hartemink and colleagues in an invited talk [11]. The first publications demonstrating the utility of this approach seem to be the independent work of Friedman et al. [8] and Hartemink et al. [14], shortly thereafter. These papers share a common theme beyond the choice of Bayesian network models: both realized that there was not enough available expression data to accurately learn large networks (a fact that has later been demonstrated repeatedly through simulation studies [18, 34]). Friedman and colleagues responded by focusing on network features that occur with high frequency in a bootstrap analysis, while Hartemink and colleagues restricted their attention to small sets of network models and investigated the utility of biologically relevant informative parameter priors.

However, RNA expression data is not the only high-throughput data available for providing insight into transcriptional regulatory networks: DNA sequence, protein-DNA binding, and protein–protein interaction data are also available. Protein-DNA binding can be assayed in vitro using a protein binding microarray (PBM) [22] or in vivo using chromatin immunoprecipitation followed by microarray (ChIP-chip; sometimes called transcription factor binding location analysis, or location analysis for short) [26]. Protein-protein interactions can be queried experimentally using techniques like yeast two-hybrid (Y2H) or affinity purification followed by mass spectrometry (AP-MS). Incorporating evidence from multiple kinds of data can often overcome the limitations of any one kind of data, because data collected using different technologies usually offer different perspectives on a problem; jointly analyzing such data in a single framework enables a consensus perspective to emerge. In addition, analysis of many kinds of data together is likely to produce more accurate results since

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

noise characteristics and biases of the various technologies should be largely independent.

Many have recognized the value in jointly analyzing disparate kinds of biological data. Marcotte and colleagues made substantial early progress in refining our understanding of protein-protein interactions by integrating multiple kinds of data (see, e.g., [21]), while Ideker and colleagues used both RNA and protein expression data in predicting the effects of perturbations on regulatory networks [19]. Our work in 2001 [12] and early 2002 [15] introduced two new concepts: first, using various kinds of data to derive informative structure priors for guiding Bayesian inference, an idea later extended by Nariai et al. [24] to protein-protein interaction data; and second, combining RNA expression data with protein-DNA binding data, an idea which has been the basis of many later developments in the field (see, e.g., [2, 28] and [32] (this volume)).

After a brief introduction to Bayesian networks and Bayesian network inference (see also [4] (this volume), which provides a more lengthy introduction), we summarize our work on the use of informative priors for learning Bayesian models of transcriptional regulatory networks. We also present examples of the application of this approach to both simulated and actual experimental data.

21.2 Bayesian Networks and Bayesian Network Inference

Imagine a set of N random variables $\boldsymbol{X} = \{X_i\}_{i=1}^{N}$ and consider how these variables may depend on one another. At one end of the spectrum, the variables may be completely independent; at the other end of the spectrum, they may be completely interdependent, which is to say that no two random variables are independent of one another, even conditioned on a subset of the other variables. *Bayesian networks* are a class of models for representing and reasoning about sets of random variables with conditional dependence relationships within this full range of possibilities – from complete independence to complete interdependence.

A Bayesian network (BN) is a graphical model: it uses a graph to represent information about the conditional independencies among random variables. In our context of modeling transcriptional regulatory networks, variables might represent RNA concentrations, protein concentrations, protein modifications or complexes, metabolites or other small molecules, experimental conditions, genotypic information, or conclusions such as diagnosis or prognosis. Variables can be discrete or continuous. To simplify exposition, we consider only discrete variables for the remainder of this chapter. Each variable is thus in one of a finite set of states, and the number of states used to model a variable represents a tradeoff between on the one hand, capturing a variable's behavior with sufficient

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107415324.004
Cambridge Books Online © Cambridge University Press, 2009
http://dx.doi.org/10.1017/CBO9780511584589.022

precision, and on the other hand, retaining an ability to interpret what the states of the variable mean, as well as managing the computational and statistical complexity of learning models over variables with large numbers of states.

Each directed edge in the graph represents a conditional dependency between a pair of random variables; more precisely, the absence of a directed edge between two vertices represents a conditional independency between the corresponding pair of random variables. These conditional independencies are all summarized by the so-called *Markov property*: variables are conditionally independent of their nondescendants given their parents. The fewer edges a model has, the more constrained the model is. Thus, a graph over completely independent random variables is empty, while a graph over completely interdependent random variables is complete. In practice, we seek sparser (simpler) models because they are able to explain "indirect" dependencies through more "direct" dependencies mediated by other variables.

In characterizing the conditional dependencies among a set of random variables, not only does a BN provide a qualitative description in the form of a graph, but it also provides a quantitative description. Following from the Markov property, the joint probability distribution over the space of variables can be factored into a product over variables, where each term is a probability distribution for that variable conditioned on the set of its parent variables:

$$\Pr(\boldsymbol{X}) \equiv \Pr(X_1, \ldots, X_N) = \prod_{i=1}^{N} \Pr(X_i \mid \mathbf{Pa}(X_i)). \tag{21.1}$$

We will denote by $\boldsymbol{\theta}$ the parameters that collectively characterize the conditional probability distributions on the right-hand side of (21.1).

Because variables can have many parents, BNs are not limited to pairwise interactions between genes, but rather can describe arbitrary combinatorial control of transcriptional regulation; this is particularly straightforward when working with discrete variables. Also, due to their probabilistic nature, BNs are robust in the face of both noisy data and imperfectly specified transcriptional regulatory networks.

21.2.1 Dynamic Bayesian Networks

A dynamic Bayesian network (DBN) [9] extends the notion of a BN to model the stochastic evolution of a set of random variables over time; the structure of a DBN thus describes the qualitative nature of the dependencies that exist between variables in a temporal process. We use $X_i[t]$ to denote the random variable X_i at time t and the set $\boldsymbol{X}[t]$ is defined analogously. Here, the evolution

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

of the temporal process is assumed to occur over discrete time points indexed by the variable $t \in \{1, \ldots, T\}$, although continuous time DBNs also exist [25]. Under such an assumption, we have $T \times N$ interacting random variables where previously we had N. The resultant joint probability distribution is

$$\Pr(\boldsymbol{X}[1], \ldots, \boldsymbol{X}[T]) = \prod_{t=1}^{T} \left[\prod_{i=1}^{N} \Pr(X_i[t] \mid \mathbf{Pa}(X_i[t])) \right]. \tag{21.2}$$

To simplify the situation, we make two further assumptions. First, we assume that each variable depends only on variables that temporally precede it. This fairly innocuous assumption still allows us to model natural cyclic phenomena like feedback loops, but guarantees that the underlying graph will be acyclic. It also greatly simplifies the computational complexity of learning. As one example, if we were to further restrict variables in our network to have at most k parents, we could find the globally optimal network in polynomial time: $O(N^{k+1})$. Second, we assume the process is a stationary first-order Markov process, which means that $\Pr(\boldsymbol{X}[t] \mid \boldsymbol{X}[t-1], \ldots, \boldsymbol{X}[1], t) = \Pr(\boldsymbol{X}[t] \mid \boldsymbol{X}[t-1])$. Given these two assumptions, the variables in $\mathbf{Pa}(X_i[t])$ are a subset of $\boldsymbol{X}[t-1]$. The underlying acyclic graph with $T \times N$ vertices can now be compactly represented by a (possibly cyclic) graph with N vertices, where an edge from X_i to X_j indicates that $X_j[t]$ depends on $X_i[t-1]$.

21.2.2 Scoring Models with the Bayesian Scoring Metric

To learn a network model from observed data, we want to maximize some scoring function that describes the ability of a network to explain the observed data simply. In the case of BNs, we can employ the Bayesian scoring metric (BSM). The scores produced by the BSM permit us to rank alternative models, and the score difference for any two models leads to a direct significance measure for determining how strongly one should be preferred over the other. According to the BSM, the score of a model is defined as the logarithm of the probability of the model being correct given the observed data. Formally,

$$\mathrm{BSM}(S) = \log \Pr(S \mid D) = \log \Pr(S) + \log \Pr(D \mid S) + c, \tag{21.3}$$

where the first term in the last expression is the log prior distribution of the model structure S, the second term is the log (marginal) likelihood of the observed data D given S, and c is a constant that does not depend on S and can thus be safely ignored when comparing structures on the basis of their scores.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

The marginal likelihood term can be expanded as

$$\Pr(D \mid S) = \int_{\boldsymbol{\theta}} \Pr(D, \boldsymbol{\theta} \mid S)\, \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \Pr(D \mid \boldsymbol{\theta}, S) \Pr(\boldsymbol{\theta} \mid S)\, \mathrm{d}\boldsymbol{\theta}, \qquad (21.4)$$

and is analytically tractable when the data are complete and the variables in the network are discrete [16], as we assume here. Marginalizing in this way introduces an inherent penalty for model complexity, thereby balancing a model's ability to explain observed data with its ability to do so simply. Consequently, it guards against overfitting models to data when data are limited.

21.2.3 Learning Networks: Model Selection and Averaging

Finding the highest-scoring model under the BSM for a given set of data is known to be NP-complete in the general case [5] (although see the discussion of DBNs above). As a result, in the general case, we resort to heuristic search strategies to find good models. Commonly used strategies include greedy hill-climbing, greedy random, genetic algorithms, Metropolis, and simulated annealing. We have implemented each of these search strategies and have observed that in our own context, simulated annealing seems to find the highest-scoring models, although in many cases greedy methods identify the same models in much less time. The temperature schedule we employ allows for "reannealing" after the temperature becomes sufficiently low.

We need not select only a single maximum a posteriori model. A more principled Bayesian approach is to compute probabilities of features of interest by averaging over the posterior model distribution. Using model averaging in this way reduces the risk of overfitting the data by considering a multitude of models when computing probabilities of features of interest. For example, if we are interested in determining whether the data D support the inclusion of an edge from variable X_i to variable X_j, we compute

$$\Pr(\mathcal{E}_{ij} \mid D) = \sum_S \Pr(\mathcal{E}_{ij} \mid D, S) \cdot \Pr(S \mid D) = \sum_S 1_{ij}(S) \cdot e^{\mathrm{BSM}(S)},$$

where $\mathcal{E}_{ij}$ is a binary random variable representing the existence of an edge from X_i to X_j, and $1_{ij}(S)$ is an indicator function that is 1 if and only if S has an edge from X_i to X_j. However, this sum is difficult to compute because the number of structures S is enormous. Fortunately, it is possible to approximate this sum by sampling, or since the vast bulk of its mass lies among the highest-scoring models, to further approximate by restricting our attention to the highest-scoring models we encounter in our search. We then compute an appropriately normalized version of the last expression using only these models.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

21.2.4 Prior Establishment

In a Bayesian setting, we need to establish prior distributions both over parameters $\boldsymbol{\theta}$ and over network structures S. In a discrete BN satisfying reasonable assumptions, the prior over parameters must be a product-of-Dirichlet distribution [17]. If prior information about parameters is available, this can be captured in the form of an equivalent prior network [17]. Otherwise, an uninformative prior is frequently employed. In either case, an "equivalent sample size" needs to be specified, which is a measure of how confident we are in the prior relative to the quantity of data.

An especially common choice for the prior over structures is to assume that it is uniform; in this case, the corresponding term in (21.3) can be safely ignored since it is the same for all structures. In the rare instance where an informative prior is chosen, it is typically hand-constructed by domain experts [16]. Here, we summarize a novel approach for automatically constructing informative priors over network structures based on evidence provided by other kinds of data.

21.3 Adding Informative Structure Priors

To complement RNA expression data, we can extend our BN framework to include data describing the genome-wide DNA binding locations of protein transcription factors. If a transcription factor is reported to bind DNA upstream of a particular gene, it provides evidence that the factor is involved in the regulation of that gene. We can incorporate this evidence when scoring our BN models by modifying the prior distribution over structures. The Bayesian methodology has a natural provision for incorporating prior information into its scoring metric; in practice, determining appropriate weights for diverse sources of information poses a significant challenge.

We can incorporate an informative structure prior in two ways. First, we can adopt what we call a "hard" prior, which is uniform except that it gives zero probability to structures that do not satisfy constraints specifying which edges are required to be present and which are required to be absent. We can implement this prior by restricting our search algorithms to move only through the space of valid network structures. While this means we search in a smaller space, it is inconsistent with the notion that high-throughput data are generally quite noisy. Second, and alternatively, we can adopt what we call a "soft" prior, with varying but positive weights on all networks, down-weighting rather than excluding structures that do not satisfy the constraints. In such a setting, protein-DNA binding location data provides evidence as to whether a regulatory relationship exists, and the more significant the location data (lower

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

the p value), the more likely the edge is to be included. As a consequence, this prior is subtler and more robust; nevertheless, it remains factorable in the context of a DBN, enabling computationally efficient local search. In fact, we can learn a DBN model using a soft informative structure prior in essentially the same amount of time as with an uninformative uniform prior. We describe this in more detail in the following sections.

21.3.1 Probability of an Edge Being Present

Transcription factor binding location data provides (noisy) evidence regarding the existence of regulatory relationships between a transcription factor and each of the genes in the genome. This evidence is reported as a p value, and the probability of an edge being present in the true regulatory network is inversely related to this p value: the smaller the p value, the more likely the edge is to exist in the true structure. In previously published work [3], we provide a detailed derivation of a function for mapping p values to corresponding probabilities of edges being present in structure S, but here we simply state the results. Let β denote $\mathsf{Pr}(\mathcal{E}_{ij})$, the prior probability that an edge exists from X_i to X_j, which we take to be constant for all i and j. Using Bayes' rule, we can show that the probability of $\mathcal{E}_{ij}$ after observing the corresponding p value is

$$\mathsf{Pr}_\lambda(\mathcal{E}_{ij} \mid \mathcal{P}_{ij} = p) = \frac{\lambda\, e^{-\lambda p}\beta}{\lambda\, e^{-\lambda p}\beta + (1 - e^{-\lambda})(1 - \beta)}, \tag{21.5}$$

where λ is a parameter controlling the amount of confidence we place in the reported p values as accurate indicators of binding and nonbinding. Some insight into the role of λ can be gained by considering the value p^* obtained by solving the equation $\mathsf{Pr}_\lambda(\mathcal{E}_{ij} \mid \mathcal{P}_{ij} = p^*) = 1/2$, which yields

$$p^* = \frac{-1}{\lambda} \log\left[\frac{(1 - e^{-\lambda})(1 - \beta)}{\lambda\beta}\right]. \tag{21.6}$$

For any fixed value of λ, an edge from X_i to X_j is more likely to be present than absent if the corresponding p value is below this critical value p^* (and vice versa). As we increase the value of λ, the value of p^* decreases and we become more stringent about how low a p value must be before we consider it as prior evidence for edge presence. Conversely, as λ decreases, p^* increases and we become less stringent; indeed, in the limit as $\lambda \to 0$, we have that $\mathsf{Pr}_\lambda(\mathcal{E}_{ij} \mid \mathcal{P}_{ij} = p) \to \beta$ independent of p, revealing that if we have no confidence in the location data, the probability of $\mathcal{E}_{ij}$ is the same value β both before and after seeing the corresponding p value, as expected. Thus, λ acts

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

as a tunable parameter indicating the degree of confidence in the evidence provided by the location data; this allows us to model our belief about the noise level inherent in the location data and correspondingly, the amount of weight its evidence should be given.

One approach to suitably weighing the evidence of the location data would be to select a single value for λ, either through parameter estimation or by some heuristic like finding the value of λ that corresponds to a certain "magic" value for p^*, such as 0.001. Instead, we adopt a Bayesian approach that places a prior on λ and then marginalizes over it. The net effect of marginalization is an edge probability that is a smoother function of the reported p values than without marginalization.

21.3.2 Prior Probability of a Structure

The prior probability of a structure $\mathsf{Pr}(S)$ is proportional to the following product over the edges in S:

$$\prod_{\{ij:\ 1_{ij}(S)=1\}} \left[\frac{\mathsf{Pr}(\mathcal{E}_{ij} \mid \mathcal{P}_{ij} = p)}{1 - \mathsf{Pr}(\mathcal{E}_{ij} \mid \mathcal{P}_{ij} = p)} \right]. \tag{21.7}$$

The normalizing constant can be safely ignored since it is the same for all structures. Analogous to the likelihood calculations, the calculations required for updating the structure prior under a local change to S are computationally efficient because the structure prior factors over the edges in S, as shown in (21.7). In particular, we need not recompute the entire prior from scratch with each local change.

Note that in the absence of location data pertaining to a particular edge, we simply use the probability $\mathsf{Pr}(\mathcal{E}_{ij}) = \beta$ for that edge. Our informative prior is thus a natural generalization of traditional priors: in the absence of any location data whatsoever, the prior probability of a network structure is exponential in the number of edges in the graph, with edges favored if we choose $\beta > 0.5$ and edges penalized if we choose $\beta < 0.5$. In the special case where $\beta = 0.5$, the prior over structures is uniform.

21.4 Applications of Informative Structure Priors

In this section, we present two examples of how BN models and informative structure priors can be used to elucidate transcriptional regulatory networks in the yeast *Saccharomyces cerevisiae*. We examine networks responsible for

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139150996.022
Cambridge Books Online © Cambridge University Press, 2009

controlling the expression of various genes that code for proteins involved in pheromone response and in cell cycle regulation. With respect to the former, the protein Ste12 is the ultimate target of the pheromone response signaling pathway and binds DNA as a transcriptional activator for a number of other genes. Transcription factor binding location data indicates which intergenic regions in the yeast genome are bound by Ste12, both in the presence and absence of pheromone [26]. With respect to the latter, a number of known cell cycle transcription factors have also been profiled by location analysis [20, 29].

To demonstrate the full range of methods discussed above, in the first example we use a model averaging approach with a hard informative structure prior to learn a static BN model, while in the second example we use a model selection approach with a soft informative structure prior to learn a dynamic BN model.

21.4.1 Pheromone Response: Static Network, Model Averaging

In the case of pheromone response, we used a set of 320 samples of unsynchronized *S. cerevisiae* populations of various wild-type and mutant strains grown under a variety of environmental conditions including exposure to different nutritive media as well as exposure to stresses like heat, oxidative species, excessive acidity, and excessive alkalinity. Genome-wide RNA expression data for each of these 320 observations were collected using four low-density 50 μm Affymetrix Ye6100 GeneChips per observation (roughly a quarter of the genome can be measured on each chip). The reported "average difference" values from these 1,280 Affymetrix GeneChips were normalized using maximum a posteriori normalization methods based on exogenous spiked controls [13].

From the 6,135 genes of the *S. cerevisiae* genome, 32 were selected either on the basis of their participation in the pheromone response signaling cascade or as being known to affect other aspects of mating response in yeast. The normalized levels of RNA expression for these 32 genes were log-transformed and discretized using discretization level coalescence methods that incrementally reduce the number of discretization levels for each gene while preserving as much total mutual information between genes as possible [12]. In this case, each gene was discretized to have four levels of discretization while preserving over 98% of the original total mutual information between pairs of genes [12]. In addition to the 32 variables representing levels of RNA expression, an additional variable named `mating_type` was considered. The variable `mating_type` represents the mating type of the various haploid strains of yeast used in the 320 observations and can take one of two values, corresponding to the MATa and MATα mating types of yeast, respectively.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

21.4.1.1 Results Using Experimental Data

We used simulated annealing to visit high-scoring regions of the model posterior and present the results of two of those runs here. In the first run, we traversed the model space with a uniform structure prior. In the second run, we incorporated a hard informative structure prior using available location data by requiring edges from STE12 to FUS1, FUS3, AGA1, and FAR1 which had p values less than 0.001.

After gathering the 500 highest-scoring models that were visited during each run of the search algorithm, we computed the probability of edges being present using model averaging, as discussed above. Thus, the estimated probability of an edge can be exactly 1 if (and only if) the edge appears in all 500 highest-scoring models.

We then compiled a composite network for each run that consists of all edges with estimated posterior probability over 0.5. These networks are shown in Figure 21.1. Nodes have been augmented with color to indicate groups of variables known in the literature to have some commonality with one another. Edges have also been augmented with color: solid black edges have posterior probability of 1, solid blue edges have probability between 1 and 0.99, dashed blue edges have probability between 0.99 and 0.75, and dotted blue edges have probability between 0.75 and 0.5. The strength of an edge does not indicate how *significantly* a parent node contributes to the ability to explain the child node but rather an approximate measure of how *likely* a parent node is to contribute to the ability to explain the child node.

In both of the networks presented in Figure 21.1, we observe a number of interesting properties. In each case, the `mating_type` variable is at the root of the graph, and contributes to the ability to predict the state of a large number of variables, which is to be expected. The links are generally quite strong indicating that their presence was fairly consistent among the 500 highest-scoring models. Almost all the links between `mating_type` and genes known to be expressed only in MATa or MATα strains occur with posterior probability above 0.99. Moreover, in both networks there exists a directly connected subgraph consisting of genes expressed only in MATa cells (magenta) and a directly connected subgraph consisting of genes expressed only in MATα cells (red). In each case the subgraph has the `mating_type` variable as a direct ancestor with strong predictive power, as expected.

The heterotrimeric G-protein complex components GPA1, STE4, and STE18 (green) form a directly connected component with the informative prior but only GPA1 and STE18 are connected with the uniform prior. Indeed, even the link between GPA1 and STE4 with the informative prior is fairly weak. On the other hand, SWI1 and SNF2 (aqua) are weakly adjacent with a uniform

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

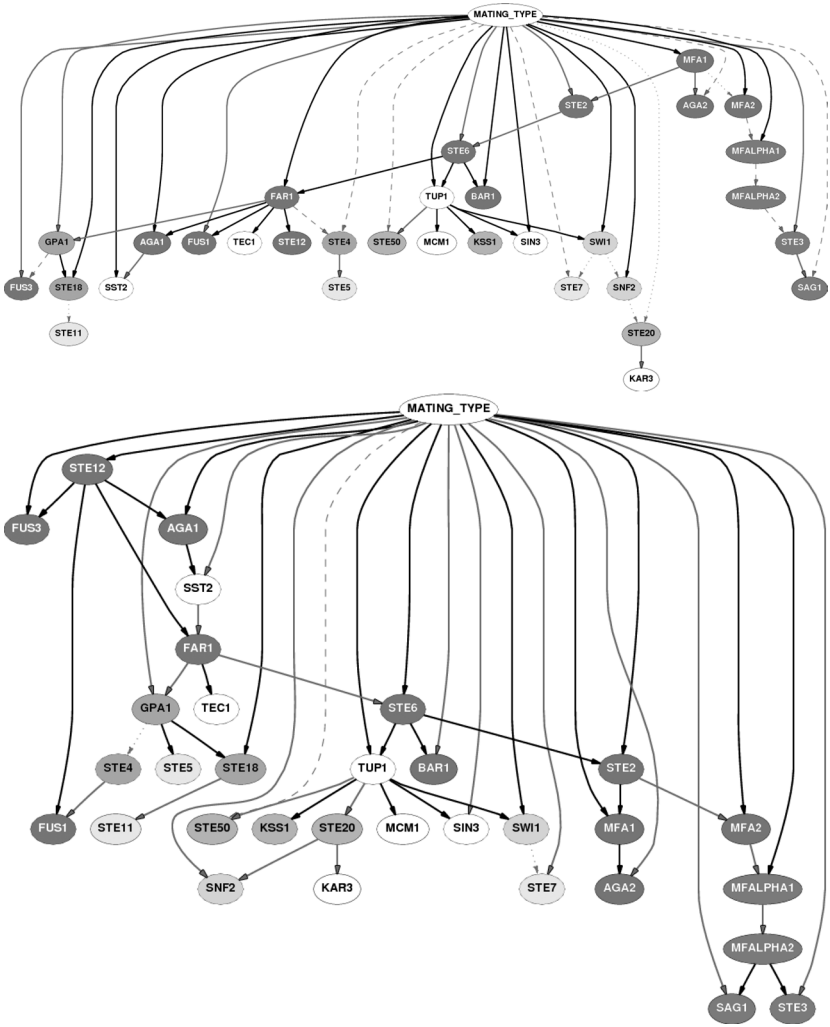

Fig. 21.1. Bayesian network models learned by model averaging over the 500 highest-scoring models visited during the simulated annealing search runs with a uniform prior and hard informative prior, respectively. Edges are included in the figure if and only if their posterior probability exceeds 0.5. Node and edge color descriptions are included in the text. (See color plate 21.1.)

prior, but not adjacent with an informative prior, though in both cases they are close descendants of TUP1. STE11 and STE5, two of the core elements of the primary signaling cascade complex (yellow), are seen as descendants of G-protein complex genes, indicating statistical dependence that may be the result of common or serial regulatory control. STE7 occurs elsewhere, however.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

Auxiliary signaling cascade genes (orange) are always descendants of TUP1, sometimes directly and sometimes more indirectly, but STE50 and KSS1 are siblings in both cases. In general, the auxiliary cascade elements do not tend to cluster with the core elements, suggesting that the regulation of their transcript levels may occur by a different mechanism than those of the genes in the core signal transduction complex.

In both networks, TUP1 appears with a large number of children, consistent with its role as a general repressor of RNA polymerase II transcription. Both networks have MCM1 and SIN3 as children of TUP1; Tup1 and Mcm1 are known to interact in the cell [10] and this result that the level of Tup1 is helpful in predicting the level of Mcm1 suggests a possible regulatory relationship between the two. FAR1 is a parent of TEC1 and GPA1 in both networks. Far1, Tec1, and Gpa1 are all known to be cell-cycle-regulated and all three are classified as being transcribed during early G_1 phase [6]. This result suggests that Far1 may play a role in regulating the expression of Tec1 and Gpa1, providing a possible mechanism for their previously observed G_1 phase coexpression.

Though it is produced at higher levels in MATa cells, it is known that Aga1 is produced in both MATa and MATα cells [27]. The networks are each consistent with this knowledge, including a frequent predictive edge from `mating_type` to AGA1, but not clustering AGA1 with other mating type specific genes (magenta and red) as it is likely regulated differently. In both networks, AGA1 and SST2 are adjacent, consistent with the fact that the two are expressed very similarly, both peaking at the M/G_1 phase of the cell cycle [31].

21.4.2 Cell Cycle: Dynamic Network, Model Selection

Turning our attention now to the cell cycle, we earlier assumed that the stochastic dynamics of variables in a DBN arise from a stationary process. This poses a bit of a problem in the case of the cell cycle since we may have a different underlying transcriptional regulatory network during each phase of the cycle. To overcome this problem, we can employ an additional variable ϕ that can be used by the model to explain how each variable's regulators depend on the cell cycle phase, allowing us to model a different stationary process within each phase. The phase variable ϕ is multinomial and the number of states is simply the number of phases we choose to model as having distinct regulatory networks. If we can label each of the time points with the appropriate phase, the inference problem remains an instance of learning network structure from complete data. We prefer this option to the alternative of learning a hidden phase variable because in our context, the quantity of available cell cycle expression data is quite limited; besides, the state of ϕ changes smoothly and predictably, so labeling each time point with the appropriate phase is straightforward. A

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

Fig. 21.2. Simplified schematic of a first-order Markov DBN model of the cell cycle. On the left, variables X_1 through X_4 are shown both at time t and $t + 1$; variable ϕ represents the cell cycle phase; dashed edges are stipulated to be present whereas solid edges are recovered by the learning algorithm. On the right, a compact representation of the same DBN model in which the cycle between X_4, X_3, and X_2 is apparent.

simplified schematic of such a DBN model of the cell cycle is depicted in Figure 21.2.

In the context of the cell cycle, we conducted tests with both simulated and actual experimental data. We used a synthetic cell cycle model to evaluate the accuracy of our algorithm and determine the relative utility of different quantities of available RNA expression data. The synthetic cell cycle model involves 100 genes and a completely different regulatory network operates in each of the three modeled phases of the cycle. The 100 genes include synthetic transcription factors, only some of which are involved in the cell cycle, and only some of which have simulated location data available. The target genes of the transcription factors are sometimes activated and sometimes repressed; some are under cell cycle control, but many are not. In addition, we include a number of additional genes whose expression is random and not regulated by genes in the model. The simulated expression data are generated using the (stochastic) Boolean Glass gene model [7]. The expression data are discretized into two states because the generating model is Boolean. Noisy p values for the simulated location data associated with a subset of the regulators are generated with noise models of varying intensity.

For experimental data, we use publicly available cell cycle RNA expression data [31] and transcription factor binding location data [20]. The expression data consist of 69 time points collected over eight cell cycles. Since these belong to different phases, the resultant number of time points in each phase is quite small. As a consequence, we choose to use only three states for the phase variable, by splitting the shortest phase G_2 in half and lumping the halves with the adjacent phases. Thus, the three states of our phase variable correspond roughly to G_1, $S + G_2$, and $G_2 + M$. We assign a phase label for each time point by examining the behavior of characteristic genes known to be regulated during specific phases [31]. This is done separately for each of the four synchronization protocols in the data set (alpha, cdc15, cdc28, and elu). We then select a set of 25 genes to model in our network, of which 10 are known transcription

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

factors for which we have available location data. The only important cell cycle transcription factor missing from this set for which we have location data is FKH2; we are not able to use it in our analysis because RNA expression levels are missing for many of the time points. The remaining 15 genes in our set are selected on the basis of their known regulation by one or more of these 10 transcription factors. The experimental expression data are discretized into three states using interval discretization [12].

Because space is quite limited here, we provide only a brief description of the basic structure of each of our experiments. The discretized data in each case are used to compute the marginal likelihood component of the BSM. The soft informative prior component is computed using (21.7), where individual edge probabilities are computed from the location data p values using (21.5) with λ marginalized out. The parameter λ is marginalized uniformly in the interval $[\lambda_L, \lambda_H]$, with $\lambda_L = 1$ to avoid problems near zero ($\lambda_L = 1$ corresponds to $p^* = 0.459$) and $\lambda_H = 10{,}000$ to avoid problems near infinity ($\lambda_H = 10{,}000$ corresponds to $p^* = 0.001$). We set $\beta = 0.5$ so that edges for which we have no location data are equally likely to be present or absent in the graph; as a consequence, without location data, edge presence in the graph depends on expression data alone. The output of our DBN inference algorithm is the network structure with the highest BSM score among all those visited by the heuristic search during its execution.

21.4.2.1 Results Using Simulated Data

We repeatedly conduct the following three experiments: score network structures with expression data alone, ignoring the log prior component in (21.3); score network structures with location data alone, ignoring the log marginal likelihood component in (21.3); and score network structures with both expression and location data. We use these experiments to evaluate the effects of location data with different noise characteristics, expression data of varying quantity, and different choices for β.

Each of our experiments is conducted on five independently generated synthetic data sets and results are averaged over those five data sets. Figure 21.3 offers a representative result. The vertical axis measures the (average) total number of errors: the sum of false positives and false negatives in the learned network. As expected, the total number of errors drops sharply as the amount of available expression data increases. The figure demonstrates that our joint learning algorithm consistently reduces the total number of false positives and false negatives learned when compared to the error rate obtained using either expression or location data alone. Also, observe that the availability of location data means that we require typically only half as much expression data to

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

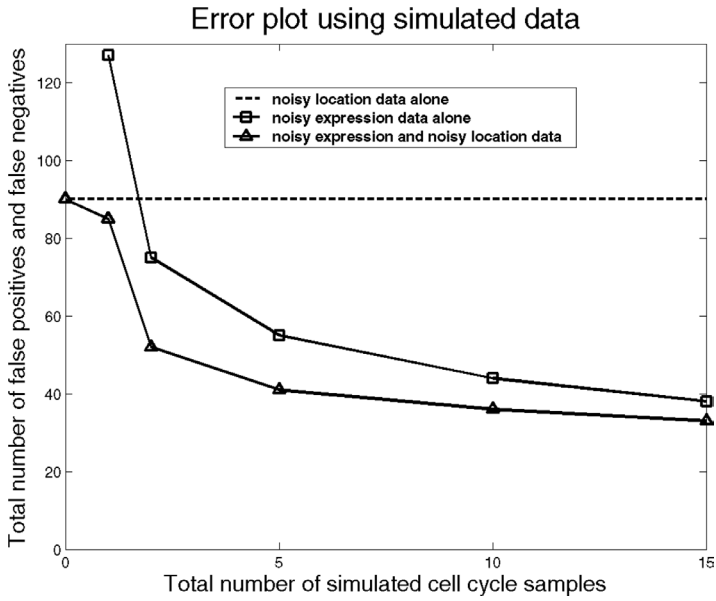

Fig. 21.3. Total number of errors while learning a synthetic cell cycle network using (noisy simulated) expression and location data, separately and with both types of data together. The graph shows the effect of increasing the number of cell cycles worth of expression data, both with and without location data. The dashed horizontal line represents learning using location data alone.

achieve the same error rate as would be achieved with expression data alone, suggesting that the availability of location data can be used to compensate for small quantities of expression data.

21.4.2.2 Results Using Experimental Data

We next apply our soft informative prior in learning networks describing the regulation of transcription during the cell cycle in yeast. As with the simulated data, we learn network structures using expression data alone, using location data alone, and jointly from both expression and location data. In the latter case, we compare our soft informative prior to our hard informative prior with a cutoff of $p = 0.001$ [15].

As an evaluation criterion (which is more difficult in this context than in the synthetic network context), we create a "gold standard" network consisting of the set of edges that are known to exist from one of the 10 transcription factors with both expression and location data to any one of the other genes in our set; we do not count edges from the other 15 genes when comparing with our gold standard since it would be difficult to determine whether recovered edges are

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

Table 21.1. *Comparison of the Highest-Scoring Networks Found in Four Different Experiments with the Gold Standard Network*

Experiment	TP	TN	FP	FN
Expression data only	7	181	20	32
Location data only	25	184	17	14
Expression and location data (hard prior)	23	187	19	11
Expression and location data (soft prior)	28	189	12	11

Note: As discussed in the text, the gold standard contains edges from only the 10 variables for which both location and expression data are available.

true or false positives, and whether omitted edges are true or false negatives. The gold standard comes from a compiled list of evidence in the literature and from the Saccharomyces Genome Database (http://www.yeastgenome.org), but we have tried to ensure that it depends on neither the specific expression data nor the specific location data used in these experiments. Note also that the gold standard is likely not the true underlying regulatory network, but rather is the best we can do given the current understanding of the yeast cell cycle (a "bronze standard"?).

With these caveats in place, Table 21.1 shows the total number of positives and negatives that are true and false for the networks found in the four experiments, with respect to the gold standard network. We see that the location data by itself does noticeably better than the expression data, suggesting that this particular set of location data are quite insightful or that this particular set of expression data are quite limited in quantity or quality. Despite the relatively poor performance of the expression data when considered in isolation, when we use our soft informative prior to include evidence from the location data along with the expression data, the number of false positives and the number of false negatives are both reduced; in contrast, the hard prior reduces the number of false negatives and increases the number of true negatives, but also increases the number of false positives and reduces the number of true positives. The soft prior uniformly outperforms the other three.

From Table 21.1, we see that combining expression and location data with our soft informative prior results in three fewer false negatives as compared to location data alone. These three are the regulation of SIC1 by ACE2 ($p = 0.010$), of ACE2 by FKH1 ($p = 0.006$), and of CLN2 by SWI4 ($p = 0.005$). These edges are detected because while the evidence of the location data in isolation is below threshold for inclusion, during the joint learning it is reinforced with evidence from expression data. In contrast, consider the regulation of transcription

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

factor FKH1 by the transcription factor MBP1: although this interaction is recovered with expression data alone, it is not included when both location and expression data are used because the corresponding p value of 0.93 is so high that the quantity of expression data is insufficient to overcome the location data evidence against inclusion of the edge. Among the supposed false positives, we observe that both location and expression data provide evidence for the regulation of cyclin PCL2 by the transcription factor SWI6, although this interaction was not reported in the gold standard network. However, SWI6 participates in the SBF complex with SWI4, and SBF is currently believed to regulate the expression of PCL2 [1], reinforcing the notion that our gold standard network is not without flaws.

21.5 Adding Informative Parameter Priors

Thus far, we have discussed the utility of informative priors over structures, but informative priors over parameters are also useful. As with priors over structures, informative priors over parameters can be formulated as a hard prior where parameters that are inconsistent with a set of known constraints are eliminated in advance [14], or as a soft prior where different sets of parameters are relatively up- or down-weighted based on their degree of conformance to the constraints. In the case of a hard prior, the data are not forced to obey the constraints (after all, the data are noisy) but the parameters that characterize the distributions used to model the data are forced to obey these constraints.

What benefit will either of these choices have? The marginal 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 $\boldsymbol{\theta}$ 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 averaged over repeated samplings to compute the marginal likelihood. 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 by a prior in the sense that the distribution of $\boldsymbol{\theta}$ has more 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 BSM. In short, if the constraint permits the model to avoid unnecessary complexity, then the model's score will increase.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

In the case of a hard prior, we simply modify the scoring metric so that the marginal 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 [14]. In the case of a soft prior, we adjust the pseudocounts associated with the Dirichlet priors over parameters so that they are not all the same [33].

21.6 Discussion

BNs have certain limitations when used to model transcriptional 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. In particular, it is important to distinguish between statistical interaction and physical interaction.

For example, if the data strongly supports the inclusion of an edge between two variables X_i and X_j, that *may* indicate a physical interaction between these two factors in the cell. Alternatively, it is possible that an unmodeled variable Y actually intermediates between X_i and X_j or is a latent common cause of X_i and X_j such that X_i and X_j exhibit statistical dependence but no physical interaction. 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_i and X_j, that *may* indicate 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. Incorporating additional complementary sources of data like transcription factor binding location can sometimes clarify the situation.

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 sufficient data do not exist 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.

Similarly, although the interactions in our dynamic models can be oriented unambiguously (because time cannot flow backwards), that does not necessarily imply that the interactions are causal since we cannot account for cellular interactions that have not been measured, as mentioned above. One of the main hopes of this line of research is that more direct causal information from alternative assays like transcription factor binding location data and

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

protein-protein interaction data will ameliorate this problem when we can include them in the analysis framework in a principled way.

From a computational perspective, BN structure inference should scale fine to networks of hundreds of interacting variables, as we have demonstrated here and elsewhere [30]. The primary factor in its ability to scale is not so much computational as statistical, and not so much with respect to the number of variables but with respect to the number of parents for each variable. As this number increases, larger and larger quantities of data are needed to learn an accurate model [35]. On a related note, while nothing precludes us computationally from modeling a higher-order Markov process in our DBNs, we are often constrained statistically by the limited quantity of available time-series expression data.

Successful elucidation of transcriptional regulatory networks will not likely be a batch learning process. Rather, we will need to increasingly consider learning that is incremental and algorithms that are online. In particular, gathering data sampled even sparsely from the joint probability space over all relevant variables in cellular regulatory networks would require an inordinate amount of data. To overcome this, 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 when learning from high-throughput data because the data are often costly to gather. Knowing in advance which are likely to be the most informative experiments to conduct for elucidating biological mechanisms of interest would be quite useful.

This field is known as "active learning" and an existing literature 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 transcriptional regulatory networks, this can be implemented by deleting a gene so that it cannot be expressed, or by constitutively overexpressing a gene from a heterologous promoter. Interventional data needs to be treated differently from observational data in the context of learning, but the framework easily extends to handle interventional data.

Finally, we should offer one last note on the viability of BNs as models of transcriptional regulatory networks in higher eukaryotes. From a biological perspective, regulation in multicellular organisms is quite a bit more complicated owing to extra spatial and temporal complexity, for example in the form of intercellular signaling and differentiation of cell types. From an experimental perspective, collecting data is often more challenging in this context as well because multiple cell types need to be profiled and because there are often technical, financial, or ethical limitations to data collection. However, provided that the models are flexible enough to capture the complexity of

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009

these organisms and provided that sufficient data can be collected, there does not seem to be any fundamental limitation of BNs as useful models for representing and elucidating transcriptional regulation in higher eukaryotes.

21.7 Availability of Papers and Banjo Software

This chapter summarizes a large body of work from our research group over the past six years. As such, it borrows heavily from papers written during this period, but offers a broader and more unified perspective on the research program as a whole. We have in some places omitted details that have been published previously; readers interested in greater detail are encouraged to read the original papers.

In addition to the work on BNs as models of transcriptional regulation summarized in this chapter, our group has undertaken research along a number of other directions related to the various topics treated elsewhere in this book. These include analysis of microarray data, analysis of proteomic spectra, modeling of the eukaryotic cell cycle, motif identification, integration of diverse kinds of data, and disease diagnosis and other classification tasks in high-dimensional systems biology. Bayesian statistical formulations and informative priors arise as common themes in all this work. Papers from our group are available from http://www.cs.duke.edu/∼amink.

Finally, we have recently developed a software package called Banjo – Bayesian network inference with Java objects – to perform network inference in static and dynamic Bayesian networks of discrete variables. Banjo is designed to be efficient, modular, and extensible. The program and complete source code are available under a noncommercial use license, and commercial licensing opportunities are available as well. For more information, visit http://www.cs.duke.edu/∼amink/software/banjo.

21.8 Acknowledgments

I am grateful to my Ph.D. advisors, David Gifford, Tommi Jaakkola, and Rick Young at MIT, for starting me on this path and for continuing to offer advice and shape my thinking. At Duke, Allister Bernard helped me formulate the soft prior over network structures, while Erich Jarvis, V. Anne Smith, and Jing Yu helped develop a simulation framework, which I have not had room to discuss here, that enabled us to improve structure inference algorithms in biological contexts with little data [35]. Jing also helped me formulate the soft monotonicity prior over BN parameters. Jürgen Sladeczek did most of the work in producing our Banjo software package. Finally, I am very thankful for the

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

ideas of many others on this topic, shared either in person or through their writing, including but not limited to David Heckerman, Nir Friedman, Daphne Koller, Greg Cooper, Peter Spirtes, Chris Meek, Max Chickering, Harald Steck, Trey Ideker, Eran Segal, and Satoru Miyano. Finally, I want to thank the editors for inviting this contribution and being patient while I crafted it, especially Marina Vannucci with whom I most frequently interacted.

Bibliography

[1] K. Baetz, J. Moffat, J. Haynes, M. Chang, and B. Andrews. Transcriptional coregulation by the cell integrity MAP kinase Slt2 and the cell cycle regulator Swi4. *Mol. Cell. Biol.*, 21:6515–6528, 2001.

[2] Z. Bar-Joseph, G. Gerber, T. Lee, N. Rinaldi, J. Yoo, F. Robert, B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, 21:1337–1341, 2003.

[3] A. Bernard and A. J. Hartemink. Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. In *Pac. Symp. Biocomp.*, Vol. 10, pp. 459–470, Jan. 2005.

[4] B. M. Broom and D. Subramanian. Computational methods for learning Bayesian networks from high-throughput biological data. In K.-A. Do, P. Müller, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, Chapter 20, pp. 385–400. Cambridge University Press, New York, 2006.

[5] D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H.-J. Lenz, editors, *Learning from Data: AI and Statistics V*, Chapter 12, pp. 121–130. Springer-Verlag, New York, 1996.

[6] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73, July 1998.

[7] R. Edwards and L. Glass. Combinatorial explosion in model gene networks. *Chaos*, 10:691–704, Sept. 2000.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *RECOMB*, Vol. 4, Apr. 2000.

[9] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. Fourteenth UAI*, pp. 139–147. Morgan Kaufmann, San Francisco, 1998.

[10] I. M. Gavin, M. P. Kladde, and R. T. Simpson. Tup1p represses Mcm1p transcriptional activation and chromatin remodeling of an a-cell-specific gene. *Embo J.*, 19:5875–5883, 2000.

[11] A. J. Hartemink. Using HDA data to statistically validate models of genetic regulatory networks in *S. cerevisiae*. EBI conference entitled *Datamining for Bioinformatics: Towards In Silico Biology*, Nov. 1999.

[12] A. J. Hartemink. Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks. PhD thesis, Massachussets Institute of Technology, 2001.

[13] 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)*, pp. 132–140. SPIE, Bellingham, WA, Jan. 2001.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139150538.022
Cambridge Books Online © Cambridge University Press, 2009

[14] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pac. Symp. Biocomp.*, Vol. 6, pp. 422–433, Jan. 2001.

[15] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Pac. Symp. Biocomp.*, Vol. 7, pp. 437–449, Jan. 2002.

[16] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pp. 301–354. Kluwer Academic Publishers, 1998.

[17] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, 20:197–243, 1995.

[18] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19:2271–2282, 2003.

[19] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292:929–934, 2001.

[20] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.

[21] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, 1999.

[22] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, 36:1331–1339, Dec. 2004.

[23] K. Murphy and S. Mian. Modelling gene expression data using dynamic Bayesian networks. Technical Report, University of California at Berkeley, 1999.

[24] N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. In *Pac. Symp. Biocomp.*, Vol. 9, pp. 384–395, Jan. 2004.

[25] U. Nodelman, C. R. Shelton, and D. Koller. Learning continuous time Bayesian networks. In *Proc. Nineteenth UAI*, pp. 451–458, 2003.

[26] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA-binding proteins. *Science*, 290:2306–2309, Dec. 2000.

[27] A. Roy, C. Lu, D. Marykwas, P. Lipke, and J. Kurjan. The AGA1 product is involved in cell surface attachment of the *Saccharomyces cerevisiae* cell adhesion glycoprotein a-agglutinin. *Mol. Cell. Biol.*, 11:4196–4206, 1991.

[28] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19:i273–i282, 2003.

[29] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511584589.022
Cambridge Books Online © Cambridge University Press, 2009

[30] V. A. Smith, E. D. Jarvis, and A. J. Hartemink. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, 18(S1):S216–S224, 2002.

[31] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[32] N. Sun and H. Zhao. A misclassification model for inferring transcriptional regulatory networks. In K.-A. Do, P. Müller, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, Chapter 18, pp. 347–365. Cambridge University Press, New York, 2006.

[33] J. Yu. Developing Bayesian Network Inference Algorithms to Predict Causal Functional Pathways in Biological Systems. PhD thesis, Duke University, May 2005.

[34] J. Yu, A. Smith, P. Wang, A. Hartemink, and E. Jarvis. Using Bayesian network inference algorithms to recover molecular genetic regulatory networks. In *Proc. International Conference on Systems Biology*, Dec. 2002.

[35] J. Yu, A. Smith, P. Wang, A. Hartemink, and E. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20:3594–3603, 2004.

Downloaded from https://www.cambridge.org/core. Duke University Libraries, on 31 May 2019 at 19:22:20, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781584589.022
Cambridge Books Online © Cambridge University Press, 2009