COMBINING LOCATION AND EXPRESSION DATA FOR PRINCIPLED DISCOVERY OF GENETIC REGULATORY NETWORK MODELS

ALEXANDER J. HARTEMINK
Duke University Department of Computer Science
Box 90129, Durham, NC 27708-0129

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

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

We develop principled methods for the automatic induction (discovery) of genetic regulatory network models from multiple data sources and data modalities. Models of regulatory networks are represented as Bayesian networks, allowing the models to compactly and robustly capture probabilistic multivariate statistical dependencies between the various cellular factors in these networks. We build on previous Bayesian network validation results by extending the validation framework to the context of model induction, leveraging heuristic simulated annealing search algorithms and posterior model averaging. Using expression data in isolation yields results inconsistent with location data so we incorporate genomic location data to guide the model induction process. We combine these two data modalities by allowing location data to influence the model prior and expression data to influence the model likelihood. We demonstrate the utility of this approach by discovering genetic regulatory models of thirty-three variables involved in *S. cerevisiae* pheromone response. The models we automatically generate are consistent with the current understanding regarding this regulatory network, but also suggest new directions for future experimental investigation.

1 Introduction

While genomic expression data has proven tremendously useful in providing insights into cellular regulatory networks, other valuable sources of data are increasingly becoming available to aid in this process. The wide range of data modalities presents a significant challenge, but also an opportunity since principled fusion of these diverse information sources will help reveal synergistic insights not readily apparent when sources are examined individually. We approach the information fusion challenge by developing principled methods for the automatic induction (discovery) of genetic regulatory network models

from both genomic location and expression data. In our modeling framework, models of regulatory networks are represented as Bayesian networks, allowing the models to compactly and robustly capture probabilistic multivariate statistical dependencies between the various cellular factors in these networks.[1,2,3,4] Here we extend our previously developed model validation framework based on Bayesian networks[1] to the context of model induction. We discover models by using a heuristic search algorithm based on simulated annealing to visit high-scoring regions of the model posterior and then using posterior model averaging to compute likely statistical dependencies between model variables. We combine genomic location and expression data to guide the model induction process by permitting the former to influence the model prior and the latter the model likelihood.

In this paper, we apply our methodology to examine the regulatory network responsible for controlling the expression of various genes that code for proteins involved in *Saccharomyces cerevisiae* pheromone response pathways. 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. Data from genomic location analysis indicates which intergenic regions in the yeast genome are bound by Ste12, both in the presence and absence of pheromone.[5] Because pheromone response and mating pathways play an essential role in the sexual reproduction of yeast and because we have access to location data regarding the binding locations of Ste12 within the yeast genome, this is a natural choice of regulatory network to examine.

We begin in Section 2 by considering various model induction methodologies. In Section 3 we discuss the collection and preparation of data for model induction in the context of pheromone response. We present various results of our model induction approach in Section 4, including the impact of using data from genomic location analysis to add edge constraints representing prior information. We conclude in Section 5 with a discussion of the results presented in this paper and offer some directions for future work.

2 Model induction

Methods for the induction of Bayesian network models from observational data generally fall into two classes: constructive methods based on the examination of conditional independence constraints that hold over the empirical probability distributions on the variables represented in the data, and search methods that seek to maximize some scoring function that describes the ability of the network to explain the observed data. We concentrate here exclusively on the latter although recent work[6] suggests that the two methods are

equivalent under reasonable assumptions. In a search context, the Bayesian scoring metric (BSM) is an especially common choice for the scoring function, although other choices can be made if the BSM is difficult to compute exactly.

We consider heuristic rather than exhaustive search strategies, since the identification of the highest-scoring model under the BSM for a given set of data is known to be NP-complete.[7] Commonly used local heuristic search algorithms include greedy hill-climbing, greedy random, Metropolis,[8] and simulated annealing; the last three are successive generalizations of one another. We have implemented these search algorithms and have observed in this particular context that simulated annealing consistently finds the highest scoring models among these algorithms. For reasons of limited space and simplified exposition, we therefore concentrate here only on the simulated annealing algorithm and results generated through its use.

The simulated annealing algorithm is so named because it operates in a manner analogous to the physical process of annealing. During the search process, the Metropolis algorithm is run as a subroutine at various temperatures T. The prevailing temperature and the score difference between graphs determine the transition probability within Metropolis, with higher temperatures indicating more permissive transitions. Initially, the temperature is set very high (allowing almost all changes to be made), but is gradually reduced according to some schedule until it reaches zero, at which point the Metropolis subroutine is equivalent to the greedy random algorithm. The schedule that the temperature is constrained to follow can be varied to produce different kinds of search algorithms. The schedule we employ allows for "reannealing" after the temperature becomes sufficiently low.

We extend our simulated annealing algorithm to search for models with constraints specifying which edges are required to be present and which are required to be absent. This allows for the incorporation of prior information about edges in the graph since this kind of constrained search algorithm is equivalent to an unconstrained search algorithm with a nonuniform prior over structures that gives zero weight to models that either include edges required to be absent or do not include edges required to be present. In this way, data from other sources (such as location data) can be easily incorporated.

We do not use our algorithm to isolate a single model because model selection tends to over-fit the data by selecting the single maximum a posteriori model and ignoring completely other models that score nearly as well. A more principled Bayesian approach is to compute probabilities of features of interest by averaging over the posterior model distribution. For example, if we are interested in determining whether the data D supports the inclusion

of an edge in graph S between two variables X and Y, we compute:

$$\mathrm{p}(E_{XY}|D) = \sum_S \mathrm{p}(E_{XY}|D,S) \cdot \mathrm{p}(S|D) \tag{1}$$

$$= \sum_S 1_{XY}(S) \cdot \mathrm{e}^{\mathsf{BSM}(S)} \tag{2}$$

where E_{XY} represents an edge from variable X to variable Y, $1_{XY}(S)$ is an indicator function that is 1 if and only if graph S includes E_{XY} as an edge, and $\mathsf{BSM}(S)$ is the Bayesian scoring metric for graph S. However, this sum is difficult to compute because the space of graphs S is enormous. Fortunately, it is possible to approximate this sum since the vast bulk of its mass lies among the highest scoring models.[a] For example, if we restrict our attention to the N highest scoring models, and index these by the variable i, we have:

$$\mathrm{p}(E_{XY}|D) \approx \frac{\sum_{i=1}^{N} 1_{XY}(S_i) \cdot \mathrm{e}^{\mathsf{BSM}(S_i)}}{\sum_{i=1}^{N} \mathrm{e}^{\mathsf{BSM}(S_i)}} \tag{3}$$

Using model averaging in this way reduces the risk of over-fitting the data by considering a multitude of models when computing the probabilities of features of interest.

3 Data preparation

3.1 Expression data

A set of 320 samples of unsynchronized *Saccharomyces cerevisiae* populations of varying genotype were observed under a diversity of experimental conditions. The set of samples ranges widely but consists primarily of observations of various wild-type and mutant strains made 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. Whole-genome expression data for each of these 320 observations was collected using four low-density 50-micron Affymetrix Ye6100 GeneChips per observation (roughly a quarter of the genome can be measured on each chip).

[a] The exponential factor in the sum has the effect of drowning out all but the highest scoring models, even though these highest scoring models are relatively infrequent.

The reported "average difference" values from these 1280 Affymetrix GeneChips were normalized using maximum a posteriori normalization methods based on exogenous spiked controls.[9] The output of this process was a 6135×320 matrix of normalized log expression values, one row for each gene in the yeast genome and one column for each experimental observation.

From the 6135 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 the mating response in yeast. Descriptions of the roles of the genes and proteins that were selected are presented in Table 1 and compiled from information from a variety of sources.[10,11]

Table 1. Descriptions of the 32 genes selected for model induction. The color mnemonics are used later in Figure 2: genes expressed only in MATa cells are magenta, genes expressed only in MATα cells are red, genes whose promoters are bound by Ste12 are blue, genes coding for components of the G-protein complex are green, genes coding for core components of the signaling cascade complex are yellow (except FUS3 which is already blue), genes coding for auxiliary components of the signaling cascade are orange, and genes coding for components of the SWI-SNF complex are aqua.

Gene	Color Mnemonic	Function of Corresponding Protein
STE2	magenta	transmembrane receptor peptide (present only in MATa strains)
STE3	red	transmembrane receptor peptide (present only in MATα strains)
GPA1	green	component of the heterotrimeric G-protein (Gα)
STE4	green	component of the heterotrimeric G-protein (Gβ)
STE18	green	component of the heterotrimeric G-protein (Gγ)
FUS3	blue	mitogen-activated protein kinase (MAPK)
STE7	yellow	MAPK kinase (MAPKK)
STE11	yellow	MAPKK kinase (MAPKKK)
STE5	yellow	scaffolding peptide holding together Fus3, Ste7, and Ste11 in a large complex
STE12	blue	transcriptional activator
KSS1	orange	alternative MAPK for pheromone response (in some dispute)
STE20	orange	p21-activated protein kinase (PAK)
STE50	orange	unknown function but necessary for proper function of Ste11
MFA1	magenta	a-factor mating pheromone (present only in MATa strains)
MFA2	magenta	a-factor mating pheromone (present only in MATa strains)
MFALPHA1	red	α-factor mating pheromone (present only in MATα strains)
MFALPHA2	red	α-factor mating pheromone (present only in MATα strains)
STE6	magenta	responsible for the export of a-factor from MATa cells (present only in MATa strains)
FAR1	blue	substrate of Fus3 that leads to G1 arrest; known to bind to STE4 as part of complex of proteins necessary for establishing cell polarity required for shmoo formation after mating signal has been received
FUS1	blue	required for cell fusion during mating
AGA1	blue	anchor subunit of a-agglutinin complex; mediates attachment of Aga2 to cell surface
AGA2	magenta	binding subunit of a-agglutinin complex; involved in cell-cell adhesion during mating by binding Sag1 (present only in MATa strains)
SAG1	red	binding subunit of α-agglutinin complex; involved in cell-cell adhesion during mating by binding Aga2 (present only in MATα strains; also known as Agα1)
BAR1	magenta	protease degrading α-factor (present only in MATa strains)
SST2		involved in desensitization to mating pheromone exposure
KAR3		essential for nuclear migration step of karyogamy
TEC1		transcriptional activator believed to bind cooperatively with Ste12 (more active during induction of filamentous or invasive growth response)
MCM1		transcription factor believed to bind cooperatively with Ste12 (more active during induction of pheromone response)
SIN3		implicated in induction or repression of numerous genes in pheromone response pathway
TUP1		implicated in repression of numerous genes in pheromone response pathway
SNF2	aqua	implicated in induction of numerous genes in pheromone response pathway (component of SWI-SNF global transcription activator complex)
SWI1	aqua	implicated in induction of numerous genes in pheromone response pathway (component of SWI-SNF global transcription activator complex)

The normalized levels of expression for these 32 genes were extracted from the 6135×320 normalization output matrix to yield a matrix of data with 32 rows and 320 columns, one row for each gene and one column for each observation. This data was then log-transformed and discretized using discretization level coalescence methods which incrementally reduce the number of discretization levels for each gene while preserving as much total mutual information between genes as possible.[2] 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.

In addition to the 32 variables representing levels of gene 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. The inclusion of this variable is necessary because, *e.g.*, the MFA1 and MFA2 genes responsible for producing the mating pheromone a-factor are expressed only in MATa strains of yeast. The data used as input for model induction was thus a matrix of 33 rows and 320 columns, 32 rows representing the discretized levels of log expression for 32 genes involved in pheromone response and one row representing the mating type of the strain in each experiment, either MATa or MATα.

3.2 Location data

Data from genomic location analysis, gathered using a chromatin immunoprecipitation assay, revealed the genes in the yeast genome whose upstream regions were bound by Ste12 under both presence and absence of pheromone. Of the 32 pheromone response genes in this paper, STE12, FUS1, FUS3, AGA1, and FAR1 promoters are all bound by Ste12, the first three being bound significantly both before and after the addition of pheromone, and the latter two being bound significantly only after the addition of pheromone. A description of the assay and a more detailed presentation of the results can be found in the paper by Ren, *et al.*[5]

4 Model averaging results

The implementation of our search algorithm is written in C and is capable of searching about 200,000-250,000 (not necessarily unique) models per minute on a 400MHz Pentium II Linux workstation. Although the code keeps a small hash-table of the scores of recently visited models, it is not especially optimized and could likely be sped up.

We used our search implementation 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 without constraints on the graph edges. In the second run, we incorporated the available location data by requiring edges from STE12 to FUS1, FUS3, AGA1, and FAR1. The top left and right histograms in Figure 1 show the distributions of scores for all models visited during the unconstrained and constrained simulated annealing runs, respectively. For comparison, the bottom histogram in Figure 1 shows the distribution of scores for all models visited when we perform a lengthy random walk through the space of models, accepting every proposed local change (equivalent to infinite-temperature Metropolis). From this figure, we see that the simulated annealing algorithm is quite effective in gradually concentrating its efforts on extremely high scoring models.

After gathering the five hundred highest scoring models that were visited during each run of the search algorithm, we computed the probability of edges being present by using the weighted average approximation shown in Equation 3 (with $N = 500$). Results of this computation for the unconstrained and constrained searches are presented in Tables 2 and 3, respectively. 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 of these that consists of all edges with estimated posterior probability over 0.5. These networks are shown in Figure 2. Graph nodes have been augmented with color information to indicate the different groups of variables with known relationships in the literature, as indicated in Table 1 and below. Graph edges have also been augmented with color information: 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 Figures 2, 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

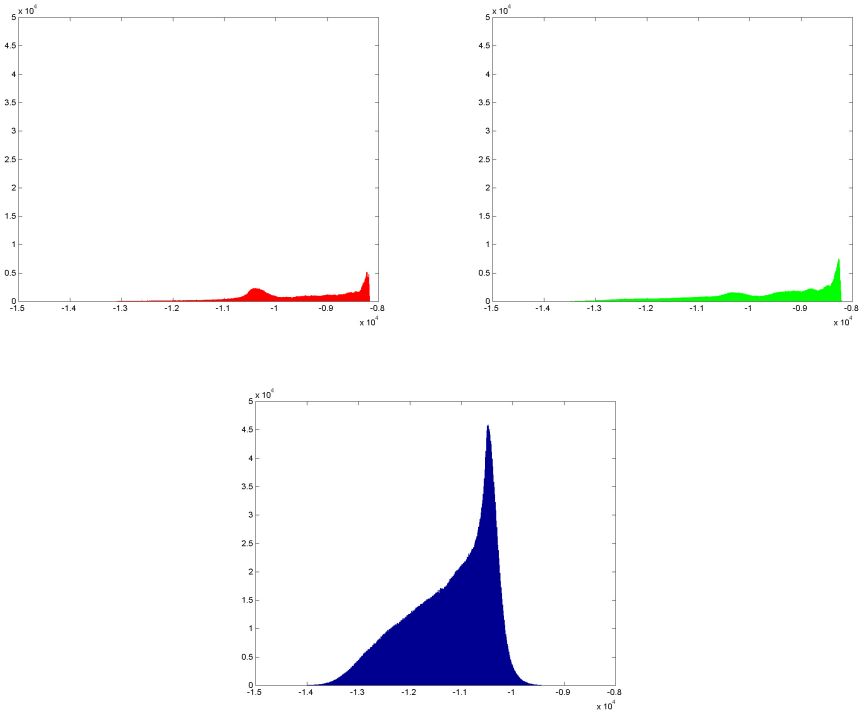

Figure 1. Histograms of scores for all models visited during simulated annealing runs (frequency versus log posterior probability of model). The top left and right histograms are for the unconstrained and constrained simulated annealing runs, respectively. For comparison, the bottom histogram was generated by a random walk through the space of models, accepting every proposed local change.

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 in the constrained graph but only GPA1 and STE18 are connected in the unconstrained graph. Indeed, even the link between GPA1 and STE4 in the constrained graph is fairly weak. On the other hand, SWI1 and SNF2 (aqua) are weakly adjacent in the unconstrained graph, but not adjacent in the constrained graph, though in both cases they are close descendants of TUP1. STE11 and STE5, two of

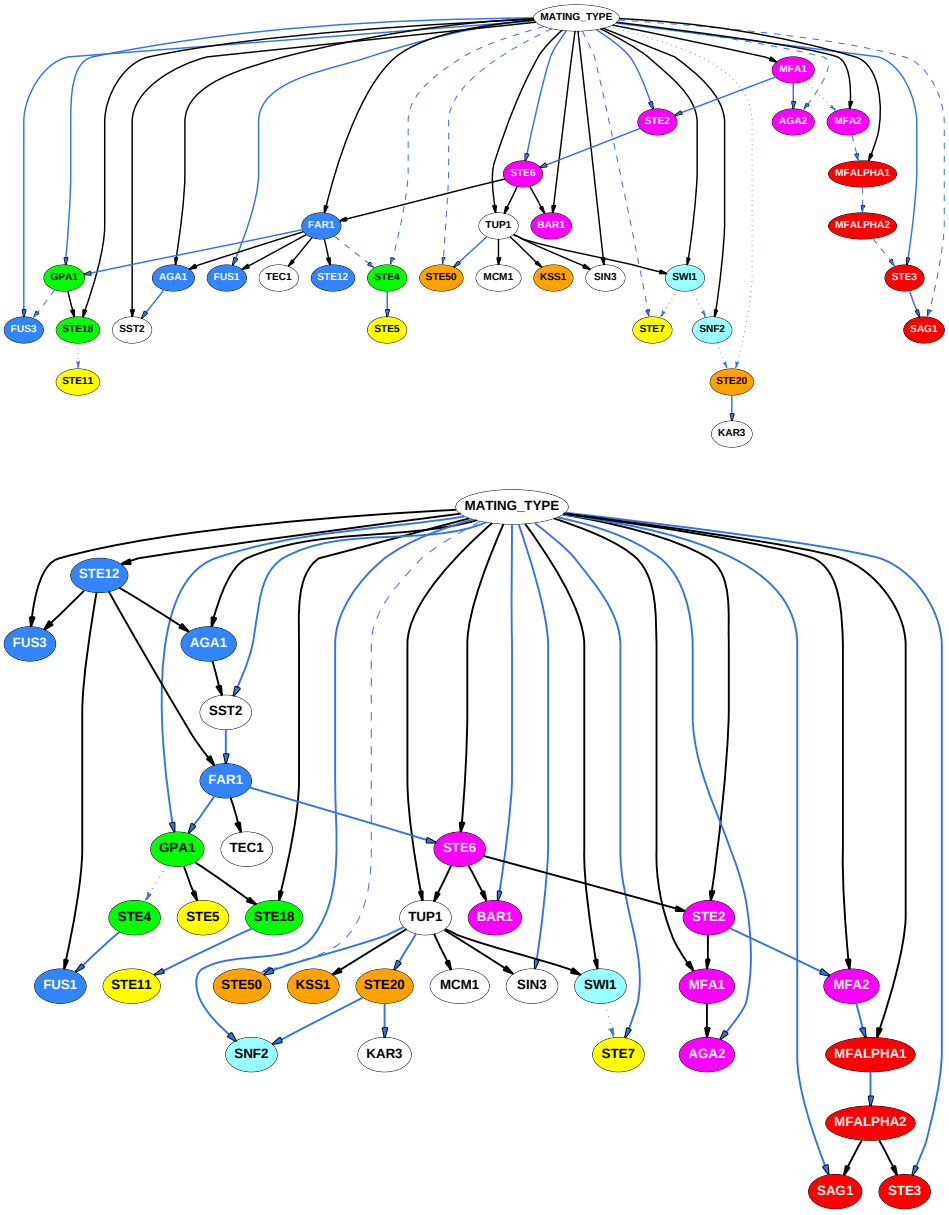

Figure 2. Bayesian network models learned by model averaging over the 500 highest scoring models visited during the unconstrained and constrained simulated annealing search runs, respectively. Edges are included in the figure if and only if their posterior probability exceeds 0.5. Node and edge color descriptions are included in the text.

Table 2. Posterior probabilities of edges being present in the unconstrained search as estimated by a weighted average over the 500 highest scoring models.

From	To	Posterior Probability	From	To	Posterior Probability
MATING_TYPE	MFA1	1.0000000	MATING_TYPE	STE4	0.9310940
STE6	TUP1	1.0000000	MATING_TYPE	SAG1	0.9191640
MATING_TYPE	TUP1	1.0000000	MATING_TYPE	STE50	0.8662340
FAR1	FUS1	1.0000000	MFA1	MFA2	0.6304400
TUP1	SWI1	1.0000000	SWI1	STE7	0.6292930
MATING_TYPE	SWI1	1.0000000	STE18	STE11	0.6159960
MATING_TYPE	MFALPHA1	1.0000000	SWI1	SNF2	0.5659030
MATING_TYPE	SNF2	1.0000000	MATING_TYPE	STE20	0.5551610
TUP1	SIN3	1.0000000	SNF2	STE20	0.5533950
MATING_TYPE	SIN3	1.0000000	STE20	SNF2	0.4340970
FAR1	AGA1	1.0000000	TUP1	STE20	0.4154630
MATING_TYPE	AGA1	1.0000000	STE50	STE11	0.3824810
MATING_TYPE	MFA2	1.0000000	STE2	MFA2	0.3695600
TUP1	MCM1	1.0000000	MATING_TYPE	STE12	0.2741810
MATING_TYPE	SST2	1.0000000	MATING_TYPE	KSS1	0.1964160
FAR1	TEC1	1.0000000	STE5	STE7	0.1710230
GPA1	STE18	1.0000000	STE50	STE7	0.1708470
MATING_TYPE	STE18	1.0000000	MATING_TYPE	MFALPHA2	0.0708669
STE6	FAR1	1.0000000	MFALPHA2	MFALPHA1	0.0596000
MATING_TYPE	FAR1	1.0000000	STE3	MFALPHA2	0.0596000
STE6	BAR1	1.0000000	GPA1	STE4	0.0594723
MATING_TYPE	BAR1	1.0000000	MFALPHA1	MFA1	0.0569844
FAR1	STE12	1.0000000	MFA1	FUS3	0.0339687
TUP1	KSS1	1.0000000	SWI1	STE20	0.0311423
MFA1	STE2	0.9998720	FUS3	MFA1	0.0255037
STE4	STE5	0.9998720	STE11	STE7	0.0090948
AGA1	SST2	0.9998720	MATING_TYPE	STE11	0.0017818
STE2	STE6	0.9998720	STE2	MFALPHA1	0.0016013
STE3	SAG1	0.9998720	SNF2	STE11	0.0015236
FAR1	GPA1	0.9998720	MFA1	MFALPHA1	0.0005666
MFA1	AGA2	0.9998720	AGA2	MFALPHA1	0.0003554
TUP1	STE50	0.9997690	STE11	KAR3	0.0002888
STE20	KAR3	0.9997110	STE6	STE2	0.0001277
MATING_TYPE	STE3	0.9995500	AGA2	MFA1	0.0001277
MATING_TYPE	GPA1	0.9994340	GPA1	STE5	0.0001277
MATING_TYPE	STE2	0.9979080	FAR1	SST2	0.0001277
MATING_TYPE	FUS3	0.9965830	SAG1	STE3	0.0001277
MATING_TYPE	STE6	0.9955280	SST2	SAG1	0.0001277
MATING_TYPE	FUS1	0.9937100	STE4	GPA1	0.0001277
MATING_TYPE	AGA2	0.9850980	STE2	AGA2	0.0001277
MATING_TYPE	STE7	0.9753010	AGA2	FUS3	0.0001277
FAR1	STE4	0.9405280	STE20	STE50	0.0001235
MFALPHA2	STE3	0.9404000	STE12	FUS3	0.0000620
MFALPHA1	MFALPHA2	0.9404000	STE18	STE50	0.0000619
GPA1	FUS3	0.9404000	MATING_TYPE	TEC1	0.0000519
MFA2	MFALPHA1	0.9349580			

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. 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

Table 3. Posterior probabilities of edges being present in the constrained search as estimated by a weighted average over the 500 highest scoring models. As the four edges required by location analysis appear in all visited graphs, their posterior probability is 1 by definition.

From	To	Posterior Probability	From	To	Posterior Probability
STE6	STE2	1.0000000	MATING_TYPE	GPA1	0.9979470
MATING_TYPE	STE2	1.0000000	MATING_TYPE	STE7	0.9959610
STE2	MFA1	1.0000000	SST2	FAR1	0.9956230
MATING_TYPE	MFA1	1.0000000	MATING_TYPE	AGA2	0.9915840
GPA1	STE5	1.0000000	MFA2	MFALPHA1	0.9911710
STE6	TUP1	1.0000000	MATING_TYPE	STE50	0.8627090
MATING_TYPE	TUP1	1.0000000	GPA1	STE4	0.7190930
STE12	FUS1	1.0000000	SWI1	STE7	0.6980530
TUP1	SWI1	1.0000000	MFA1	FUS3	0.3288450
MATING_TYPE	SWI1	1.0000000	FAR1	STE4	0.2809070
MATING_TYPE	MFALPHA1	1.0000000	MATING_TYPE	STE4	0.2809060
TUP1	SIN3	1.0000000	MATING_TYPE	KSS1	0.1904580
STE12	AGA1	1.0000000	STE50	STE7	0.1808790
MATING_TYPE	AGA1	1.0000000	AGA2	FUS3	0.1517050
MATING_TYPE	MFA2	1.0000000	STE5	STE7	0.0452417
TUP1	MCM1	1.0000000	STE11	STE7	0.0114721
AGA1	SST2	1.0000000	MATING_TYPE	MFALPHA2	0.0081962
MFALPHA2	STE3	1.0000000	MFA1	MFALPHA1	0.0044375
MATING_TYPE	STE6	1.0000000	MATING_TYPE	FAR1	0.0043766
FAR1	TEC1	1.0000000	STE2	MFALPHA1	0.0024661
GPA1	STE18	1.0000000	MATING_TYPE	STE5	0.0008114
MATING_TYPE	STE18	1.0000000	STE18	STE50	0.0004177
MFALPHA2	SAG1	1.0000000	AGA2	MFALPHA1	0.0003914
STE12	FAR1	1.0000000	MATING_TYPE	KAR3	0.0003614
STE6	BAR1	1.0000000	STE20	STE50	0.0003008
MATING_TYPE	STE12	1.0000000	SWI1	SNF2	0.0002817
TUP1	KSS1	1.0000000	SNF2	STE20	0.0002817
MFA1	AGA2	1.0000000	MATING_TYPE	STE20	0.0001019
STE12	FUS3	1.0000000	STE11	KAR3	0.0000791
MATING_TYPE	FUS3	1.0000000	STE6	MFALPHA1	0.0000523
MATING_TYPE	SNF2	0.9999970	STE4	GPA1	0.0000420
FAR1	STE6	0.9999950	MFA1	MFA2	0.0000371
MATING_TYPE	BAR1	0.9999950	STE7	STE50	0.0000239
MFALPHA1	MFALPHA2	0.9999950	MATING_TYPE	STE11	0.0000221
STE4	FUS1	0.9999860	MATING_TYPE	TEC1	0.0000164
MATING_TYPE	SIN3	0.9999840	STE50	STE11	0.0000148
STE18	STE11	0.9999790	FAR1	FUS1	0.0000138
STE2	MFA2	0.9999630	SNF2	STE7	0.0000085
FAR1	GPA1	0.9999580	MFALPHA2	MFALPHA1	0.0000050
MATING_TYPE	STE3	0.9998550	GPA1	STE6	0.0000050
STE20	SNF2	0.9997180	MFA2	MFALPHA2	0.0000050
TUP1	STE20	0.9997180	STE18	KAR3	0.0000038
STE20	KAR3	0.9995980	STE11	STE50	0.0000024
MATING_TYPE	SAG1	0.9994460	SNF2	STE11	0.0000020
TUP1	STE50	0.9983870	MFALPHA2	STE50	0.0000018
MATING_TYPE	SST2	0.9979600			

are known to interact in the cell[12] 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.[13] 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 co-expression.

Though it is produced at higher levels in MATa cells, it is known that Aga1 is produced in both MATa and MATα cells.[14] The graphs 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 graphs, 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.[15]

5 Discussion

When we interpret automatically generated Bayesian networks, it should be remembered that edges indicate a statistical dependence between the transcript levels of genes, but do not necessarily specify the form or presence of a physical dependence. For example, a variable X can seem to be influencing a variable Z if a critical intermediating variable Y remains unmodeled. As another example, in both networks in Figure 2, a link appears between MFA2 and MFALPHA1. even though these mating factors are never both expressed in haploid *S. cerevisiae* strains. However, cells expressing one are less likely, statistically, to be expressing the other; hence the link. The weakness of the link indicates that other variables such as `mating_type` are frequently successful in explaining away this statistical dependence. In general, multiple biological mechanisms may map to the same set of statistical dependencies and thus be hard to distinguish on the basis of statistical tests alone. Moreover, if there is not sufficient data to observe a system in a numberof different configurations, we may not be able to uncover certain dependencies at all.

The composite network resulting from unconstrained search based only on genomic expression data has a few apparent limitations. Most strikingly, the search method is unable from expression data alone to learn the correct regulatory relationships between Ste12 and its targets. By fusing expression data with location data, the constrained search is able to consider statistical dependencies in the expression data that are consistent with the physical relationships already identified in the location data. In this way, location data proves to be quite complementary to expression data: since it can help identify network edges directly, location data dramatically decreases the amount of expression data needed to discover regulatory networks by statistical methods.

When genomic location data suggests that particular edges should be present, our algorithms currently modify the model prior so that graphs lacking these suggested edges have zero weight. However, we know that location data can be noisy. We can relax our assumption of zero weight, and instead modify the model prior so that graphs lacking these suggested edges have small but positive weight. This is permissible within our framework but adds the extra complication that the relative weight of models lacking suggested edges needs to be specified (presumably based on the degree of confidence in

the location data). Values for this weight that are too high or low lead to the under- or over-inclusion of suggested edges, respectively.

Additionally, it is possible that a protein may bind DNA but have no impact on downstream gene regulation. Location data provides information about physical relationships while expression data provides information about statistical relationships; the two are not guaranteed to agree.

There remain a number of ways to extend this work in the future. Among these are the use of search algorithms that more frequently visit high scoring regions of the model search space, incorporation of data from other sources besides expression and location data, leveraging time-series data and dynamic Bayesian networks to model feedback processes, leveraging interventional data to uncover causal processes, and adding the ability to discover annotated network edges refining the type of relationship learned between model variables.[1]

Acknowledgments

The authors wish to thank Tomi Silander for access to B-Course source code and anonymous reviewers for their helpful comments. In addition, Hartemink gratefully acknowledges support through the Merck/MIT Graduate Fellowship in Informatics.

References

1. A. J. Hartemink, *et al.* In *Pac. Symp. Biocomp.*, 6:422–433, 2001.
2. A. J. Hartemink. *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks.* PhD thesis, MIT, 2001.
3. N. Friedman, *et al.* In *RECOMB 2000.* ACM-SIGACT, Apr. 2000.
4. D. Pe'er, *et al.* In *ISMB 2001.* ISCB, Jul. 2001.
5. B. Ren, *et al. Science*, 290(5500):2306–2309, Dec. 2000.
6. R. G. Cowell. In *UAI 2001.* Morgan Kauffman, Jul. 2001.
7. D. M. Chickering. In D. Fisher and H.-J. Lenz, eds, *Learning from Data: AI and Statistics V*, chap. 12, 121–130. Springer-Verlag, 1996.
8. N. Metropolis, *et al. J. Chem. Phys.*, 21:1087–1091, 1953.
9. A. J. Hartemink, *et al.* In *BiOS 2001*, 132–140. SPIE, Jan. 2001.
10. E. A. Elion. *Curr. Opin. Microbiology*, 3(6):573–581, Dec. 2000.
11. M. Costanzo, *et al. Nuc. Acids Res.*, 29(1):75–79, 2001.
12. I. Gavin, M. Kladde, and R. Simpson. *Embo J.*, 19:5875–5883, 2000.
13. R. J. Cho, *et al. Mol. Cell*, 2:65–73, Jul. 1998.
14. A. Roy, *et al. Mol. Cell. Biol.*, 11(8):4196–4206, Aug. 1991.
15. P. T. Spellman, *et al. Mol. Biol. Cell*, 9:3273–3297, 1998.