INFORMATIVE STRUCTURE PRIORS: JOINT LEARNING OF DYNAMIC REGULATORY NETWORKS FROM MULTIPLE TYPES OF DATA *SUPPLEMENTAL SECTION*

ALLISTER BERNARD, ALEXANDER J. HARTEMINK
Duke University, Dept. of Computer Science, Box 90129, Durham, NC 27708

1 Bayesian marginalization over parameter λ

Bayesian marginalization leads to an edge probability distribution that tapers more gradually than would have been the case without using marginalization as depicted in Figure 1 (this distribution has thicker tails than the exponential distribution).

2 Results using simulated data

We use simulated data from a synthetic cell cycle model to evaluate the accuracy of our algorithm and determine the relative utility of different quantities of available gene expression data. The synthetic cell cycle is quite complex involving 100 genes operating in three phases of the cycle as shown in Figure 2. This synthetic cell cycle consists of cell cycle transcription factors with/without location data, non cell cycle transcription factors with location data, activated/repressed genes as well as genes not involved at all in the cell cycle process. In all the network has 54 true positives and 9846 true negatives that can be learnt. Scalability of learning a DBN with uninformative priors has been previously examined[1] and so we do not examine issues of scale as the computationally efficient prior ensures our method is no different from

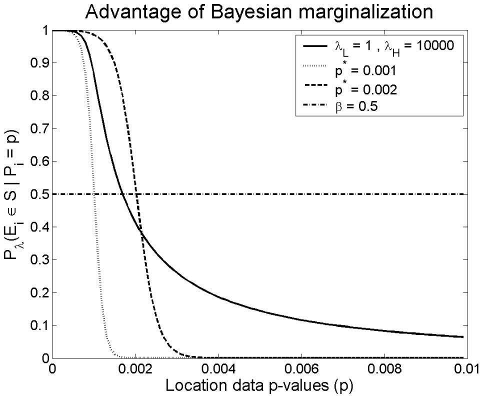

Figure 1. The effect of Bayesian marginalization leads to a thicker tailed distribution.

regular DBN learning.

2.1 Generating synthetic expression data

The simulated gene expression data is generated using the (stochastic) Boolean Glass gene model described elsewhere.[2] In the *Boolean Glass gene* model,[2] gene expression values are governed by the following equation

$$G_i^t = G_i^{t-1} + rate_1 * (-G_i^{t-1}) + rate_2 * [F_i(parents(G_i^t) \backslash G_i^{t-1})] + \epsilon \qquad (1)$$

where G_i^t and G_i^{t-1} are the expression values of gene G_i at times t and $t-1$ respectively, $0 < rate_1, rate_2 \leq 1$ and ϵ is an error term drawn from a normal distribution, F_i is a Boolean function defined over the parents of gene G_i at time t. All genes were initialized by sampling from a normal $\mathcal{N}(0, 0.5)$ distribution, with the exception of transcription factors active in the first phase. These were given higher initial expression values by sampling from a normal $\mathcal{N}(3, 0.5)$ distribution. After initialization, all variables were updated according to the Boolean Glass gene equations corresponding to the first phase for 30 time points. Then the update equations were changed to reflect the structure of the second phase, as shown in Figure 2, and another 30 time points were generated. The update equations changed once more for the third phase, and another 30 time points were generated. So each simulated cell cycle had a total of 90 time points in 3 phases, during each of which a different set of update equations prevailed. This process was repeated to generate more expression data as needed; we experimented with various amounts of data, from 1 to 15 cell cycles worth.

2.2 Generating synthetic location data

To generate synthetic location data, we need to generate synthetic p-values for the subset of genes that are transcription factors in the model. For the edges from each of the transcription factors that do not exist in the regulatory network in Figure 2 (i.e., the non-targets of each transcription factor), we generate p-values from a uniform distribution $\mathcal{U}[0, 1]$, since this is the definition of a p-value. For edges that exist (i.e., the targets of each transcription factor), we use an exponental distribution with scale parameter λ over the interval $[0, 1]$. The greater the value of λ, the more reliable is the location data generated in the sense that the p-values for true edges are more likely be close to zero. In our experiments we chose $\lambda \in \{1, 10, 20, 100\}$, resulting in 4 *complete* location data sets and for ease of understanding we named these sets *extremely nosiy location data, moderately noisy location data, fair location data* and *excellent location data* respectively.

2.3 Experimental validation

We conduct the following three experiments: score network structures with expression data alone, ignoring the log prior component $\log \mathcal{P}(S)$; score network structures with location data alone, using the prior component of the score and ignoring the log likelihood component $\log \mathcal{P}(D|S)$; and score network structures with both expression and location data. Figure 4 summarizes the results. The vertical axis measures the total number of errors: the sum of false positives and false negatives in the learned network; the total number of errors relative to the synthetic network can range from 0 to 10000. As expected, the total number of errors drops sharply as the amount of available expression data increases. Together, the curves show 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 achieve the same error rate as would be achieved with expression data in isolation, suggesting that the availability of location can be used to compensate for small quantities of expression data. Figure 3 compares the effect of learning the network structure when using location data having different noise characteristics. Observe that for less noisy location data, joint learning with limited expression data can result in worse results as compared to using no expression data. Finally in Figure 5 we show that by varying the prior probability β we can control the sensitivity and specificity of the learned network structure.

3 Results using experimental data

3.1 Assigning phase labels

We used publicly available cell cycle gene expression data[3]. The gene expression data consists of 69 time points collected over 8 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. To generate a phase label for each time point, we select characteristic genes known to be regulated during specific cell cycle regulatory regions namely M/G_1,G_1,S,S/G_2 and G_2/M.[3] Guided by the expression of these characteristic genes, we can assign a phase label to each time point. This is done separately for each of the four synchronization protocols in the dataset (alpha, cdc15, cdc28, and elu). Table 1 shows the phase label assignments for each synchronization protocol.

3.2 Variable selection and analysis

Table 2 shows the list of cell cycle transcription factors used. These transcription factors have location data available for them. Table 3 shows the list of regulated genes used for our experimental analysis. In all we used a total of 25 genes including the cell cycle transcription factors. Before we ran our method on the data for these

Table 1. Phase label assignments

Synchronization Protocol	**Timepoints and their corresponding phase labels**
alpha	G_1 : alpha7, alpha14, alpha21, alpha70, alpha77, alpha84
	$S + G_2$: alpha28, alpha35, alpha42, alpha 49, alpha84, alpha91
	$G_2 + M$: alpha0, alpha56, alpha63, alpha105, alpha112, alpha115
cdc15	G_1 : cdc15_30, cdc15_50, cdc15_130, cdc15_140, cdc15_150, cdc15_230, cdc15_270
	$S + G_2$: cdc15_70, cdc15_80, cdc15_160, cdc15_170, cdc15_180, cdc15_190, cdc15_200, cdc15_290
	$G_2 + M$: cdc15_10, cdc15_90, cdc15_100, cdc15_110, cdc15_120, cdc15_210, cdc15_220, cdc15_230, cdc15_240
cdc28	G_1 : cdc28_10, cdc28_20, cdc28_100, cdc28_110, cdc28_120
	$S + G_2$: cdc28_30, cdc28_40, cdc28_50, cdc28_60, cdc28_130, cdc28_140
	$G_2 + M$: cdc28_0, cdc28_70, cdc28_80, cdc28_90, cdc28_150, cdc28_160
elu	G_1 : elu30, elu60, elu90, elu120
	$S + G_2$: elu150, elu180, elu210, elu240, elu270
	$G_2 + M$: elu0, elu300, elu330, elu360, elu390

25 selected genes the following steps were carried out

- Precompute the marginalization integral for the p-values of the selected location data
- Discretize the selected expression data into three states using interval discretization

Using simulated annealing as our heuristic search method we then identified network structures with high scores. Figure 7 depicts the complete regulatory network obtained after running our algorithm on the expression and location data for the 25 selected genes.

References

1. J. Yu, A. Smith, P. Wang, A. Hartemink, and E. Jarvis. *Bioinformatics*, page to appear, 2004.
2. R. Edwards and L. Glass. *Chaos*, 10:691–704, Sep. 2000.
3. P. T. Spellman et al. *Mol. Biol. Cell*, 9:3273–3297, 1998.

Table 2. List of cell cycle transcription factors with location data

Transcription Factor – Standard Name	**GO Biological Process**
ACE2	G1-specific transcription in mitotic cell cycle
ASH1	Pseudohyphal growth, regulation of transcription, mating-type specific
FKH1	chromatin silencing at silent mating-type cassette, pseudohyphal growth , regulation of cell cycle
MBP1	DNA replication , regulation of cell cycle
MCM1	DNA replication initiation , regulation of transcription from Pol II promoter
NDD1	G2/M-specific transcription in mitotic cell cycle
STB1	G1/S transition of mitotic cell cycle
SWI4	G1/S transition of mitotic cell cycle, cell cycle, transcription
SWI5	G1-specific transcription in mitotic cell cycle
SWI6	G1/S-specific transcription in mitotic cell cycle, meiosis, transcription

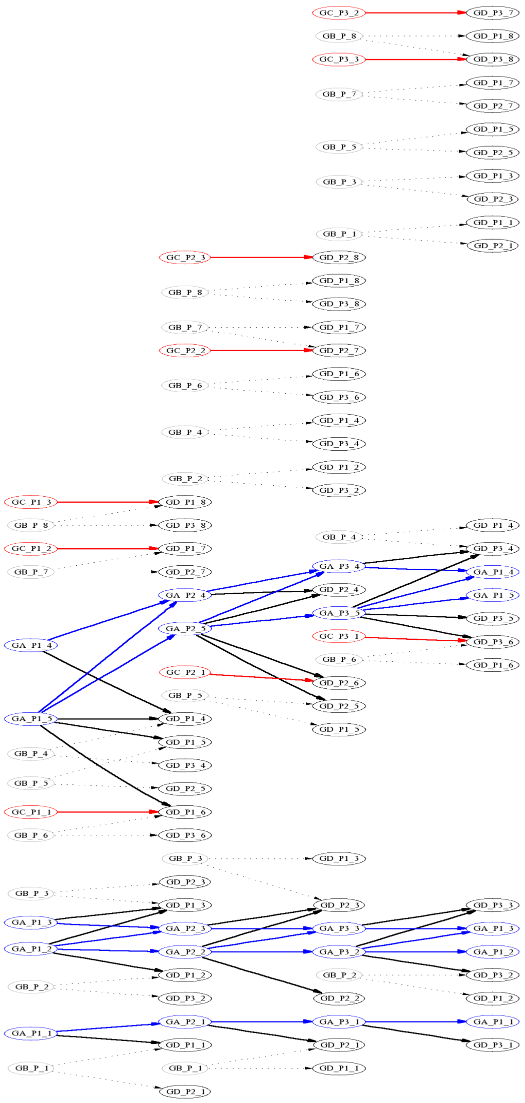

Figure 2. The figure shows the true structure of the simulated network of 100 nodes. Each column represents a particular instance in time. Each column should have 100 nodes but for clarity we only depict nodes in a column if they have an outdegree greater than zero. Blue nodes are cell cycle transcription factors for which location data is available, red nodes are cell cycle transciption factors for which no location data is available, grey nodes are non cell cycle transcription factors for which location data is available. All black nodes are either activated or repressed by one or more of these transcription factors. Blue edges represent activation of cell cycle transcription factors by other transcription factors and for which we have location data. Black edges represent activation/repression of black nodes by cell cycle transcription factors and for which we have location data. Red edges represent activation/repression of black nodes by cell cycle transcription factors and for which we have no location data. Dotted edges represent activation/repression of black nodes by non cell cycle transcription factors and for which we have location data.

Table 3. List of regulated genes

Gene – Standard Name	**GO Biological Process**
ALG7	N-linked glycosylation
CDC20	cyclin catabolism, mitotic metaphase/anaphase transition, mitotic sister chromatid segregation, mitotic spindle elongation, ubiquitin-dependent protein catabolism
CDC21	DNA-dependent DNA replication, dTMP biosynthesis
CDC5	DNA-dependent DNA replication, protein amino acid phosphorylation
CDC6	pre-replicative complex formation and maintenance
CLB2	G2/M transition of mitotic cell cycle , regulation of cyclin dependent protein kinase activity
CLB5	G1/S transition of mitotic cell cycle, G2/M transition of mitotic cell cycle, premeiotic DNA synthesis, regulation of cyclin dependent protein kinase activity
CLN1	regulation of cyclin dependent protein kinase activity
CLN2	re-entry into mitotic cell cycle after pheromone arrest, regulation of cyclin dependent protein kinase activity
CTS1	cytokinesis, completion of separation
EGT2	cytokinesis
FAR1	cell cycle arrest, signal transduction during conjugation with cellular fusion
HTA1	chromatin assembly or disassembly
PCL2	cell cycle
SIC1	G1/S transition of mitotic cell cycle, regulation of cyclin dependent protein kinase activity

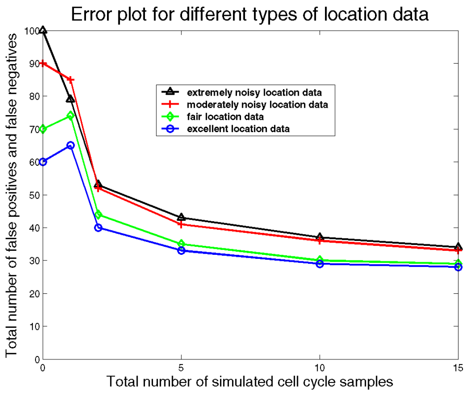

Figure 3. Total number of errors while learning a synthetic cell cycle network using (noisy simulated) expression and location data together. The graph shows the effect of using location data having different noise characteristics.

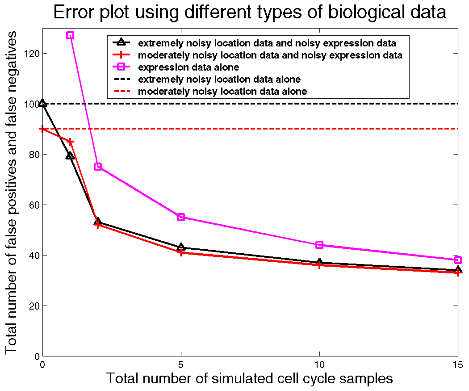

Figure 4. Total number of errors while learning a synthetic cell cycle network using (noisy simulated) expression and location data, separately and with both types of data together. The graph shows the effect of increasing the number of cell cycles worth of expression data, both with and without location data. The dashed horizontal line represents learning using location data alone.

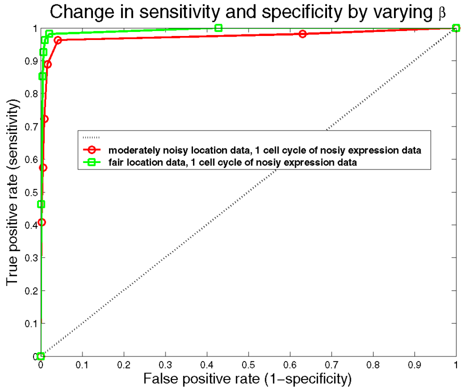

Figure 5. The graph shows how varying β effects the sensitivity and specificity of the learned network structure

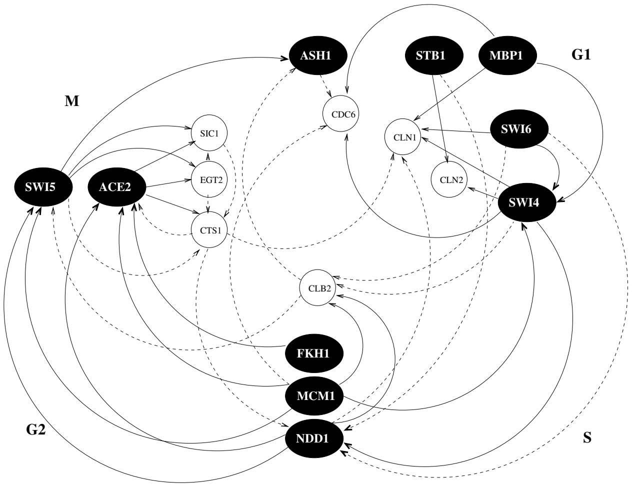

Figure 6. Partial regulatory network recovered using expression data from Spellman et al. and location data from Lee et al. Shaded elliptical nodes are transcription factors for which location data is available. Unshaded circular nodes are genes for which no location data is available. Solid edges represent interactions that have been verified in the literature. Dashed edges represent interactions that have not been verified; either the edge is incorrect or the evidence from the literature is inconclusive. Observe the cyclic regulation of transcription factors across phases of the cell cycle.

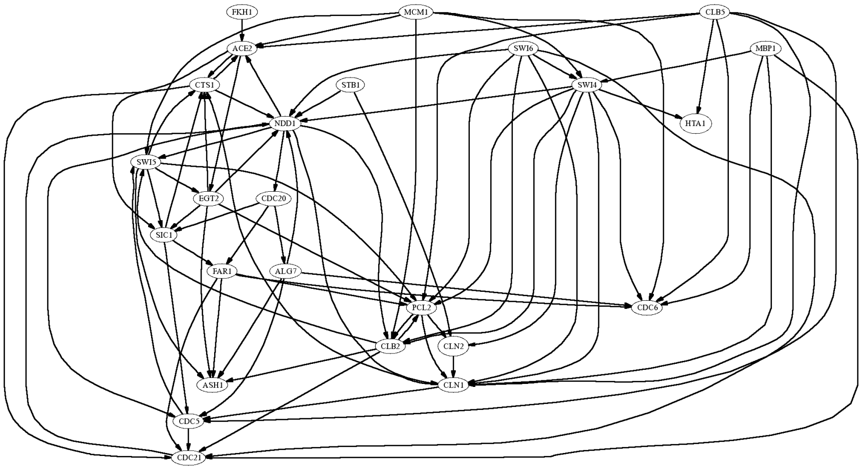

Figure 7. Complete regulatory network recovered using expression data from Spellman et al. and location data from Lee et al.