INFERRING TRANSCRIPTIONAL REGULATION THROUGH LOGICAL NETWORKS FROM TEMPORAL MOUSE BRAIN GENE EXPRESSION DATA

Mingzhou (Joe) Song, [*] **Chris K. Lewis,** [*] **Eric R. Lance,** [*]
Elissa J. Chesler, [**] **Roumyana Kirova,** [**]
Michael A. Langston, [***] **and Susan E. Bergeson** [****]

[*] *New Mexico State University, Las Cruces, NM 88003, USA*
[**] *Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[***] *University of Tennessee, Knoxville, TN 37996, USA*
[****] *University of Texas, Austin, TX 78712, USA*

Abstract: The problem of computing logical network models to account for temporal dependencies among interacting genes and environmental stimuli from high-throughput transcriptomic data is addressed. A logical network reconstruction algorithm was developed that uses the statistical significance as a criterion for network selection to avoid false interactions arising from pure chance. Using temporal gene expression data collected from the brains of alcohol-treated mice in an analysis of the molecular response to alcohol, this algorithm identified several genes from a major neuronal pathway as putative components of the alcohol response mechanism. Three of these genes have known specific associations with alcohol response as reported in the literature. Several other potentially relevant genes were also highlighted, in agreement with independent results from literature mining. These genes may play a role in the response to alcohol. Additional, previously-unknown interactions were discovered in the logical network that, subject to biological verification, may offer new clues in the search for the elusive molecular mechanisms of alcoholism.

Keywords: Logical networks, Transcriptional regulation, Alcohol response.

1. INTRODUCTION

The regulation of transcription occurs in an intriguingly complex system involving multiple interacting regulatory processes. Modeling transcriptional regulation requires algorithms that retain information about regulatory interactions. We reconstruct logical networks (LNs) from trajectories of discrete random variables, in order to uncover temporal dependencies among genes and environmental stimuli. The temporal interactions serve as a model for transcriptional gene regulatory networks (GRNs). We have developed an algorithm and evaluated it theoretically, quantitatively, and empirically, using transcriptome profiles from the brains of alcohol-treated mice, providing tremendous opportunities and insight to the diverse underlying molecular mechanisms in alcoholism. We inspect through the LN model temporal dependencies among key genes in response to alcohol in mice.

Temporal dependency is a function of causal interactions among processes in a regulatory network. Only when the function is nearly linear, does a temporal dependency substantially reflect a causal interaction. When linearity does not hold, a causal interaction may not be observed as a strong temporal dependency. System modeling may be further complicated by incomplete observations as are typical in biological data, for example, missing protein concentrations and small molecular messengers in a regulatory network. However, consistent temporal dependency must arise from a causal interaction, even with incomplete ob-

servations. Therefore, statistically significant temporal dependencies among genes and environmental stimuli constitute a foundation to establish causalities.

An LN is a system model to characterize interactions among discrete variables over discrete time. It is a directed graph, with each node in the graph being a discrete variable plus a logical function. The logical function for a node X can be described by a generalized truth-table, mapping all possible combinations of parent node values to values of X. The Boolean network (BN) (Liang et al., 1998; Akutsu et al., 2003; Pal et al., 2005; Klamt et al., 2006), where each variable takes the value of either 0 or 1, is a special case of an LN. An LN can be reconstructed from observed trajectories of a system under perturbed conditions. There are two important issues: One is how to compute efficiently the best among feasible LN candidates; the other one is how to determine the probability that the best candidate arises out of the randomness caused by noise and sampling errors in a network where no nodes interact. The first issue depends on how one handles the combinatorial computational cost, often *NP*-complete, incurred by reconstructing an LN. The second issue has been gaining attention such as in BN fitting (Kim et al., 2007). By computing the statistical significance of an LN based on multinomial tests at each node, we are able to resolve both issues in one step. We also established an approach to calculate the statistical power for LN modeling given number of time points and replicas per time point in an experiment, which can guide the design of time-course experiments.

Although other modeling methodologies have been developed, discrete dynamic system models including LNs and BNs are advantageous given the increased availability of experimental designs that collect temporal gene expression data at the whole genome scale. A BN represents, however, each gene expression level in two states: on and off, thereby limiting its capacity in discriminating quantitative changes in gene expression levels under gradually perturbed situations. The LN extension enables 1) finer description for the dynamics of genes, and 2) richer interaction patterns among genes. More importantly, we will present a one-step determination of statistical significance for an LN based on multinomial hypothesis testing, which addresses the goodness-of-fit and over-fitting issues tackled as separate steps in other approaches such as the application of coefficient of determination with user specified network complexity (Shmulevich et al., 2002). Bayesian networks and dynamic Bayesian networks (DBNs) have been used for GRN modeling, see (Imoto et al., 2003; Friedman, 2004) and (Ong et al., 2002), respectively. A DBN describes statistical dependencies among genes temporally, by incorporating time transitions between Bayesian networks at consecutive time points. Since a DBN does not explicitly describe functional relations among genes, it is not a direct tool to understand GRN dynamics, though certain DBNs can indeed be converted to probabilistic BNs (Lähdesmäki et al., 2006).

We applied the LN to study the alcohol influence on gene expression in mouse brains. The effects of alcohol on functions of gene products and the corresponding effect on gene expression is an active research area, particularly in the inflammatory and neural plasticity processes that result in lasting brain changes in response to alcohol. We believe the LNs resulting from our work will provide highly relevant clues to discover biologically important gene interactions involved in the molecular mechanisms of alcoholism.

2. DEFINITION OF LOGICAL NETWORK

An LN is a directed graph with a logical function attached to each node, a discrete variable that changes over discrete time. Each logical function for a node X can be described by a generalized truth-table that maps all possible combinations of parent node values to values of X. Let X have Q quantization levels and K parents $\pi_1, \pi_2, \ldots, \pi_K$ of $Q_1, Q_2, \ldots, Q_K$ quantization levels, respectively. The *complexity* of an LN is the maximum number of incoming edges a node can have. A *synchronous* LN updates the values of all the nodes simultaneously. An LN is a dynamic modeling tool because its current state can be computed from the previous state using the logical functions for each node. An LN is *first-order* if the value of each node at time t only involves the parent values at time $t-1$. Our discussion is restricted to synchronous first-order LNs. An LN allows richer interaction patterns than a BN. All types of pairwise interactions in a BN are illustrated in Fig. 1. A nonlinear pairwise interaction

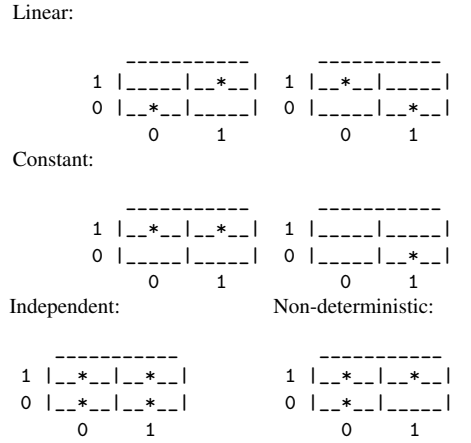

Fig. 1. All types of pairwise interactions in a BN.

between two logical variables, shown in Fig. 2, is impossible with a BN.

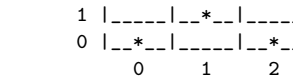

Fig. 2. A nonlinear interaction pattern in an LN.

3. STATISTICAL POWER FOR DETECTING A LOGICAL NETWORK

Given an observed trajectory of an LN and the sample size per time point, one is statistically limited in detecting an LN up to a certain complexity. The probability of accepting an LN some of whose nodes are indeed interacting given the observed trajectory is called the *statistical power*, which is determined by the logical functions, the distribution of each variable, sample size, Type I error, and the effect strength. The statistical power is independent of the inference approach used to recover an LN from the trajectory. With statistical power, one can answer the question whether the amount of data in the trajectory can statistically support any LN for certain complexity at all. Thus, statistical power is an important tool in experimental design to determine appropriate sample sizes.

We assume that each entry in a truth-table has a binomial distribution. The alternative hypothesis is that the binomial distribution has a success rate parameter of $p_a = 0.8$, versus the null hypothesis that the binomial distribution has $p_n = 0.5$. In other words, if the chances are equal for the entry in the truth-table for success or failure, there is no evidence for this truth-table to reflect any interaction. The test is two-sided. The Type I error rate $\alpha = 0.05$ is adjusted to α' considering multiple testing effect. Let n_- and n_+ be the decision boundary: if $n < n_-$ or $n > n_+$, reject null hypothesis, or equivalently the rejection region is $(0, n_-)$ and (n_+, N), where N is the total number of trials. The decision boundaries n_- and n_+ are determined such that $\sum_{n=0}^{n_-} B(N,n,p_n) + \sum_{n=n_+}^{N} B(N,n,p_n) = \alpha'$ and $B(N,n_-,p_n) = B(N,n_+,p_n)$, where the binomial distribution is defined as $B(N,n,p) = \binom{N}{n} p^n (1-p)^{N-n}$. The statistical power is $\sum_{n=0}^{n_-} B(N,n,p_a) + \sum_{n=n_+}^{N} B(N,n,p_a)$. Figure 3 plots the maximal power as a function for the complexity of an LN given different lengths of trajectories. The curve demonstrates that the more complex the network is, the lower the statistical power, given the same experimental factors. A (maximal) 68% power is possible if we use 5 time points for each condition with 7 replicas at each time point with a network of 20 genes, a complexity of 6, and a Type I error rate of 0.05. For a typical statistical power cutoff of 60%, our microarray experiment in Section 5 was justified. The Type I error α adjustment may be conservative as dependency may exist among time points. Although the binomial distribution can be replaced with a multinomial one to calculate the statistical power for a generalized truth-table, this study establishes the minimal requirements.

4. LOGICAL NETWORK INFERENCE BY MULTINOMIAL TESTS

Various criterions for goodness-of-fit have been used in inference of an LN from observed trajectories.

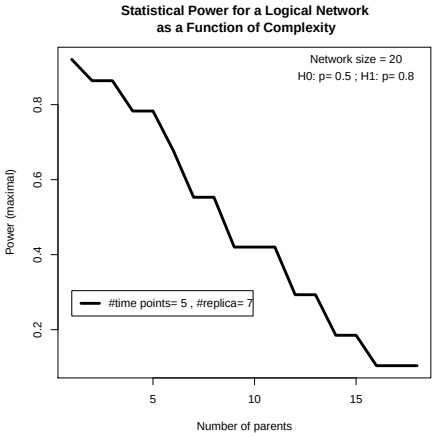

Fig. 3. Statistical power for detecting an LN as a function of its complexity, given number of time points (5), number of replicas per time point (7), network size (20) and hypotheses $p_a = 0.8$ (alternative) vs. $p_n = 0.5$ (null).

Mutual information among variables has been employed in interaction graphs (Margolin et al., 2006); likelihood is used to determine network structure for Bayesian networks (Friedman and Goldszmidt, 1996); coefficient of determination has been created for BNs (Shmulevich et al., 2002). These measures, however, do not penalize the complexity of a network and overfitting has to be resolved separately. Instead, we derived the statistical significance of an LN as a measure for both goodness-of-fit and model selection. It is derived from multinomial hypothesis testing. We move forward from existing approaches in LNs that often do not assess the probability of an LN arising by chance.

Table 1 shows the transition table of a single node X, which can also be considered a contingency table. The

Table 1. The transition table of node X.

row	$\pi_1[t-1]$ $Q_1=2$	… …	$\pi_K[t-1]$ $Q_K=3$	#0	$X[t]$ …	#$Q-1$
0	0	…	0	$n_{0,1}$	…	$n_{0,Q-1}$
1	0	…	1	$n_{1,1}$	…	$n_{1,Q-1}$
		⋮			⋮	
$R-1$	1	…	2	$n_{R-1,1}$	…	$n_{R-1,Q-1}$

number of rows in the table is $R = Q_1 Q_2 \cdots Q_K$. $n_{r,c}$ is the number of observations in which the parents take the values in the r-th row at $t-1$ and X takes the value of c at t. Let $n_{\cdot,c}$ be the summation of the c-th column. Let $n_{r,\cdot}$ be the summation of the r-th row. Let n be the total number of observations. The following hypotheses are designed for each row:

Null hypothesis: $n_{r,0} : n_{r,1} : \ldots : n_{r,Q-1} = n_{\cdot,0} : n_{\cdot,1} : \ldots : n_{\cdot,Q-1}$;

Alternative hypothesis: $n_{r,0} : n_{r,1} : \ldots : n_{r,Q-1} \neq n_{\cdot,0} : n_{\cdot,1} : \ldots : n_{\cdot,Q-1}$.

This hypothesis test determines if X depends on parent values in row r. It is in essence a multinomial test with the probability parameters $\frac{n_{\cdot,0}}{n}, \frac{n_{\cdot,1}}{n}, \ldots, \frac{n_{\cdot,Q-1}}{n}$. A multinomial test inspects the chi-square statistic

$\chi^2(r) = \sum_{c=0}^{Q-1} \frac{(n_{r,c}-\bar{n}_{r,c})^2}{\bar{n}_{r,c}}$, where $\bar{n}_{r,c} = \frac{n_{r,\cdot} n_{\cdot,c}}{n}$ is the expected count. Asymptotically $\chi^2(r)$ has a chi-square distribution with $Q-1$ degrees of freedom. $\chi^2(r)$ can be computed for each row r in the table. By the property of the chi-square distribution, summation of independent chi-squares is still chi-square whose degrees of freedom are the summation of each individual degrees of freedom. However, when we sum up all $\chi^2(r)$ over r, we loose $Q-1$ degrees of freedom because each column has a fixed total. Thus the transition table statistic $\chi^2 = \sum_{r=0}^{R-1} \chi^2(r)$ is chi-square distributed with $(R-1)(Q-1)$ degrees of freedom. Let χ_i^2 with degrees of freedom ν_i be the statistic for the transition table of the i-th node. We define the test statistic for an LN with N nodes as $\chi_{LN}^2 = \sum_{i=1}^{N} \chi_i^2$. Under the null hypothesis of no interaction, $\chi_1^2, \chi_2^2, \ldots, \chi_N^2$ are all independent. Thus, χ_{LN}^2 has chi-square distribution with ν_{LN} degrees of freedom by summing up the degrees of freedom for each transition table ν_i, i.e., $\nu_{LN} = \sum_{i=1}^{N} \nu_i$. A p-value can be computed for χ_{LN}^2 to indicate the statistical significance of the LN model. The p-value provides a means to trade-off between goodness-of-fit and complexity. Therefore, LN reconstruction is to find an LN with the minimum p-value. Since the χ_i^2 statistics for the transition tables at each node are independent of each other, minimization of the overall p-value reduces to minimizing the p-values for individual transition tables at each node.

Once an optimal set of transition tables at each node are identified, truth-tables can be derived by maximum likelihood estimation of probabilities for the multinomial distribution on each row. Each row will be assigned a truth value that corresponds to the maximum probability parameter in its multinomial distribution. Although not implemented in this paper, a probabilistic LN can be reconstructed, not by setting a truth-table, but by keeping the probability parameters in the multinomial distribution for each row. The LN reconstruction algorithm is presented as Algorithm 1. It searches for each node an optimal truth-table that minimizes the p-value with up to M parents. Its complexity is $O\left(N \sum_{i=1}^{M} Q_{\max}^i \binom{N}{i}\right)$, where $Q_{\max}$ is the maximum quantization level of all nodes.

Algorithm 1 Logical-Network-Construction

```
for each node do
  for m ← 0 to M do
    for each possible selection of m parents do
      Accumulate a transition table
      Current p-value ← Perform multinomial test
      if the p-value is smaller than the current minimum p-value for the current node then
        minimum p-value ← current p-value
      end if
    end for
  end for
  Convert the transition table with the minimum p-value to a truth-table for node m by maximum likelihood estimation of multinomial parameters.
end for
Compute the p-value for the logical network
```

5. COMPUTATIONAL MODELING RESULTS

We demonstrate an LN reconstructed from temporal gene expression in mouse brains in response to alcohol to uncover known interactions curated in PathwayArchitect™ (STRATAGENE, La Jolla, CA).

Animal husbandry and cDNA microarray – Thirty-five adult DBA/2J (D2) mice were housed on a 12:12 light:dark cycle and given food and water ad libitum. The mice were habituated for three days to i.p. injections of saline and on the forth day were injected with 20% alcohol in saline in a total dose of 4 g/kg. D2 mice are exquisitely sensitive to alcohol dependence, and at this dose show physical signs consistent with dependence from about 4-10 hours after injection. Brains were removed and anterior cortex tissue was dissected at 2, 7, 12, 24 hours following the alcohol injection with 7 biological replicas at each time point. cDNA microarrays were hybridized using the Array 350 microarray labeling kit.

Initial gene screening and quantization – Through ANOVA across time course, post hoc t-tests and partial least squares analyses, a total of 392 differentially expressed genes were selected because they exhibit both temporal and alcohol related expression variation. Missing gene expression values are imputed using SAM software (Tusher et al., 2001). Among the 392 selected genes, we performed maximum likelihood joint quantization (unpublished) to obtain a list of 19 genes for LN reconstruction. The quantization levels for each dimension were between 1 and 4. The 19 selected genes ended up with exactly 2 quantization levels, while the 373 ignored genes were all quantized to a single level. The alcohol node is assigned based on the experimental condition: 1 for alcohol-injected samples and 0 for control samples.

Logical network reconstruction – A reconstructed LN is shown in Fig. 4. The size of the test is 0.05. The maximum number of parents per node is 6. The overall p-value of the inferred network is 3.6e-05 and the p-values for truth-tables at each node are given in Table 2. In this LN, *Idh3g, Smarce1, 1700029101Rik, Gm740, MGC40675, Fosb, Ckap1*, and *Camk2b* are the most influential gene nodes. In fact, a major neural pathway is represented. The interaction with alcohol for *Smarce1* (Ozimek et al., 2004), *Fosb* (Bachtell et al., 1999), and *Camk2b* (Winston and Maro, 1995) are biologically verified. In addition, nine out of the 19 nodes in our LN (Fig. 5) have been identified as interacting with alcohol from biology literature by PathwayArchitect. This indicates that our approach was indeed successful in capturing significant causal interactions through temporal dependencies. More importantly, however, new hypotheses for several genes that had never before been implicated in alcoholism were generated. The molecular mechanisms of alco-

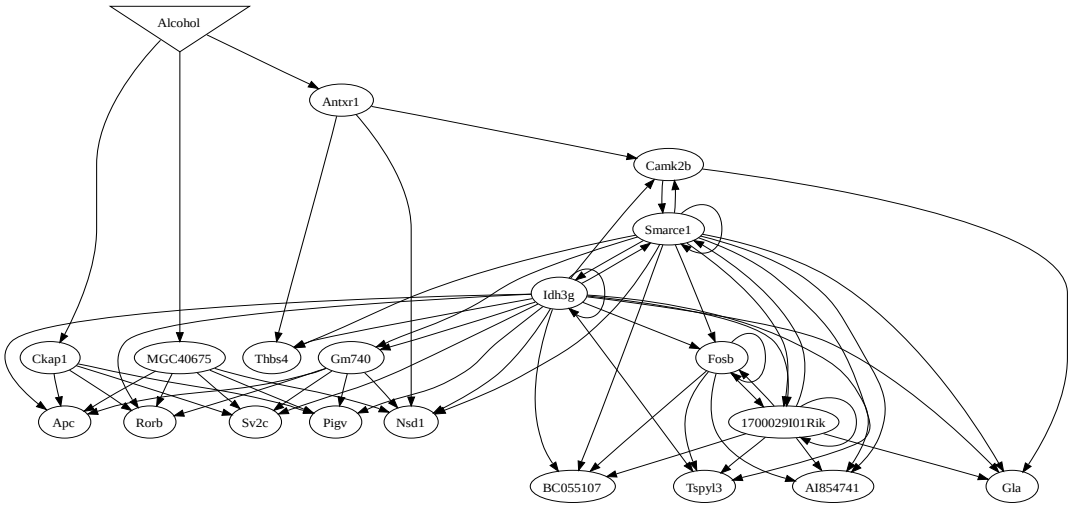

Fig. 4. An inferred LN (p-value=3.6e-05). The oval nodes represent genes and the inverse triangle alcohol.

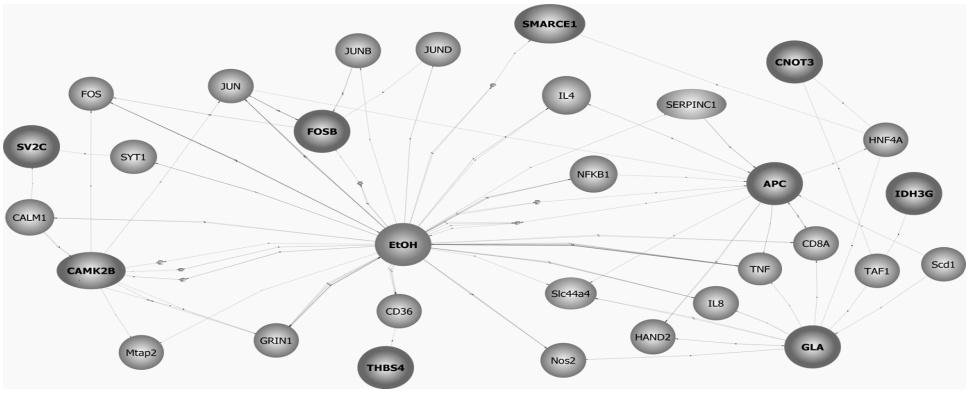

Fig. 5. Genes responsive to alcohol (EtOH) uncovered by PathwayArchitect from literature. The darker nodes were identified in Fig. 4.

Table 2. The p-values and number of parents for each node in the LN.

Node	Symbol	#parents	p-value
1	Alcohol	-	-
2	Idh3g	2	0
3	Rorb	4	2.88658e-15
4	AI854741	4	0
5	Nsd1	5	0
6	Gla	4	0
7	Camk2b	3	4.35629e-12
8	Sv2c	4	0
9	Fosb	4	0
10	Gm740	2	3.08642e-14
11	MGC40675	1	4.996e-15
12	BC055107	4	2.12399e-10
13	Tspyl3	4	0
14	1700029I01Rik	4	0
15	Smarce1	4	3.52554e-05
16	Antxr1	1	3.91767e-11
17	Pigv	4	0
18	Thbs4	3	0
19	Ckap1	1	5.73303e-07
20	Apc	4	1.35336e-13

holism are complex and poorly understood. These results demonstrated that our algorithm can generate and prioritize new hypotheses for understanding complex traits such as alcoholism.

6. CONCLUSIONS AND FUTURE WORK

Our LN reconstruction algorithm identifies significant associations among a subset of genes to a target gene by performing the multinomial test, derived from a statistical property regarding the summation of independent chi-squares. Thus we have offered a unique framework to extract LNs to characterize temporal interactions from time-course gene expression data. Although the alcohol influence on gene expression in mouse brains remains an open problem for current biological investigation, we are among the first to inspect the temporal patterns in gene expression by reconstructing an LN to account for the observed data and to offer a possible explanation for the underlying causal interactions among genes involved in response to alcohol. Some of the inferences made on temporal dependencies corroborate with present knowledge on gene regulations in mouse. Many of the other inferences will be subject to further biological verification, which we expect major discoveries to arise.

The challenge of GRN reconstruction from microarray data is that typically one must select a subset of interesting genes to render the model computable. Approaches which filter genes or gene-gene relations

have been applied. While this leads to the improved signal in the data, it may neglect extensive information on highly relevant genes which are exhibiting subtle variation in the same temporal patterns as other connected genes. Rather than filtering based on statistical effects, one could develop LN models from known pathways and evaluate how they respond and interact with pharmacological perturbations. This strategy can be implemented by inferring LNs from GRNs established by literature mining such as Ingenuity Pathways Knowledge Base (Ingenuity Systems, Redwood City, CA) and PathAssist (JusticeTrax Inc., Mesa, AZ). This will possibly allow the modeling to begin at a more realistic starting point, and will reserve statistical power for the strong plausible relations that are previously reported.

Another important future direction is to incorporate a more diverse set of nodes. The biological relevance of an inferred LN can be substantially improved if simultaneous measurements of the proteome, the metabolome, and the transcriptome are available, without major modifications to the current algorithms. Once data are properly scaled, the method is highly generalizable and has significant potential for inferring temporal relations among widely diverse biological processes. The illustration of the validity of our results from a small time-course gene expression study indicates substantial potential for denser sampling, and for the incorporation of additional data representing other aspects of the neurobiological response to alcohol, including neurohormonal, physiological, and behavioral measures.

ACKNOWLEDGMENTS

MS, CKL, and ERL were supported by the NSF-DOE Faculty and Student Team grant NSF HRD-0420407. EJC was supported by DOE-ORNL. RK was supported by BISTI and DOE. MAL was supported in part by the NIH under grants 1-P01-DA-015027-01, 5-U01-AA-013512 and 1-R01-MH-074460-01, by the DOE under the EPSCoR Laboratory Partnership Program, by the Australian Research Council, and by the European Commission under the Sixth Framework Programme. Additionally, EJC and MAL were supported by NIH/NIAAA INIA Bioinformatics Core and INIA Bioinformatics Pilot; EJC, MAL, and RK by NICDC. SEB was supported by NIH grants AA013182, AA013403, and AA013475.

REFERENCES

T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model. *Theoretical Compu. Sci.*, 298(1): 235–251, 2003.

R.K. Bachtell, Y.M. Wang, P. Freeman, F.O. Risinger, and A.E. Ryabinin. Alcohol drinking produces brain region-selective changes in expression of inducible transcription factors. *Brain Res.*, 847(2): 157–165, 1999.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning Bayesian networks. In *Proc. of Int'l Conf. on Machine Learning*, pages 157–165, 1996.

S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. of Bioinformatics & Compu. Bio.*, 1(2):231–252, 2003.

H. Kim, J.K. Lee, and T. Park. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*, 8 (37):doi:10.1186/1471–2105–8–37, 2007.

S. Klamt, J. Saez-Rodriguez, J.A. Lindquist, L. Simeoni, and E.D. Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(56), 2006.

H. Lähdesmäki, S. Hautaniemi, I. Shmulevich, and O. Yli-Harja. Relationships between probabilistic Boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal Proc.*, 86(4):814–834, 2006.

S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp. on Biocomputing*, 3:18–29, 1998.

A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):663–672, 2006.

I.M. Ong, J.D. Glasner, and D. Page. Modelling regulatory pathways in E. *coli* from time series expression profiles. *Bioinformatics*, 18:S241–S248, July 2002.

P. Ozimek, K. Lahtchev, J.A. Kiel, M. Veenhuis, and I.J. van der Klei. Hansenula polymorpha swi1p and snf2p are essential for methanol utilisation. *FEMS Yeast Res.*, 4(7):673–682, 2004.

R. Pal, I. Ivanov, A. Datta, M.L. Bittner, and E.R. Dougherty. Generating Boolean networks with a prescribed attractor structure. *Bioinformatics*, 21: 4021–4025, 2005.

I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002.

V. Tusher, R. Tibshirani, and C. Chu. Significance analysis of microarrays applied to ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

N.J. Winston and B. Maro. Calmodulin-dependent protein kinase II is activated transiently in ethanol stimulated mouse oocytes. *Dev Biol.*, 170(2):350–352, 1995.