BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 6 2009, pages 772–779
doi:10.1093/bioinformatics/btp028

Systems biology

Gene network reconstruction from transcriptional dynamics under kinetic model uncertainty: a case for the second derivative

David R. Bickel[1,2,*], Zahra Montazeri[2], Pei-Chun Hsieh[3], Mary Beatty[4], Shai J. Lawit[4] and Nicholas J. Bate[4]

[1]Ottawa Institute of Systems Biology, [2]Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, ON K1H 8M5, Canada, [3]Graduate Institute of Systems Biology and Bioinformatics, National Central University, No. 300, Jhongda Road, Jhongli City, Taoyuan County 32001, Taiwan (R.O.C.) and [4]Pioneer Hi-Bred International, Inc., 7300 NW 62nd Avenue, PO Box 1004, Johnston, Iowa, IA 50131-1004, USA

Received on May 9, 2008; revised on December 4, 2008; accepted on January 12, 2009

Advance Access publication February 13, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Measurements of gene expression over time enable the reconstruction of transcriptional networks. However, Bayesian networks and many other current reconstruction methods rely on assumptions that conflict with the differential equations that describe transcriptional kinetics. Practical approximations of kinetic models would enable inferring causal relationships between genes from expression data of microarray, tag-based and conventional platforms, but conclusions are sensitive to the assumptions made.

Results: The representation of a sufficiently large portion of genome enables computation of an upper bound on how much confidence one may place in influences between genes on the basis of expression data. Information about which genes encode transcription factors is not necessary but may be incorporated if available. The methodology is generalized to cover cases in which expression measurements are missing for many of the genes that might control the transcription of the genes of interest. The assumption that the gene expression level is roughly proportional to the rate of translation led to better empirical performance than did either the assumption that the gene expression level is roughly proportional to the protein level or the Bayesian model average of both assumptions.

Availability: http://www.oisb.ca points to R code implementing the methods (R Development Core Team 2004).

Contact: dbickel@uottawa.ca

Supplementary information: http://www.davidbickel.com

1 INTRODUCTION

1.1 Transcriptional network reconstruction

Much of the recent interest in biomolecular network reconstruction is motivated by the desire to map microscopic interactions to macroscopic traits that are of interest to the medical and food industries (Peccoud *et al*., 2004). For example, pharmaceutical companies have an interest in reverse-engineering molecular networks to find druggable targets (Hopkins and Groom, 2002; Schadt *et al*., 2007) or otherwise find genes that strongly influence disease and that could respond to therapy (Chen *et al*., 2008). Markowetz and Spang (2007) provide an introduction to the literature.

Transcriptional networks have been reconstructed from gene expression measured at a snapshot in time, often in response to some set of perturbations or treatments. Expression time-course experiments have also raised the prospect of inferring not only the existence of causal relationships between genes, but also the direction of causality from regulating genes to regulated genes, without requiring the manipulation of genes one by one.

1.2 Bayesian inference of biomolecular networks

1.2.1 Bayesian statistics Approximate Bayesian inference tends to achieve a level of conservatism between, on one hand, hypothesizing the network deemed most likely irrespective of the degree of uncertainty in that network (e.g. Friedman *et al*., 2000; Husmeier, 2003) and, on the other hand, correcting *P*-values for multiple testing.

A recent hierarchical model for inferring regulatory networks via Bayes's theorem is a case in point; information that provides evidence of transcription factors is represented in terms of a prior distribution, whereas the evidence associated with gene expression data remains in the likelihood function (Jensen *et al*., 2007). Such modeling of transcription factors seeks a more detailed understanding than does allowing unknown mediating genes that may encode transcription factors and other intermediate connections between genes in the network to remain unmodeled.

For the purpose of inferring previously unknown influences between transcriptional network components given appropriate data and any well specified model of such influences, many of the most important causal models may be classified either as directed acyclic graphs or as kinetic models. The use of Bayesian probability theory

*To whom correspondence should be addressed.

results in Bayesian networks when applied to directed acyclic graphs but can result in the methodology of this araticle when applied to kinetic models.

1.2.2 Bayesian networks A Bayesian network is a directed acyclic graph whose nodes are random variables and whose edges are associations expressed in terms of conditional probabilities (Jensen, 2001). Bayesian networks interpreted causally (Pearl, 2000) have often been the tool of choice for biomolecular network reconstruction; they have, for example, been employed to uncover evidence of physical relationships between proteins after applying several perturbations (e.g. Sachs *et al*., 2005). Bayesian networks have also been applied to time course data when generalized to networks called dynamic because each node corresponds to a gene at a given time (e.g. Husmeier, 2003; Kim *et al*., 2003). If the conditional independence topology of the network is unknown, as is the case in frontier network reconstruction, then edge indicators, conveniently labeled as 0 or 1 for absent or present edges, also have a joint prior distribution. In principle, the problem is solved with Bayes's theorem by computing the joint posterior distribution of the latent variables, edge indicators and parameters conditional on the observed data. Mathematical difficulties make the exact solution unattainable, sometimes leading to computational searches for a single network that, although optimal in some sense, may have very low posterior probability.

The simplicity of Bayesian networks have made them natural preliminary tools for applying Bayesian inference to the problem of biomolecular network reconstruction. However, the Bayesian statistical framework is general enough to instead incorporate knowledge of the dynamics of physical gene–gene interactions.

1.2.3 Bayesian inference of kinetic models In common with other statistical approaches, Bayesian inference can give misleading results when used without adequate modeling. Consequently, rather than using Bayesian networks, we apply Bayes's theorem to a well studied class of kinetic models to infer causal relationships between genes. Such models have attracted recent attention, but usually without computing the posterior distributions of their parameters (e.g. Chen *et al*., 1999; de Hoon *et al*., 2002; Vander Velden and Peccoud, 2003). A discussion of two important non-Bayesian approaches to kinetic models appears in Section S5 of the Supplementary Material (Bonneau *et al*., 2006; Bonneau *et al*., 2007; Gardner *et al*., 2003). Modeling at a sufficiently high level is supported by the finding that modeling the dynamical system in too much detail can lead to differential equation or stochastic process parameters that can only be identified, even in the absence of statistical error, when there are not only measurements of the abundance of transcripts over time, but also other information such as knowledge of specific interactions between promoters and transcription factors (Zak *et al*., 2002). More generally, model complexity at an inappropriate level for the data at hand often leads to wasted analyst effort, to computational intractability and to parameter overfitting.

Bridging the gap between high-level statistical methods and low-level differential equation models, we herein describe a Bayesian method of inferring causal relationships between genes on the basis of gene expression measurements that have little or no replication and that only roughly reflect numbers of molecules. This is accomplished by carefully approximating both the kinetic models that describe transcriptional dynamics and the posterior probabilities of gene–gene influences based on such models. Compared with other Bayesian methods of inferring kinetic model parameters (Wilkinson, 2006), our approach is simple in that it does not require advanced computation such as that of Markov chain Monte Carlo simulations. Select approximations may be relaxed for more precise inference once higher quality expression data or reliable information from other sources becomes available.

In the transcriptional network reconstruction method, our propose is applicable to studies of gene expression measured at four or more consecutive points separated by equal intervals of time (e.g. Serban and Wasserman, 2003; Spellman *et al*., 1998), provided that such intervals are small enough to capture the transcriptional dynamics, that the total time of measurement is large enough to capture translation, and that there is only one dominant cell type in the tissue samples. While the method requires neither replication (unless there are fewer than five time points) nor information about which genes encode transcription factors, straightforward ways to incorporate either or both into the data analysis are provided. Replication is handled at the level of the statistical model, whereas transcription factor information becomes part of a prior distribution. These necessary requirements for application of the proposed methodology are not sufficient to ensure that there are enough biological samples to obtain reliable predictions of regulation; statistical power depends on the extent of biological variability as well as the sample size. However, the methods are designed to be robust to insufficient data in the sense that all the reported probabilities of regulatory relationships would in that case tend to be very small, thereby helping prevent unwarranted predictions. Section S4 of the Supplementary Material supplies details on the number of time points needed and on the reliability of network inference at different levels of biological variability. Section S5 relaxes the requirement of equal sampling times.

Section 2 describes the kinetic models that, under the conditions specified, hold in all cell systems. Section 3 introduces the regression model and prior distribution used to infer the model parameters representing gene–gene influences. The Supplementary Material reports the findings of a simulation study and illustrates this gene network reconstruction methodology by applying it to the replicated plant cell culture experiment that initially motivated the methodology. The results of applications to data of non-replicated yeast and bacteria experiments are also presented in the Supplementary Material and are summarized Section 4. In the Section 5, we draw general conclusions.

2 TRANSCRIPTIONAL NETWORK MODELS

Consider the set of genes any of which might be the dominant regulator of any gene of interest i. The number m of such potentially regulating genes may equal the genome size in the absence of adequate information about transcription factors. Let $x_i(t)$ denote the transcript abundance of the i-th gene at time t, β_{ij} correspond to a real-valued strength of influence associated with a product of the j-th gene affecting the i-th gene and D_i correspond to the non-negative degradation rate of the transcript of the i-th gene. $\beta_{ij} > 0$ corresponds to activation, whereas $\beta_{ij} < 0$ corresponds to repression.

The linear transcription model of Chen *et al*. (1999) reduces to

$$\frac{dx_i(t)}{dt} = \sum_{j=1, j\neq i}^{m} \beta_{ij} x_j(t) - D_i x_i(t), \quad (1)$$

under the assumption that the transcript concentrations are proportional to *the concentrations* of the proteins they encode, or to

$$\frac{d^2 x_i(t)}{dt^2} \approx \sum_{j=1}^{m} \beta_{ij} x_j(t), \quad (2)$$

under the assumption that the transcript concentrations are proportional to *the rates of concentration change* of the proteins they encode, as found empirically (Chechik *et al*., 2008) and as expected from mass-action kinetics, given that such rates dominate degradation. Due to their linearity, these equations may alternatively be derived as approximations to other models in the literature. For example, in the case of the first-order model (1), $dx_1(t)/dt = \beta_{1,2} x_2(t) - D_1 x_1(t)$ approaches a rate-law kinetic model of transcription (Gardner and Faith, 2005) if $x_2(t)$ is proportional to the concentration of the transcription factor, if that concentration is far from saturation and if the transcription factor binds to a single motif. Equation (1) may be written more concisely by letting $\beta_{ii} = -D_i$:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{m} \beta_{ij} x_j(t), \quad (3)$$

equivalent, in the absence of perturbations, to a model applied to a transcript network in bacteria (Gardner *et al*., 2003). In contrast, some of the simplest corresponding static and dynamic Bayesian network models are

$$x_i(t) = \sum_{j=1, j\neq 1}^{m} \beta_{ij} x_j(t) \quad \text{and} \quad (4)$$

$$x_i(t+\tau) = \sum_{j=i, j\neq 1}^{m} \beta_{ij}(t+\tau) x_j(t+\tau) + \sum_{j=1}^{m} \beta'_{ij}(t) x_j(t), \quad (5)$$

respectively, where τ is the time between each measurement. The parameters of the latter may be practically unidentifiable due to the strong autocorrelation of $x_j(t)$ unless the $\beta_{ij}(t)$s and $\beta'_{ij}(t)$s have highly informative priors. With no more parameters than model (4), model (3) is much more realistic biologically since it has the change in transcript concentration, rather than the absolute transcript concentration, in direct proportion to the concentration of the transcripts of the regulating genes, those for which $\beta_{ij} \neq 0$. The less realistic of these two assumptions is in effect implicitly made whenever standard Bayesian networks (e.g. Sachs *et al*., 2005) or other association networks (e.g. Bickel, 2005; Schäfer and Strimmer, 2005) are interpreted causally. Further, like other causal models based on differential equations (Jensen, 2001), kinetic models (2) and (3) allow feedback loops, regardless of whether their parameters are inferred with Bayes's theorem, whereas Bayesian networks (4) are notorious for their inability to handle feedback loops.

The complete Bayesian solution to model (3) would be the joint posterior distribution of β_{ij} for all values of i and j computed on the basis of the observations, one or more error models, and a joint prior distribution encoding all relevant biological knowledge and its uncertainties. While that ideal cannot be attained, it supplies guidance for achieving approximate solutions and, as necessary, indicates a direction in which they may be improved.

3 STATISTICAL METHODS

3.1 First-order difference equations

3.1.1 Regression framework To apply model (3) to gene expression data, the concentrations $x_j(t)$ are replaced by their observed values $y_j(t)$ after averaging over any technical replicates. These measurements are considered approximately proportional to the transcript copy numbers of their genes. For example, with a microarray platform, $y_j(t)$ could be a hybridization intensity (or a monotonic transformation of such an intensity) deemed roughly proportional to the mRNA concentration level; $y_j(t)$ may be more accurately estimated given platform-specific information (Frigessi *et al*., 2005). With a tag-based method of measuring expression, $y_j(t)$ is the abundance of tags corresponding to the j-th gene (Gainetdinov *et al*., 2007; Hu and Polyak, 2006). Also replacing the time derivative with the first-forward difference yields, at times $t \in \{\tau, 2\tau, \ldots, t_{\max} - \tau, t_{\max}\} = \mathrm{T}$,

$$\Delta y_i(t) \equiv y_i(t+\tau) - y_i(t) = \sum_{j=1}^{m} \beta_{ij} y_j(t) + \beta_i + \varepsilon_i(t),$$

where β_{ij} has absorbed the constant sampling time τ, the intercept β_i represents the unexplained linear trend over time, and $\varepsilon_i(t)$ represents the error due to biological variability. If there is biological replication, the replicates are denoted by $k \in \{1, \ldots, n\}$ and the observed values by $y_{jk}(t)$. In a repeated measures design, there would be a total of n individual organisms or cell cultures, and each value of k would refer to the same individual over all points in time, leading to the autoregressive model

$$y_{ik}(t+\tau) - y_{ik}(t) = \sum_{j=1}^{m} \beta_{ij} y_{jk}(t) + \beta_i + \varepsilon_i(t).$$

However, in most studies of gene expression on time scales of biochemical reactions, measuring the same individual over time is impractical, either because each individual organism must be sacrificed to take the tissue sample or because such sampling will perturb an individual's future gene expression dynamics, thereby introducing a systematic bias that increases over time. Thus, the common situation is that a value of k at one time does not correspond to the same value of k at a later time, and, unless some data are missing, there will be a total of $(t_{\max}/\tau)n$ individuals. A computationally efficient tactic for applying the forward difference approximation to data of this structure reduces them by averaging over replicates at each point in time. This yields $\bar{y}_i(t)$, the mean expression intensity of the i-th gene over the n replicates at time t. A simplistic data reduction method would then stipulate model

$$\bar{y}_i(t+\tau) - \bar{y}_i(t) = \sum_{j=1}^{m} \beta_{ij} \bar{y}_j(t) + \beta_i + \varepsilon_i(t),$$

which forfeits most of the degrees of freedom and thus fails to adequately take advantage of the biological replication. Data reduction may be performed without such substantial information

loss by instead assuming the regression model

$$\Delta y_{ik}(t) \equiv \bar{y}_i(t+\tau) - y_{ik}(t) = \sum_{j=1}^{m} \beta_{ij}\bar{y}_j(t) + \beta_i + \varepsilon_{ik}(t). \quad (6)$$

Here, $\varepsilon_{ik}(t)$ is a residual assumed drawn from a zero-mean normal distribution of standard deviation (SD) σ_i and density f. The unknown quantities, β_{ij}, β_i and $\varepsilon_{ik}(t)$, are random variables in the Bayesian framework, whereas the design matrix $\mathbf{y}$ is fixed by the observations $\bar{y}_i(t)$ for all t; it has $(t_{\max}/\tau - 1)n = (|\mathrm{T}|-1)n$ rows corresponding to the forward differences $\Delta y_{ik}(t)$, also fixed, and $m+1$ columns corresponding to a unit intercept column and the m regression coefficients. The following methodology is developed with biological replication ($n \geq 2$) in mind for the sake of generality, but it applies equally to studies without replication since $n = 1$ implies

$$\bar{y}_j(t) = y_{i1}(t) = y_i(t); \varepsilon_{i1}(t) = \varepsilon_i(t); \Delta y_{i1}(t) = \Delta y_i(t).$$

Even though no individual is measured over time, this approach can be informative if there are some common aspects of gene expression dynamics that are reflected across a sufficient number of the individuals measured and if t denotes the elapsed time after some common perturbation.

3.1.2 Case of complete measurements In the case of no missing regressors, expression measurements are available over time for all m potentially regulating genes. Unless all m regressor genes may be considered regulators for gene i, each regression coefficient β_{ij} equals 0 with some non-zero prior probability. Thus, facilitating the selection of non-zero coefficients, Equation (6) may conveniently be rewritten as

$$\Delta y_{ik}(t) = \sum_{j=1}^{m} \alpha_{ij}\beta_{ij}\bar{y}_j(t) + \beta_i + \varepsilon_{ik}(t); \alpha_{ij} = I(\beta_{ij} \neq 0), \quad (7)$$

where $I(\cdot)$ is the indicator function mapping to 1 if its argument is true or to 0 if it is false. In terms of model selection or Bayesian model averaging for the i-th response gene, the 2^m possible values of the m-tuple $(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im})$ span the sub-model space. Restricting this sub-model space such that $\sum_{j=1}^{m} \alpha_{ij} \leq M$ for some $M \leq (|\mathrm{T}|-1)n-2$ makes inference manageable. This facilitates leveraging the conceptual separation between sub-model selection/averaging and inference about the parameters conditional on each of the sub-models in the restricted space, e.g. $(\beta_{i7}, \beta_i, \sigma_i)$ for sub-model $\alpha_{ij} = I(j = 7)$ or $(\beta_{i1}, \ldots, \beta_{iM}, \beta_i, \sigma_i)$ for sub-model $\alpha_{ij} = I(j \in \{1, \ldots, M\})$. Conveniently, the number of regression parameters per sub-model is no more than the number of observations per sub-model even if $m >> (|T|-1)n$. Were the goal to find a set of predictive sub-models, we could proceed by stochastic search through the restricted sub-model space (see Casella and Moreno, 2006) or perhaps by conventional stepwise selection. However, Bayesian quantification of network uncertainty instead requires the computation of $P(\alpha_{ij} = 1|\mathbf{y})$.

In the simplest situation, one regulating gene dominates all others and expression measurements are available for all m potentially regulating genes. The prior distribution of α_{ij} then satisfies $\forall_{j,j' \neq j} P(\alpha_{ij} = 1, \alpha_{ij'} = 1) = 0$, i.e. $\sum_{j=1}^{m} \alpha_{ij} = 1$with probability 1, so that only m sub-models have non-zero probability. Unless there is evidence outside the expression experiment favoring some genes above others as the regulator of gene i, each possible sub-model has equal prior probability: $\forall_{j \in \{1,\ldots,m\}} P(\alpha_{ij} = 1) = 1/m$. This uniformity of the prior does not represent all states of previous information; for example, given the results of a transcription factor prediction algorithm, the prior probability for each potentially regulating gene would increase monotonically with the algorithm's reported evidence that it codes for a transcription factor. To simplify notation, we express the posterior probability that the j-th gene is the regulator of the i-th gene in terms of Bayes factors BF_{ij} with respect to any arbitrarily chosen sub-model of positive probability:

$$P(\alpha_{ij} = 1|\mathbf{y}) = \chi_i \mathrm{BF}_{ij} = \left(1 + \sum_{j'=1, j' \neq 1}^{m} \frac{\mathrm{BF}_{ij'}}{\mathrm{BF}_{ij}}\right)^{-1}$$

for proportionality constant χ_i. In terms of BIC_{ij}, the Bayesian information criterion (Schwarz, 1978), the BF can be approximated up to constant factor c_i as

$$\mathrm{BF}_{ij} \approx \mathrm{BF}_{ij}^{\mathrm{Schwarz}} = c_i \exp\left(\frac{-\mathrm{BIC}_{ij}}{2}\right) = c_i \sup_{\beta_{ij}, \beta_i, \sigma_i} f\left(\mathbf{y}|\alpha_{ij} = 1\right),$$

in this case proportional to the profile likelihood ratio used in the likelihood ratio test statistic since the number of free parameters is the same in each sub-model. Since f is the normal probability density function, the BF is approximated in terms of the ratio of the sums of squared residuals, resulting in

$$P\left(\alpha_{ij} = 1|\mathbf{y}\right) \approx \left(1 + \sum_{j'=1, j' \neq j}^{m} (\hat{\sigma}_{ij}/\hat{\sigma}_{ij'})^{(|\mathrm{T}|-1)n}\right)^{-1}, \quad (8)$$

where the maximum likelihood estimator of the observation SD, conditional on the dominance of the j-th gene, equals the estimated root mean square error (RMSE):

$$\hat{\sigma}_{ij} = \sqrt{(|\mathrm{T}|-1)^{-1} n^{-1} \sum_{k \in \{1,\ldots,n\}, t \in \{\tau, \ldots, t_{\max}-\tau\}} \hat{\varepsilon}_{ijk}^2(t)}.$$

This use of the BIC enables genome-scale analyses by obviating the computationally prohibitive simulations required by many other Bayesian methods, and does so without unduly compromizing the reliability of the resulting inferences. In fact, $\lim_{n \to \infty} \mathrm{BF}_{ij}^{\mathrm{Schwarz}}/\mathrm{BF}_{ij} = 1$ if $\mathrm{BF}_{ij}^{\mathrm{Schwarz}}$ is the BIC approximation of the BF and BF_{ij} is the BF that corresponds to a nested null model defined by a single value of the parameter of interest, in this case β_{ij}, that has a normal prior with information equal to that of one observation (Kass and Wasserman, 1995). Further, the approximation $\mathrm{BF}_{ij} \approx \mathrm{BF}_{ij}^{\mathrm{Schwarz}}$ does not require unrealistic amounts of data: a simple normal model with known SD has found the BIC approach to the BF excellent even for samples each of as few as five observations (Kass and Wasserman, 1995). Less formally, there is typically no reason to approximate the BF to greater accuracy than a factor of two (Jeffreys, 1948), and the BIC has been found to work about as well for model averaging as computation based on the specification of the joint prior distribution of the parameters (Hastie *et al*., 2001). Clyde and George (2004) compare the BIC approximation with other methods of computing BF.

For purposes of inferring biological networks, the main advantage of genome-wide platforms such as those of microarrays and tag-based methods may be that enough of the genome is represented

that $P(\alpha_{ij}=1|\mathbf{y})$ and analogous quantities can often attain high values even without previous information on the gene network. Such posterior probabilities provide informative upper bounds of how much confidence one may reasonably place in causal relationships between genes on the basis of observed data when honestly accounting for network uncertainty. In contrast, conventional technologies such as RT-PCR may suffer from the missing regulator problem of the next subsection. Although microarray platforms do not enable direct comparison of intensities between genes, that limitation does not affect our approach since gene-to-gene variations in intensity are included in the β_{ij} s.

3.1.3 Generalization to missing regulators To generalize the above methodology, consider the set of genes that could be the regulator of any gene of interest i. Suppose m', the number of those genes with expression measurements over time, is less than m, the number of potentially regulating genes, which might be the genome size or the number of putative transcription factors. (If the uncertainty in m is substantial, it may be assigned a prior distribution to propagate that uncertainty to the posterior probability of each model.) Although the SD estimates of the missing $m-m'$ genes cannot be known, none of them would be greater than $\hat{\sigma}_{i0}$, the SD estimates based on the intercept term of Equation (7) with $\forall_{j\in\{m'+1,\ldots,m\}}\alpha_{ij}=0$. Then the data for the first m' genes provide an approximate upper bound on the posterior probability of each of their models:

$$\underset{j\in\{1,\ldots,m'\}}{\forall} P\left(\alpha_{ij}=1|\mathbf{y}\right) \tag{9}$$

$$\leqslant\left(1+\sum_{j'=1,j'\neq j}^{m'}\left(\hat{\sigma}_{ij}/\hat{\sigma}_{ij'}\right)^{(|\mathrm{T}|-1)n}+\left(m-m'\right)\left(\hat{\sigma}_{ij}/\hat{\sigma}_{i0}\right)^{(|\mathrm{T}|-1)n}\right)^{-1}$$

To the extent that $m'<<m$, this bound differs from the approximation made by optimistically assuming that none of the non-measured genes is the dominant gene that regulates gene i, i.e. $\forall_{j\in\{m'+1,\ldots,m\}}P(\alpha_{ij}=1|\mathbf{y})=0$:

$$\underset{j\in\{1,\ldots,m'\}}{\forall} P_{\text{naive}}\left(\alpha_{ij}=1|\mathbf{y}\right)\approx\left(1+\sum_{j'=1,j'\neq j}^{m'}\left(\hat{\sigma}_{ij}/\hat{\sigma}_{ij'}\right)^{(|\mathrm{T}|-1)n}\right)^{-1}. \tag{10}$$

Assumptions of this kind, usually implicit, plague attempts to reconstruct transcriptional networks from expression measured on small fractions of the set of potentially regulating genes. If there are expression measurements for $\sim m$ genes, the effect of using (10) instead of (9) may be quantified by randomly selecting, without replacement, subsets each of m' genes for the computations.

3.2 Second-order difference equations

If doubling the mRNA copy number of a regulator doubles *a rate of translation* rather than doubling the *amount of a protein* in the cell, the methods of Section 3.1 require modification. Using the second-order derivative Equation (2) instead of the first-order derivative Equation (3) puts the second-forward difference equation

$$\begin{aligned}\Delta^2 y_{ik}(t)&\equiv\bar{y}(t+2\tau)-2\bar{y}_i(t+\tau)+y_{ik}(t)\\&=\sum_{j=1}^{m}\tilde{\beta}_{ij}\bar{y}_j(t)+\tilde{\beta}_i+\tilde{\varepsilon}_{ik}(t)\end{aligned} \tag{11}$$

in place of the first-forward difference, Equation (6). The data analysis would then proceed similarly to the first-order derivative analysis described above if the researchers are certain that the assumptions behind Equation (2) are more realistic than those behind Equation (3). Then the equivalent of Equation (8) would be

$$P\left(\tilde{\alpha}_{ij}=1|\mathbf{y}\right)\approx\left(1+\sum_{j'=1,j'\neq 1}^{m}\left(\hat{\tilde{\sigma}}_{ij}/\hat{\tilde{\sigma}}_{ij'}\right)^{(|\mathrm{T}|-2)n}\right)^{-1}. \tag{12}$$

for $\tilde{\alpha}_{ij}=I(\tilde{\beta}_{ij}\neq 0)$. If they lack such certainty, their uncertainty can be reflected in the amount of prior probability assigned to each model, as developed below.

3.3 Uncertainty in the difference equation order

To reflect complete uncertainty about whether doubling the mRNA copy number of a regulator doubles a rate of translation rather than doubling the amount of a protein in the cell, we assigned 50% of the prior probability to the model of Equation (2) and 50% to that of Equation (3). Then the model selection problem in the complete measurement case may be framed in terms of this analog of Equation (7):

$$\begin{aligned}&\left(\sum_{j=1}^{m}\alpha_{ij}\right)\Delta y_{ik}(t)+\left(\sum_{j=1}^{m}\tilde{\alpha}_{ij}\right)\Delta^2 y_{ik}(t)\\&=\sum_{j=1}^{m}\alpha_{ij}\beta_{ij}\bar{y}_j(t)+\sum_{j=1}^{m}\tilde{\alpha}_{ij}\tilde{\beta}_{ij}\bar{y}_j(t)+\beta_\iota+\varepsilon_{ik}(t)\end{aligned} \tag{13}$$

given that one regulating gene dominates all the others for a given gene assumed to be regulated and that the domination is described either by Equation (2) or by Equation (3), but not by both, so that only a single influence coefficient will be non-zero for each value of i. As before, each regulating gene has equal prior probability in the absence of information favoring any over others. These considerations constrain the model indicator prior by $\forall_{i\in\{1,\ldots,m\}}P(\sum_{j=1}^{m}\alpha_{ij}+\sum_{j=1}^{m}\tilde{\alpha}_{ij}=1)=1$ and

$$\underset{(i,j)\in\{1,\ldots,m\}^2}{\forall} P(\alpha_{ij}=1)=P(\tilde{\alpha}_{ij}=1)=(2m)^{-1}.$$

Averaging over two models for each (i, j) pair, the posterior probability that gene j is the regulator of gene i is $P(\alpha_{ij}=1\cup\tilde{\alpha}_{ij}=1|\mathbf{y})=P(\alpha_{ij}=1|\mathbf{y})+P(\tilde{\alpha}_{ij}=1|\mathbf{y})$. Then, following the same approximations that led to Equation (8),

$$\pi_{ij}\equiv P\left(\alpha_{ij}=1\cup\tilde{\alpha}_{ij}=1|\mathbf{y}\right)\approx\left(\widehat{\sigma}_{ij}^{-(|\mathrm{T}|-2)n}+\widehat{\tilde{\sigma}}_{ij}^{-(|\mathrm{T}|-2)n}\right)C_i, \tag{14}$$

where C_i is the proportionality constant such that $\sum_{j'=1}^{m}\pi_{ij'}=1$. The exponents were adjusted since there are only $\left(|\mathrm{T}|-2\right)n$ terms in the computation of the RMSE $\hat{\tilde{\sigma}}_{ij}$ for the second-forward difference model since each time t in the sum is one of $|\mathrm{T}|-2$ elements of $\{\tau,\ldots,t_{\max}-2\tau\}$.

The posterior probability that the first-order difference model (6) is correct for the presumed regulated gene i can also be computed by averaging, now over all the possible regulating genes instead of

Table 1. The AUC estimates for the four datasets

Dataset	First-order AUC	Second-order AUC	Average AUC
Kao *et al.* (2004)	0.47	0.60	0.69
Bansal *et al*. (2006)	NA	0.39	0.45
Spellman *et al*. (1998)	0.40	1.00	0.61
de Lichtenberg *et al*. (2005)	0.20	0.76	0.20

The AUC estimates under the second-order model are consistently greater than those under the first-order model, suggesting that the former predicts putative transcription factors better than the latter.

over the two model orders:

$$P\left(\sum_{j'}\alpha_{ij'}=1\middle|\mathbf{y}\right)\approx C_i\sum_{j'}^{m}\widehat{\sigma}_{ij'}^{-(|\mathrm{T}|-2)n}. \tag{15}$$

Since by construction those model orders are mutually exclusive and jointly exhaustive, the posterior probability that the second-order difference model (11) is correct is

$$P\left(\sum\tilde{\alpha}_{ij'}=1\middle|\mathbf{y}\right)=1-P\left(\sum_{j'}\alpha_{ij'}=1\middle|\mathbf{y}\right). \tag{16}$$

4 VALIDATION USING PUBLIC DATA

We applied our models to two yeast datasets (de Lichtenberg *et al*., 2005; Spellman *et al*., 1998) and two bacteria dataset (Bansal *et al*., 2006; Kao *et al*., 2004). Each dataset is from a different strain.

For each regulated gene of interest, we found the gene with the highest posterior probability of being its dominant regulating gene. We then checked the annotation of those probability-maximizing genes in EchoBASE, GeneDB and in the *Saccharomyces* Genome Database and noted which among them were putative transcription factors. To quantify performance, we estimated the Area Under the receiver operating characteristic Curve (AUC) (Green and Swets, 1966) between the probability-maximizing genes that encode putative transcription factors and the probability-maximizing genes that do not. The AUC measures how accurately the models predicted putative transcription factors. An AUC of 0 indicates that low probabilities perfectly predict putative transcription factors (least desirable), an AUC of 0.5 indicates that there is no predictive power (also undesirable) and an AUC of 1 indicates that higher probabilities perfectly predict putative transcription factors (ideal). (The AUC has also been applied to the problem of determining which genes are differentially expressed between groups (Bickel, 2004; Pepe *et al*., 2003)).

Table 1 summarizes the results for the four datasets. Figure 1, Figure 2 and Figure 3 show the probabilities of the genes that maximized the probability under the first-order, second-order and averaged models. The Supplementary Material supplies additional details about our analyses of these data.

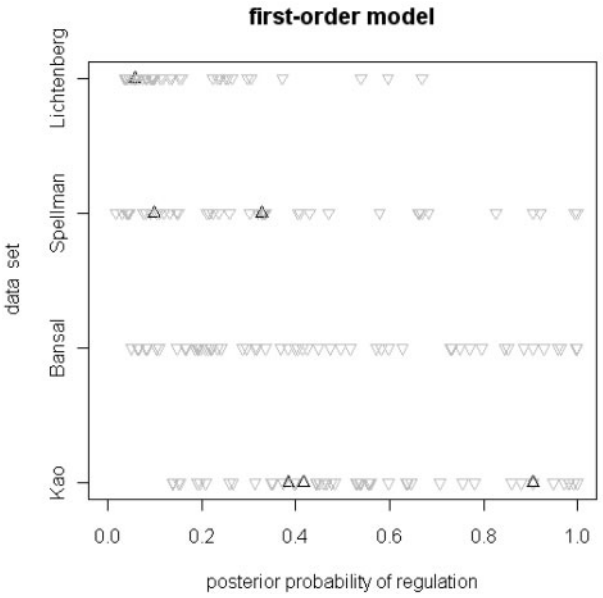

Fig. 1. Posterior probabilities based on the first-order model [Equation (8)] for each of the four datasets. Black triangles represent genes encoding putative transcription factors and gray triangles represent the other probability-maximizing genes. The Supplementary Material reports the numeric values.

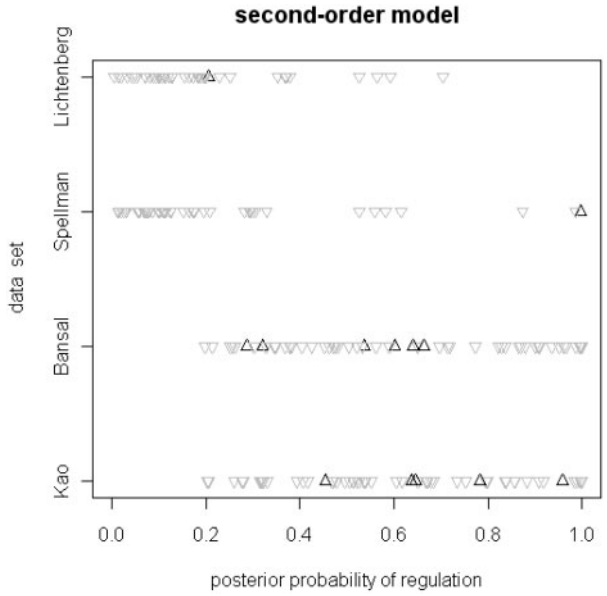

Fig. 2. Posterior probabilities based on the second-order model [Equation (12)] for each of the four datasets. Black triangles represent genes encoding putative transcription factors and gray triangles represent the other probability-maximizing genes. The Supplementary Material reports the numeric values.

5 DISCUSSION

We highlight three unique aspects of our choosing prior distributions to represent available information for causal network inference. First, under substantial uncertainty about which differential equation best models the transcriptional influences on the transcription of a given gene of interest, we assign each differential equation a prior probability followed by averaging the posterior probabilities of influences over all differential equations considered. This strategy,

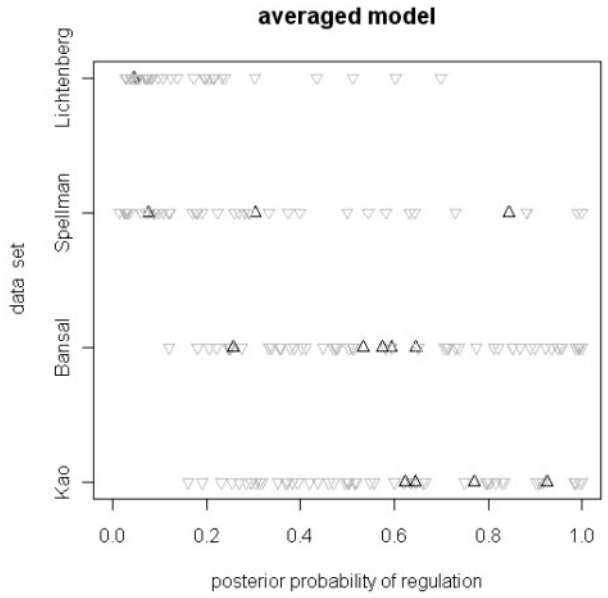

Fig. 3. Posterior probabilities based on the average model [Equation (14)] for each of the four datasets. Black triangles represent genes encoding putative transcription factors and gray triangles represent the other probability-maximizing genes. The Supplementary Material reports the numeric values.

unlike the analysis of the data on the basis of a single kinetic model, propagates the uncertainty in the kinetic model to the uncertainty in the causal relationships between genes inferred in the analysis [Equation (14)]. Second, the proposed method of assigning a prior distribution is general enough to cover both the absence and presence of information about which genes encode transcription factors. This is accomplished by using previous information to assign a prior probability to each gene that could be the one that dominates the expression of the gene of interest. For example, lacking such information, every gene in the genome of the species studied has a prior probability equal to $1/m$, the reciprocal of the genome size, thereby guarding against the practice of automatically inferring a causal network even when expression has been measured only over a small fraction of the genome. On the other hand, a selection of genes to measure guided by previous transcription factor information can be expected to yield better results. Our approach reflects this by assigning a prior probability for each gene selected on the order of $1/m$, where m in this case is the number of transcription factors. Third, the distinction between m, the number of genes that might dominate the expression of the gene of interest, and m', the number of genes with measured expression levels, generalizes the methodology without introducing a bias toward overconfidence in network connections. In other words, unless the number of genes measured is comparable to the number of genes that have not been ruled out as the gene dominating the gene of interest, adequate specification of the prior distribution makes it unlikely that any measured gene will have 50% or more posterior probability of influencing the expression of the gene of interest [Equation (9)]. This applies separately to each of the regulated genes of interest, the number of which is limited only by the availability of expression measurements and of computational resources.

This methodology may seem overly stringent when compared with algorithms that would hypothesize networks of hundreds of genes even from data of questionable adequacy. While we have indeed guarded against putting undue confidence in causal interpretations of estimated association networks, we have not embraced the conservatism in traditional hypothesis testing that seeks to avoid all false discoveries. Based in part on the results of the Supplementary Material, we take the moderate position that some aspects of causal gene networks may be inferred with at least some degree of confidence using current technology. This is accomplished by coherently inferring parameters of kinetic models as protection against overstating or understating how much can be learned from gene expression time courses. Systems biology as a field is discredited by the publication of more and more large transcriptional networks without quantifying the extent to which such networks are justified by experimental data and what is already known about the systems.

Even with such precautionary measures, more thorough modeling of the uncertainty may yield lower posterior probabilities of gene–gene influences than those of Equation (14):

(1) The linearity between the transcript abundance and the microarray intensity becomes a less adequate approximation at high-mRNA copy numbers, as the fluorescence becomes saturated (K. V. Velden, personal communication). The possible sensitive dependence of conclusions on this and other linearities of the models can in principle be mitigated by the introduction of non-linearity parameters and their prior distributions, preferably informed by empirical studies of the relationship between the copy number and the intensity (e.g. Frigessi *et al*., 2005). Alternatively, tag-based platforms may be employed for more direct measurement of mRNA copy numbers (Gainetdinov *et al*., 2007; Hu and Polyak, 2006). However, a change in platform will prove insufficient to the extent that transcription factors do not regulate their targets linearly.

(2) As noted in the caption of Figure S4 of the Supplementary Material, leaving prior time scale information out of the model approximation may in some cases result in misleadingly high posterior probabilities.

(3) While the models were inspired largely by dynamics expected of transcript levels within an individual cell, the microarray measures total gene expression over a population of several cells. The deviation of single-cell dynamics from the linearity of the kinetic models would have to be summed over all cells of the population to determine the adequacy of those models at the population level. Modeling the relationship between individual-cell expression and microarray measurements would require the introduction of parameters such as the number of cells per population, which, when marginalized over to compute to posterior probabilities of interest, may add considerable uncertainty to the conclusions.

To the extent that the last of these considerations affects the probability of regulation, network reconstruction breakthroughs may occur in the short term more by advances in within-cell measurement technology than by those in statistical modeling and computation. Further, no gene expression platform can detect post-transcriptional modifications known to be important in regulation. These kinds

of limitations apply more generally: since models can always be improved, statistics tends to only provide lower bounds on the uncertainty of inferences about complex systems (cf. Cox, 2001).

Nonetheless, the second-order model performed well in validation (Table 1). One AUC estimate is 100%, reflecting the case in which the only putative transcription factor has a higher posterior probability of being the dominant regulator of a gene of interest than any of the other 42 genes.

ACKNOWLEDGEMENTS

We are very grateful to Maria Fedorova for information on BMS cell lines; Hai Zhu for computational support; George Casella and Merlise Clyde for conversations regarding Bayesian statistics; Jon Lightner, Chris Martin, Mark Whitsitt and Bob Merrill for institutional support; Jennifer Chung for assistance with sample collection; Brian Zeka and John Nau for profiling technical support; Danielle Dewar-Darch for helpful information about yeast; Kent Vander Velden and Corey Yanofsky for insightful comments on the manuscript; and Carolina Perez-Iratxeta for a discussion on the merits of tag-based measurements of gene expression. In addition, Jean Peccoud provided indispensable leadership of the plant case study during the brainstorming and experimental design phases. We also thank Mark Cooper, Tim Helentjaris, Chris Zinselmeier and Dave Selinger for field-trial, association-network discussions that prompted us to seek causal interactions between genes. Finally, the suggestions of all of the anonymous reviewers lead to a much more thorough manuscript.

Funding: Supported in part by the Canada Foundation for Innovation (CFI16604) and the Ministry of Research and Innovation (MRI16604).

Conflict of Interest: none declared.

REFERENCES

Bansal,M. *et al*. (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Bickel,D.R. (2004) Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics*, **20**, 682–688.

Bickel,D.R. (2005) Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics*, **21**, 1121–1128.

Bonneau,R. *et al*. (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Bonneau,R. *et al*. (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell*, **131**, 1354–1365.

Casella,G. and Moreno,E. (2006) Objective Bayesian variable selection. *J. Am. Stat. Assoc.*, **101**, 157–167.

Chechik,G. *et al*. (2008) Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat. Biotechnol.*, **26**, 1251–1259.

Chen,T. *et al*. (1999) Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput.*, **4**, 29–40.

Chen,Y. *et al*. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.

Clyde,M. and George,E.I. (2004) Model uncertainty. *Stat. Sci.*, **19**, 81–94.

Cox,D.R. (2001) Comment on 'Statistical Modeling: The Two Cultures' (Leo Breiman). *Stat. Sci.*, **16**, 216–231.

de Hoon,M.J.L. *et al*. (2002) Inferring gene regulatory networks from time-ordered gene expression data using differential equations. In Lange,S. *et al*. (eds.), *Proceedings of the 5th International Conference on Discovery Science*. Springer-Verlag, Berlin, pp. 267–274.

de Lichtenberg,U. *et al*. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, **21**, 1164–1171.

Friedman,N. *et al*. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Frigessi,A. *et al*. (2005) Genome-wide estimation of transcript concentrations from spotted cDNA microarray data. *Nucleic Acids Res.*, **33**, 1–13.

Gainetdinov,I.V. *et al*. (2007) Use of short representative sequences for structural and functional genomic studies. *Biochemistry (Moscow)*, **72**, 1179–1186.

Gardner,T.S. and Faith,J.J. (2005) Reverse-engineering transcription control networks. *Phys. Life Rev.*, **2**, 65–88.

Gardner,T.S. *et al*. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Green,D.M. and Swets,J.A. (1966) *Signal Detection Theory and Psychophysics*. John Wiley and Sons, Inc., New York.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.

Hu,M. and Polyak,K. (2006) Serial analysis of gene expression. *Nat. Protoc.*, **1**, 1743–1760.

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

Jeffreys,H. (1948) *Theory of Probability*. 2nd edn. Oxford University Press, London.

Jensen,F.V. (2001) *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York.

Jensen,S.T. *et al*. (2007) Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Stat.*, **1**, 612–633.

Kao,K.C. *et al*. (2004) Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc. Natl Acad. Sci. USA*, **101**, 641–646.

Kass,R.E. and Wasserman,L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.*, **90**, 928–934.

Kim,S.Y. *et al*. (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinformatics*, **4**, 228–235.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks - A review. *BMC Bioinformatics*, **8**.

Pearl,J. (2000) *Causality*. Cambridge University Press, Cambridge.

Peccoud,J. *et al*. (2004) The selective values of alleles in a molecular network model are context dependent. *Genetics*, **166**, 1715–1725.

Pepe,M.S. *et al*. (2003) Selecting differentially expressed genes from microarray experiments. *Biometrics*, **59**, 133–142.

Sachs,K. *et al*. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Schadt,E.E. *et al*. (2007) Computer systems and methods for associating genes with traits using cross species data. United States Patent 20070166707.

Schäfer,J. and Strimmer,K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Serban,N. and Wasserman,L. (2003) Identifying genes altered by a drug in temporal microarray data: a case study. In *Proceedings of the Joint Statistical Meetings of the American Stastical Association*. Biopharm. Student Paper Competition Award.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Vander Velden,K. and Peccoud,J. (2003) Modeling networks of molecular interactions in the living cell: structure, dynamics, and applications. In Sanders,W.H. and Ciardo,G. (eds.), *Proceedings of the10th International Workshop on Petri Nets and Performance Models*, IEEE Computer Society, Washington, DC, pp. 2–10.

Wilkinson,D.J. (2006) *Stochastic Modelling for Systems Biology*. Taylor & Francis, Boca Raton, Florida.

Zak,D.E. *et al*. (2002) Local identifiability: when can genetic networks be identified from microarray data? In *Proceedings of the Third International Conference on Systems Biology*.