APPROXIMATE BAYESIAN APPROACHES FOR REVERSE ENGINEERING BIOLOGICAL NETWORKS

Andrea Rau

Florence Jaffr´ezic

Jean-Louis Foulley

R. W. Doerge

See next page for additional authors

Recommended Citation

Rau, Andrea; Jaffr´ezic, Florence; Foulley, Jean-Louis; and Doerge, R. W. (2010). "APPROXIMATE BAYESIAN APPROACHES FOR REVERSE ENGINEERING BIOLOGICAL NETWORKS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1067

Author Information

Andrea Rau, Florence Jaffr´ezic, Jean-Louis Foulley, and R. W. Doerge

APPROXIMATE BAYESIAN APPROACHES FOR REVERSE ENGINEERING BIOLOGICAL NETWORKS

Andrea Rau[1,2*], Florence Jaffrézic[2], Jean-Louis Foulley[2], and R. W. Doerge[1,3]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.
[2]INRA AgroParisTech, Animal Genetics and Integrative Biology, Populations Statistics Genomes, 78350 Jouy-en-Josas, France
[3]Department of Agronomy, Purdue University, West Lafayette, IN 47907, U.S.A.
[*]Corresponding author: arau@stat.purdue.edu

ABSTRACT: Genes are known to interact with one another through proteins by regulating the rate at which gene transcription takes place. As such, identifying these gene-to-gene interactions is essential to improving our knowledge of how complex biological systems work. In recent years, a growing body of work has focused on methods for reverse-engineering these so-called gene regulatory networks from time-course gene expression data. However, reconstruction of these networks is often complicated by the large number of genes potentially involved in a given network and the limited number of time points and biological replicates typically measured. Bayesian methods are particularly well-suited for dealing with problems of this nature, as they provide a systematic way to deal with different sources of variation and allow for a measure of uncertainty in parameter estimates through posterior distributions, rather than point estimates. Our current work examines the application of approximate Bayesian methodology for the purpose of reverse engineering regulatory networks from time-course gene expression data. We demonstrate the advantages of our proposed approximate Bayesian approaches by comparing their performance on a well-characterized pathway in *Escherichia coli*.

1 Introduction

The development of microarray technology in the mid-1990s (Schena et al., 1995, 1996; Lipschutz et al., 1999) made possible large-scale studies of gene expression. Since that time, microarrays have become a popular platform to study the behavior of genes during specific biological processes in a variety of organisms, such as the cell cycle in *Saccharomyces cerevisiae* (Spellman et al., 1998) and the life cycle of *Drosophila melanogaster* (Arbeitman et al., 2002). By collecting tissue samples from an organism and measuring gene expression over several time points, gene expression profiles can be assembled to elucidate information about the relationships occurring among genes in an organism during a particular biological process. Although microarrays are presently the most prominent and least expensive platform for such time-course studies, the decreasing cost and refinement of next generation sequencing (NGS) technologies (Mardis, 2008) suggests that time-course expression profiles may be studied with sequence-based approaches in the near future.

In spite of the abundance of data generated from time-course studies of gene expression, it can be very difficult to unravel the complexity of the chemical dynamics that occur within a cell. One reason for this is that cell development is regulated by well-orchestrated patterns of expression

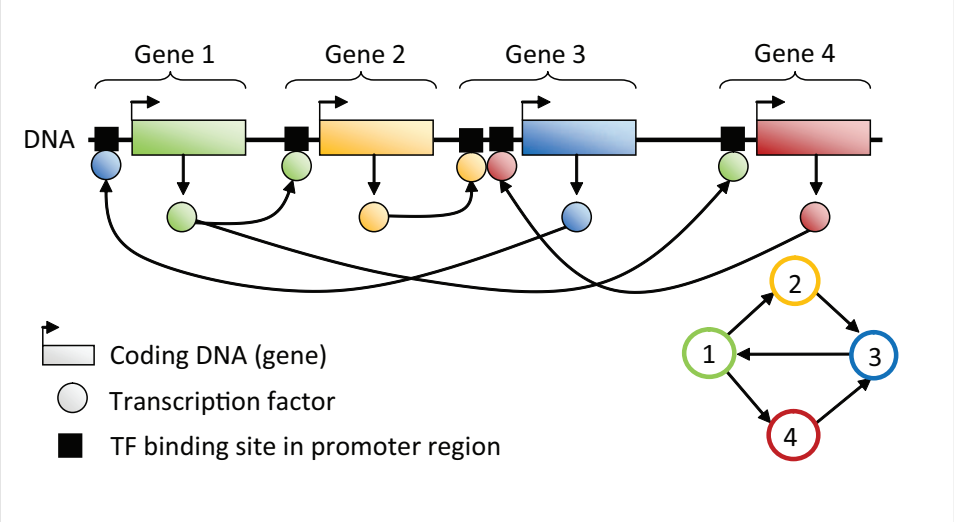

Figure 1: A simple gene regulatory network made up of four genes, represented by colored boxes. In this example, each gene is transcribed and translated into a transcription factor protein (colored circles), which in turn binds to the promoter regions of genes (rectangles) in the network to regulate their expression. The gene regulatory network may be represented using the graph in lower right corner, made up of four nodes (genes) and five directed edges (gene-to-gene interactions). Image taken from Rau (2010).

among groups of genes, often referred to as *gene regulatory networks* (Friedman, 2004; Wilkinson, 2009). Gene regulatory networks are generally believed to govern the rate at which genes in the network are expressed, and as such they often play a critical role in the control of complicated cellular functions. Correctly identifying the components of gene regulatory networks and the gene-to-gene interactions contained therein is thus essential to understanding how complex biological systems work.

Within regulatory networks, genes interact with one another indirectly through proteins known as transcription factors (TF). By binding to the promoter region (i.e., an upstream region of DNA that facilitates transcription) of a gene, a TF controls the transfer of information during transcription by promoting (activating) or blocking (repressing) RNA polymerase, which in turn affects the level of expression of that gene (Schlitt and Brazma, 2007, see Figure 1). Graphs are often used to visualize gene regulatory networks, where nodes represent genes and edges represent interactions among the genes (Figure 1, bottom right). In addition, gene regulatory networks are often characterized by a set of properties common to such biological pathways (Figure 2). First, feedback loops (Figure 2, images 1 and 2) are common motifs that are able to shape signalling responses over time or given particular cellular conditions (Brandman and Meyer, 2008). Second, gene regulatory networks tend to be composed of spoke-and-hub type structures (Figure 2, image 3), where regulated genes are one step away from their regulator (Alon, 2007). Third, gene networks are typically sparse, and genes are often regulated by a limited number of other genes (Leclerc, 2008). In terms of a graphical structure, this means that the fan-in (or in-degree) of each node is typically small

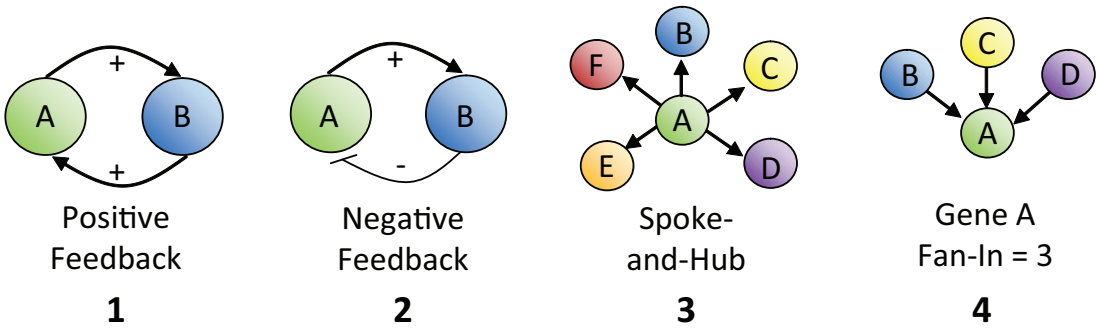

Figure 2: Illustration of characteristics of gene regulatory networks. (1) A positive feedback loop, where gene A activates gene B (represented by an edge with an arrowhead), and gene B in turn activates gene A. (2) A negative feedback loop, where gene A activates gene B while gene B represses gene A (represented by an edge ending in a bar). (3) A spoke-and-hub type structure, with gene A acting as a central regulator gene. (4) The fan-in (number of regulators) for gene A is 3. Image taken from Rau (2010).

(Figure 2, image 4). Finally, because it can be difficult to measure the abundance of a particular TF experimentally, the level of expression of its corresponding gene (i.e., the gene that produces the TF through the process of transcription and translation) is typically used as a proxy.

Understanding how genes interact with one another during a biological process is currently a major goal of the systems biology community. Two basic types of approaches are used to identify the gene-to-gene interactions present in a set of observed gene expression data (Tegnér et al., 2003): the forward engineering approach, which identifies and quantifies fundamental equations of gene regulation based on principles of biochemistry, and the reverse engineering approach, which attempts to discover gene-to-gene interactions (i.e., the structure of the corresponding graph) from a set of gene expression data. We focus our attention on the latter goal, where gene expression is measured within the same organism across time. In this context, statistical methods are used to infer either an adjacency matrix or a parameter matrix from time-course gene expression data (Figure 3). An adjacency matrix is composed of ones and zeroes that indicate the presence or absence of a gene-to-gene interaction (edge) in the network (graph), respectively. A parameter matrix contains additional information about the magnitude and type (i.e., positive for activations and negative for repressions) of each gene-to-gene interaction (i.e., non-zero edges in the graph). That is, the larger the magnitude of a particular element of the parameter matrix, the larger the regulatory effect of the gene-to-gene interaction (and consequently, the thicker the edge in the graph).

The process of reverse engineering gene regulatory networks from time-course gene expression data is a challenging task for several reasons. First, in time-course studies of gene expression using high-throughput technologies (e.g., microarrays), the number of genes observed is typically far greater than the number of samples (e.g., biological replicates or time points). For this reason, the task of reverse engineering gene regulatory networks from such data falls squarely within the $n \ll p$ paradigm typical of genomic studies. As such, to avoid an explosion in model complexity,

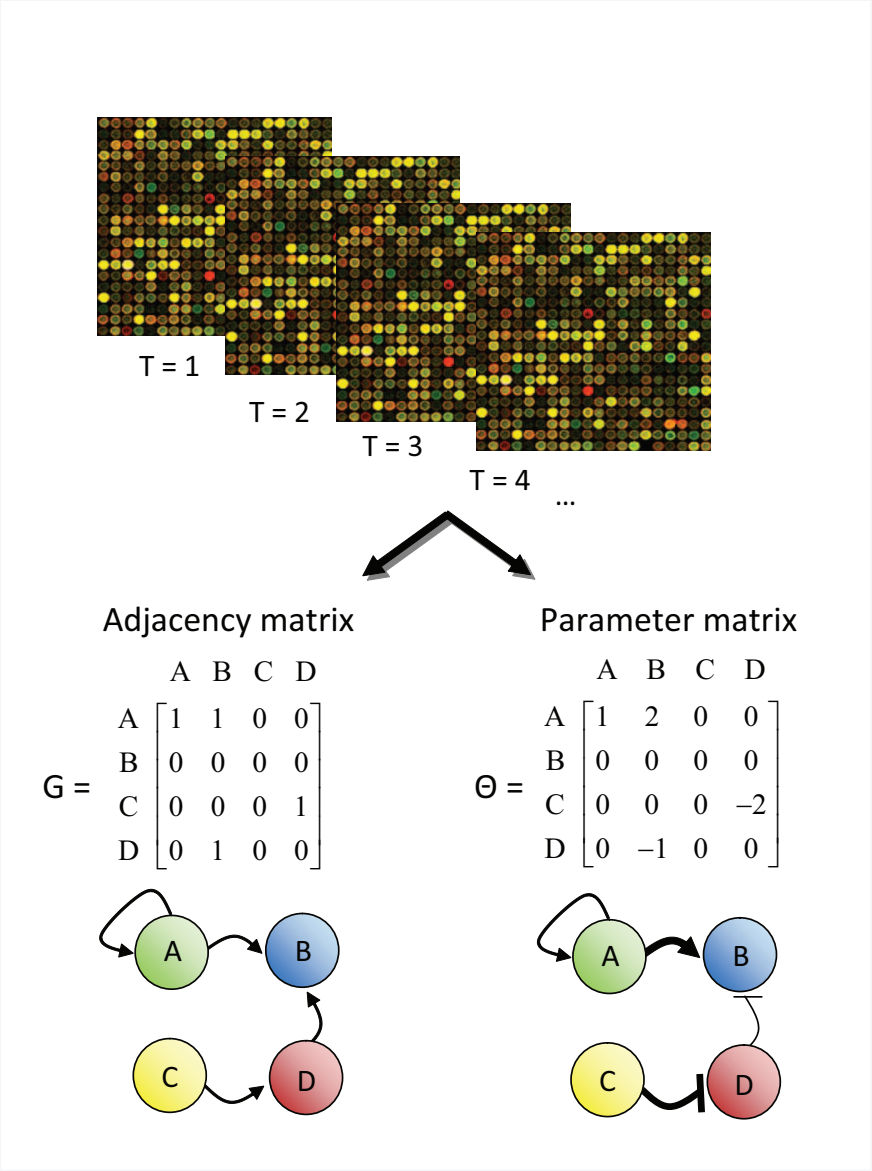

Figure 3: An illustration of the process of reverse engineering a gene regulatory network from longitudinal data. High throughput technologies (e.g., microarrays) are used to measure the gene expression in biological samples taken across several time points. Statistical methods then may infer a network adjacency matrix (bottom left), where ones and zeroes indicate the presence or absence of an edge in the graph (i.e., gene-to-gene interaction in the network), respectively. Alternatively, other approaches also include more detailed descriptions of network structure through a parameter matrix (bottom right), where non-zeroes indicate the magnitude and type (activation or repression) of interactions present in the network, and zeroes represent the absence of an interaction. In this representation, thick edges in the graph represent stronger effects, arrowheads activations, and barred lines repressions. Image taken from Rau (2010).

model parameters are typically set such that the transition probabilities between time points $t-1$ and t are the same for all t (Husmeier et al., 2005). In addition, because genes within a regulatory network interact with one another while reacting to the cellular environment, the structure of the network can inherently be very complex itself. A direct consequence of this network structure is that the resulting expression data often exhibit high multicollinearity. In this work, two approaches based in approximate Bayesian methodology (Carlin and Louis, 2000; Beaumont et al., 2002) are described to reverse engineer gene regulatory networks from time-course gene expression data. The two methods are jointly applied to analyze a well-characterized pathway in the model organism *Escherichia coli*.

2 Approximate Bayesian Methods

The Bayesian framework is well-suited to the inference of gene regulatory networks for a variety of reasons. First, as the number of genes potentially involved in a given network increases, the number of possible networks structures increases exponentially (Husmeier et al., 2005). This problem of high dimensionality is exacerbated by the limited number of biological replicates and time points measured in most real data. In addition, it is often the case that many network structures may yield similarly high likelihoods, making it difficult to determine a single globally optimal structure. As such, posterior distributions for network structures may be more informative about network structures, as well as specific gene-to-gene interactions. Finally, a Bayesian framework also enables *a priori* biological information (e.g., from bioinformatics databases) to be incorporated in a particular model via the prior distribution structure.

In the Bayesian paradigm, a model $f(Y|\Theta)$ is fit to observed data Y, where the parameters Θ are also random variables following a prior distribution, $\pi(\Theta)$. Inference is then passed on the conditional distribution of the parameters given the observed data, $\pi(\Theta|Y) \propto f(Y|\Theta)\pi(\Theta)$, also referred to as the posterior distribution. Because it can be difficult to conduct a full Bayesian analysis in closed form for complex models, such as for gene regulatory networks, we focus here on two approaches based in approximate Bayesian methodology: namely, empirical Bayes methods (Carlin and Louis, 2000) and Approximate Bayesian Computation (Beaumont et al., 2002).

2.1 Empirical Bayes Methods

In the first approach, we develop an empirical Bayes estimation procedure to perform network inference (Rau et al., 2010; Rau, 2010). This method was motivated by that of Beal et al. (2005), based on variational Bayesian learning of linear feedback state space models (SSM). Under the SSM framework, a pair of linear equations is used to describe the interactions occurring among a set of genes and a set of hidden states from one time point to the next. Specifically, consider observed time-course gene expression data $Y = \{\mathbf{y}_{tr}\}$ and unobserved hidden states $X = \{\mathbf{x}_{tr}\}$ with P genes, K hidden states, T time points, and R biological replicates. Let $\mathbf{y}_{tr}$ and $\mathbf{x}_{tr}$ represent the expression of the sets of genes and hidden states, respectively, in replicate r at time t. The state

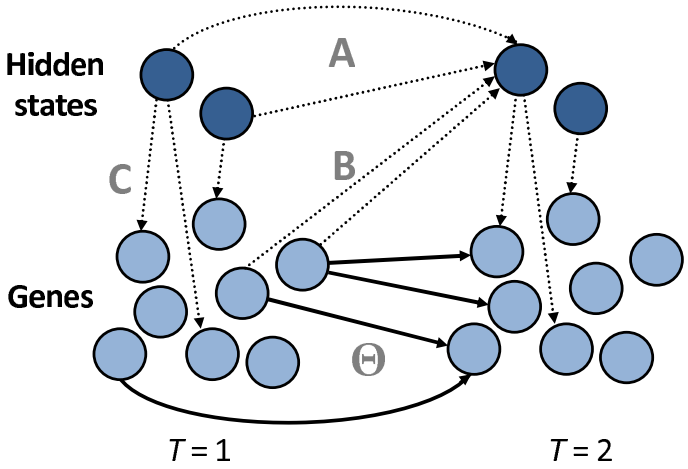

Figure 4: A visual representation of the linear feedback state space model, with the observed expression of a set of genes (light blue nodes) and the unobserved expression of a set of hidden states (dark blue nodes) at two time points, $T = 1$ and $T = 2$, where A, B, C, and Θ correspond to the matrices in Equation 1. The solid arrows, representing the nonzero elements of Θ, correspond to the direct gene-gene interactions that make up the gene regulatory network. Image taken from Rau (2010).

and observation equations, respectively, for the SSM are:

$$\begin{aligned} \mathbf{x}_{tr} &= A\mathbf{x}_{t-1,r} + B\mathbf{y}_{t-1,r} + \mathbf{w}_{tr} \\ \mathbf{y}_{tr} &= C\mathbf{x}_{tr} + \Theta\mathbf{y}_{t-1,r} + \mathbf{z}_{tr}. \end{aligned} \tag{1}$$

where $\mathbf{w}_{tr} \sim N(0, I)$ and $\mathbf{z}_{tr} \sim N(0, V^{-1} = \text{diag}(\mathbf{v}^{-1}))$, with $\mathbf{v}$ being a P-dimensional vector of gene precisions, for $t = 1, ..., T$ and $r = 1, ..., R$ (see Figure 4). Due to its restrictive distributional (Gaussian) assumptions, the SSM is best suited to exploratory analyses of gene regulatory networks where little *a priori* biological information is known.

The primary entity of interest in the SSM of Equation 1 is the parameter matrix Θ, which encodes the direct gene-to-gene interactions from one time to the next. In addition, a hierarchical Bayesian framework may be defined on the SSM of Equation 1 (Beal et al., 2005; Rau et al., 2010). That is, let $\mathbf{a}_{(j)}$, $\mathbf{b}_{(j)}$, $\mathbf{c}_{(j)}$, and $\boldsymbol{\theta}_{(j)}$ denote vectors made up of the jth rows of matrices A, B, C, and Θ, respectively. Then

$$\begin{aligned} \mathbf{a}_{(j)}|\boldsymbol{\alpha} &\sim N(0, \text{diag}(\boldsymbol{\alpha})^{-1}) \\ \mathbf{b}_{(j)}|\boldsymbol{\beta} &\sim N(0, \text{diag}(\boldsymbol{\beta})^{-1}) \\ \mathbf{c}_{(i)}|\boldsymbol{\gamma}, v_i &\sim N(0, v_i^{-1}\text{diag}(\boldsymbol{\gamma})^{-1}) \\ \boldsymbol{\theta}_{(i)}|\boldsymbol{\delta}, v_i &\sim N(0, v_i^{-1}\text{diag}(\boldsymbol{\delta})^{-1}) \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^T$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]^T$, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_K]^T$, $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_P]^T$, v_i is the ith component of vector $\mathbf{v}$, $j = 1, \ldots, K$ and $i = 1, \ldots, P$. Thus, we have a set of parameters

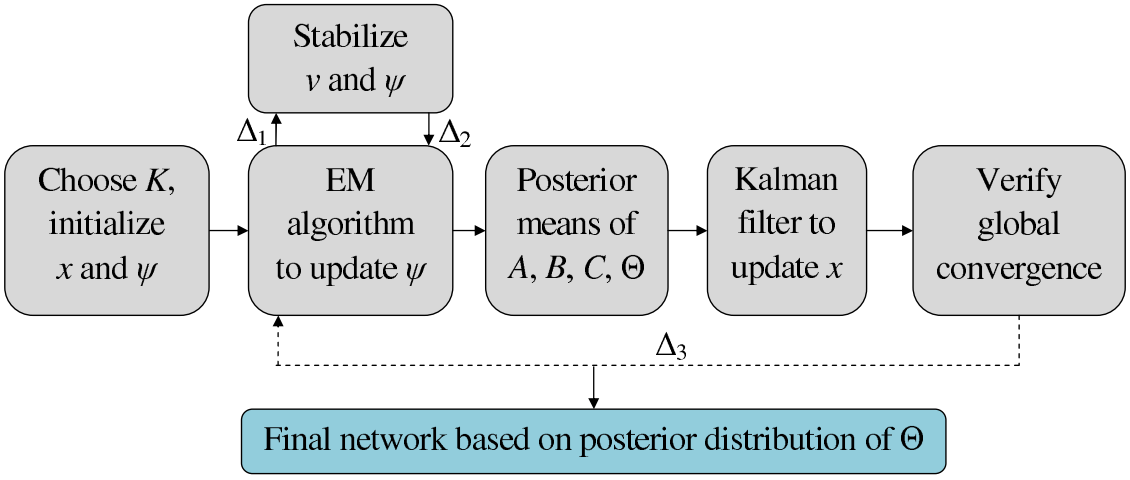

Figure 5: Visual representation of the typical workflow of the EBDBN algorithm. After selecting the hidden state dimension K, two sub-loops of the EM algorithm are used to update model hyperparameters ψ (using convergence criteria Δ_1 and Δ_2). Posterior means of the model parameters and Kalman filter estimates of the hidden states are subsequently calculated. When global convergence is attained (based on convergence criterion Δ_3), the posterior distribution of matrix Θ may be obtained. Image taken from Rau (2010).

$\{A, B, C, \Theta, \mathbf{v}\}$ and a set of hyperparameters $\psi = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$ describing the *a priori* precisions of the parameter set.

The Empirical Bayes Dynamic Bayesian Network (EBDBN) method is an iterative procedure used to infer gene regulatory networks by obtaining the posterior distribution of parameter matrix Θ (Figure 5). To do, first the dimension of the hidden state (i.e., K) is chosen using a time series method for model selection, based on the autocovariances between observations (see Bremer, 2006; Bremer and Doerge, 2009; Rau et al., 2010; Rau, 2010, for more details). Second, a set of recursive calculations known as the Kalman filter and smoother (Kalman, 1960) is used to estimate the values of the hidden states (Bremer and Doerge, 2009; Rau et al., 2010), given the current values of the model parameters. Third, posterior distributions for the model parameters A, B, C, and Θ are calculated based on a two-step Expectation-Maximization (EM) estimation (Dempster et al., 1977) of model hyperparameters ψ. See Rau et al. (2010) for additional details on the EBDBN algorithm.

2.2 Approximate Bayesian Computation

In the second approach, we apply a simulation-based Bayesian method to conduct a detailed analysis of small, well-characterized pathways under fewer model assumptions. By exploiting the capabilities of modern computing, this method makes possible inference on the posterior distribution of gene networks, even in cases where the likelihood is intractable or difficult to calculate. In this approach, we focus on a first-order vector autoregressive (VAR) model (i.e., a multivariate autoregressive model) as an approximation to the dynamics of the gene regulatory network occurring

Algorithm 1 The ϵ-Tolerance Rejection Sampler.

0. Set $i = 0$.
1. Sample a candidate parameter vector $\Theta^\star$ from prior distribution $\pi(\Theta)$.
2. Simulate data $Y^\star$ from the model described by conditional probability distribution $f(\cdot|\Theta^\star)$.
3. Compare simulated data $Y^\star$ to the observed data Y using a distance function ρ and tolerance ϵ. If $\rho(Y^\star, Y) \leq \epsilon$, accept $\Theta^\star$, otherwise reject.
4. If $i < N$ (a pre-set number of acceptances), return to 1.

in time-series expression data. Although such models are popular in econometric analyses (Enders, 2004), they have seen limited application in modeling gene regulatory network, due in part to the difficulty in estimation model parameters for sparse, high-dimensional data using standard statistical approaches (Opgen-Rhein and Strimmer, 2007).

Specifically, let $Y = \{\mathbf{y}_{tr}\}$ represent the observed time-course gene expression data as before. The first-order VAR model, denoted VAR(1), may be written as follows:

$$\mathbf{y}_t = \Theta\mathbf{y}_{t-1} + \mathbf{z}_t \tag{3}$$

where $\mathbf{z}_t$ is a vector of white noise such that $E(\mathbf{z}_t) = 0$, $E(\mathbf{z}_t^2) = \sigma^2$, and $E(\mathbf{z}_t\mathbf{z}_s)$ = 0 for $t \neq s$, and Θ is a $P \times P$ coefficient matrix representing the direct gene-to-gene interactions as before. Note that the VAR(1) model is a simplification of the SSM in Equation 1, where $A = B = C = 0$ and the distributional assumption on $\mathbf{z}_t$ is removed. For notational simplicity, in this section we focus on the case where only one biological replicate is available ($R = 1$), but the extension to multiple replicates is straightforward.

Because no distributional assumptions are made on the error terms $\mathbf{z}_t$ in Equation 3, it may be impossible or computationally prohibitive to compute the likelihood $L(\Theta|Y) = f(Y|\Theta)$ of a given network. In such cases, a sampling-based approach known as Approximate Bayesian Computation (ABC) can enable Bayesian inference (Beaumont et al., 2002; Marjoram et al., 2003; Ratmann et al., 2007). At their core, all ABC methods follow the same general form (Pritchard et al., 1999), known as the ϵ-tolerance rejection sampler (Algorithm 1), where a distance function ρ and tolerance ϵ are used to determine whether simulated and observed data are "close" to one another. In the context of gene regulatory networks, gene expression data for a given network $\Theta^\star$ are simulated using one-step ahead predictors based on Equation 3, such that $\mathbf{y}_t^\star = \Theta^\star\mathbf{y}_{t-1}$.

When $\epsilon > 0$, Algorithm 1 is approximate and its output amounts to simulating from the prior when $\epsilon \to \infty$. When $0 < \epsilon < \infty$, Algorithm 1 results instead in a sample of parameters from the distribution $\pi(\Theta|\rho(Y^\star, Y) \leq \epsilon)$. If ϵ is sufficiently small, then this approximate posterior distribution is a good approximation to the true posterior distribution $\pi(\theta|Y)$. However, in practice a balance must be achieved between a small enough tolerance to obtain a good approximation to the posterior, and a large enough tolerance to allow for feasible computation time and acceptance rates. Typically, ABC algorithms must be repeated a large number of times (on the order of $N = 1 \times 10^6$ or more), and only parameter values corresponding to the smallest $\alpha\%$ (e.g., 1%) of ϵ are used for inference (Beaumont et al., 2002).

Several adaptations and improvements have been proposed to the standard ABC form, including methods based on summary statistics calculated on the simulated and observed data (Beaumont

et al., 2002), post-adjustments of the approximate posterior distributions based on the calculated distances (Beaumont et al., 2002; Leuenberger and Wegmann, 2009), and Monte Carlo techniques to improve efficiency (Marjoram et al., 2003; Sisson et al., 2007; Beaumont et al., 2009). In this work, we focus on an adaptation of the ABC-Markov chain Monte Carlo (MCMC) technique of Marjoram et al. (2003) called the ABC-MCMC Network (ABC-Net) method (Rau, 2010). In the ABC-Net method, a Markov chain is constructed using the Metropolis-Hastings algorithm (Hastings, 1970) where the equilibrium distribution of the chain is chosen to be the approximate posterior distribution $\pi(\Theta|\rho(\mathbf{y}^\star, \mathbf{y}) \leq \epsilon)$. If $\pi(\Theta)$ and $q(\Theta^\star|\Theta^i)$ represent the prior distribution of the network parameter matrix and the probability of moving from Θ^i to $\Theta^\star$ at the $(i+1)^{\text{st}}$ iteration, then a proposed network parameter matrix $\Theta^\star$ is accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(\Theta^\star)q(\Theta^i|\Theta^\star)}{\pi(\Theta^i)q(\Theta^\star|\Theta^i)}\mathbf{I}(\rho(\mathbf{y}^\star, \mathbf{y}) \leq \epsilon)\right\} \tag{4}$$

where $\mathbf{I}(\cdot)$ is an indicator function. Implementing the ABC-Net method in practice requires several adaptations specific to gene regulatory networks, in particular defining an appropriate distance function ρ, threshold ϵ, prior distribution $\pi(\cdot)$, and proposal distribution $q(\cdot|\cdot)$. See Rau (2010) for additional details on these considerations.

3 Data Analysis

The S.O.S. DNA repair system of *Escherichia coli* is a small, well-characterized gene network that is responsible for repairing DNA after damage (Ronen et al., 2002). Specifically, under normal cellular conditions a master repressor called lexA suppresses the expression of genes within the S.O.S. system (Figure 6). However, when DNA damage is detected by one of the S.O.S. proteins (recA), it becomes activated and provokes the autocleavage of lexA, which subsequently drops in abundance. This in turn suspends the repression of the remaining S.O.S. genes, and these genes become activated. When the detected DNA damage has been repaired, the levels of the recA protein drop, allowing lexA to reaccumulate in the cell and suppress the S.O.S. genes. At this point, the cell returns to its original state.

The S.O.S. DNA repair system is a useful example to illustrate the complementary nature of the EBDBN and ABC-Net methods, as it is a benchmark dataset in which specific gene-to-gene regulatory interactions are well understood. In particular, we make use of data collected by Ronen et al. (2002), which focus on a sub-network within the S.O.S. DNA repair system made up of eight genes: uvrD, lexA, umuD, recA, uvrA, uvrY, ruvA, and polB. Using green fluorescent protein (GFP) reporter plasmids, Ronen et al. (2002) measured the expression of these eight S.O.S. genes at fifty time points (every six minutes following ultraviolet irradiation of the cells to provoke DNA damage). The authors performed a set of four experiments for each of two different intensities of ultraviolet light; in this work, we focus on the data collected for Experiment 3 (using 20 JM^{-2}). The data are directly available at the authors' website (`http://www.weizmann.ac.il/mcb/UriAlon`).

Because the gene-to-gene interactions in the S.O.S. DNA repair system are well-defined and no hidden states are believed to be involved in the network, we apply the EBDBN method with

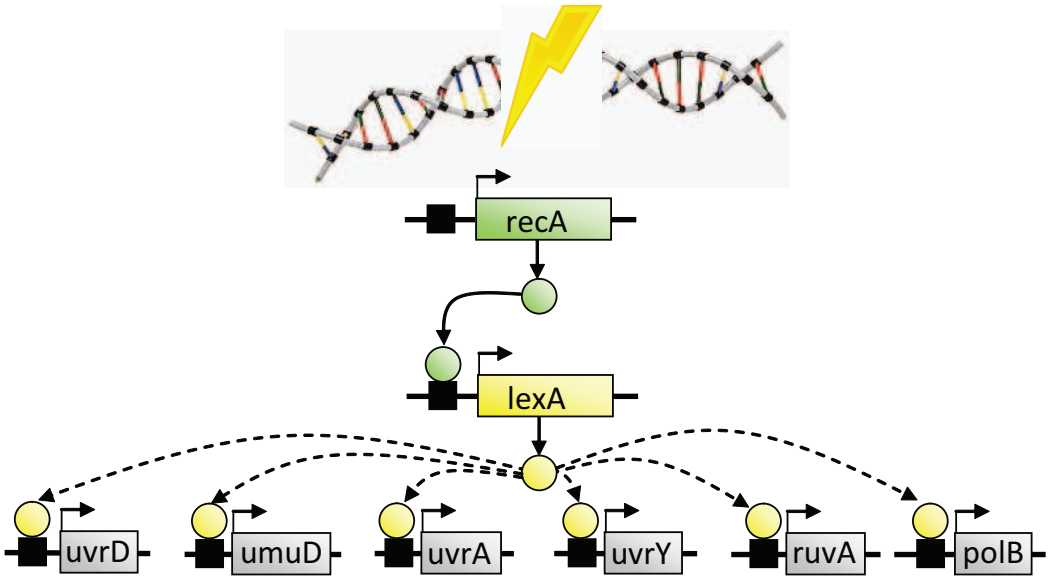

Figure 6: The S.O.S. DNA repair system of *E. coli*. Under normal conditions, the master repressor lexA represses the expression of the S.O.S. genes (uvrD, umuD, uvrA, uvrY, ruvA, and polB) responsible for DNA repair. When DNA damage is detected by the protein recA, it becomes activated and provokes the autocleavage of lexA. This in turn provokes the de-repression of the S.O.S. genes. After DNA damage is repaired, the level of recA drops, lexA reaccumulates in the cell, and the S.O.S genes return to their original state. Image taken from Rau (2010).

a hidden state dimension of K = 0, where a 99.9% cutoff is used as a threshold for the z-scores of the edges. We also apply the ABC-Net method to these data. We set the Gaussian proposal standard deviation of the ABC-Net method to 0.5 (see Rau, 2010, for more details), and we ran the algorithm for ten independent chains of length 1×10^6, with a thinning interval of 50. The VAR(1) simulator of Equation 3 was used to generate simulated data $Y^\star$, and the bounds of the prior distribution on the parameter matrix Θ were set to (-2,2). A Euclidean distance function was used to compare simulated and observed data, where the threshold ϵ was selected based on the 1% quantile of distances based on 5000 randomly generated networks. Due to the small size of the network, the maximum fan-in of each node in the graph was constrained to 2 or less (i.e., each gene has a maximum of two regulators). For additional details on the analysis, see Rau (2010).

The results of the EBDBN and ABC-Net methods for the S.O.S. repair system are shown in Figure 7. In this figure, blue and red solid edges represent "true positives" and "false postives" identified using the EBDBN method, according to the previously described behavior of the S.O.S. network. However, note that we use these terms somewhat loosely, as even for well-understood networks such as the S.O.S. DNA repair system, the absence of a particular gene-to-gene interaction in the literature cannot indicate with absolute certainty that such a relationship is absent. Gray dotted lines represent gene-to-gene interactions supported by the literature that are not identified by the EBDBN method. In Figure 7, we also include the marginal approximate posterior

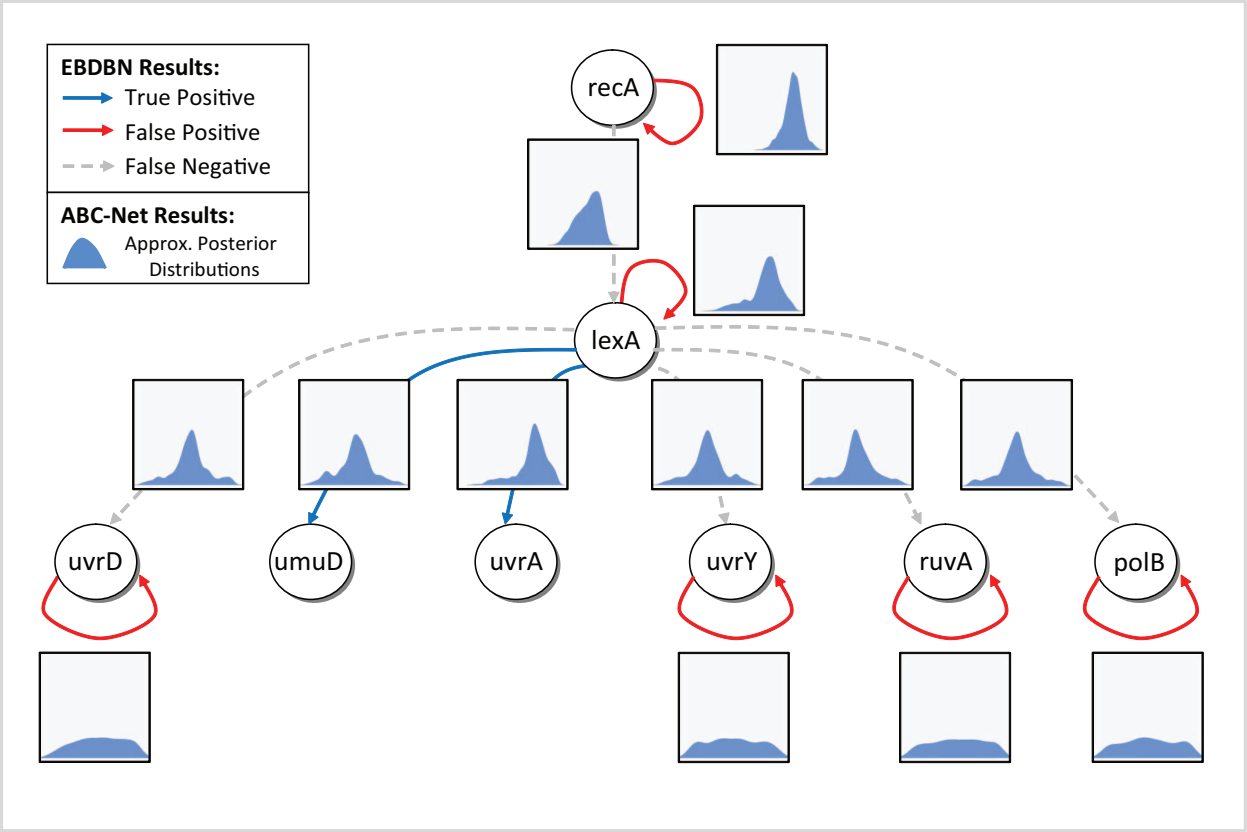

Figure 7: Results for the S.O.S DNA repair system for the EBDBN and ABC-Net methods. Blue and red solid edges in the network represent gene-to-gene interactions identified by the EBDBN method that are "true positives" and "false positives", according to the known behavior of genes in the S.O.S. network. Dotted gray lines represent gene-to-gene interactions supported by the literature that are not identified by the EBDBN method. Blue-filled densities represent the marginal approximate posterior distributions found through the ABC-Net method. The feedback loops on the S.O.S. genes (uvrD, uvrY, ruvA, and polB) appear to be flexible edges, while other identified edges exhibit greater rigidity. Image taken from Rau (2010).

distributions for each of these gene-to-gene interactions obtained by the ABC-Net method. These approximate posterior distributions seems to fall into two categories: diffuse distributions (e.g., the feedback loops on uvrD, uvrY, ruvA, and polB) and peaked distributions (the remaining interactions). We refer to these types of posterior distributions as flexible and rigid edges within the graph, respectively.

By examining the results of the EBDBN and ABC-Net methods in tandem, the information gleaned from each approach individually about the structure of the gene regulatory network can be augmented. For example, gene-to-gene interactions identified by the EBDBN method with rigid approximate posterior distributions from the ABC-Net method appear to be supported by fairly substantial evidence, as those particular interactions are restricted to a smaller range of values in

their posterior distributions. On the other hand, gene-to-gene interactions identified by the EBDBN method that are associated with flexible approximate posterior distributions may indeed represent false positives, as those parameters may take on a wider range of values without negatively impacting the proximity of simulated and observed data in the ABC-Net method. In this way, the distinctive results of the EBDBN and ABC-Net methods are able to yield complementary information about specific gene-to-gene interactions within the S.O.S. DNA repair system, as well as the overall dynamics of the full biological system.

4 Summary

Reverse engineering the structure of gene regulatory networks from longitudinal expression data is an intrinsically difficult task, given the complexity of network topology within biological systems, the large number of potential gene-to-gene interactions in typical networks, and the small number of replicates and time points available in real data. In this paper, we discussed two approximate Bayesian methods to reverse engineer regulatory networks from time-course gene expression data. The two approaches, while not comparable, are complementary, and help illustrate the need for a variety of network inference methods adapted for different contexts. The first proposed approach, known as the EBDBN method, is based on an empirical Bayes estimation procedure using a linear Gaussian state space model. In the second approach, known as the ABC-Net approach, we apply a simulation-based Bayesian method to conduct a detailed analysis of small, well-characterized pathways under less restrictive model assumptions.

Continuing to improve knowledge of gene regulatory networks is an important goal in agricultural studies, as gene regulatory networks are implicated in the coordination of genes underlying complex traits, including those that may be of economic value. Although the example in this paper of the S.O.S. DNA repair system in *Escherichia coli* is based on a very simple model organism, the methods illustrated in this paper could be adapted to the analysis of agriculturally relevant networks. For example, in the model plant *Arabidopsis thaliana*, gene regulatory networks are known to be involved in starch metabolism during the diurnal cycle (Smith et al., 2004; Opgen-Rhein and Strimmer, 2007), dehydration stress tolerance (Shinozaki and Yamaguchi-Shinozaki, 2006), flowering time control (Welch et al., 2003), cold acclimation (Chawade et al., 2007), and nitrogen response affecting growth and development (Gifford et al., 2006). As time-course studies of such networks become increasingly cost-effective, we anticipate that the reverse engineering approaches presented in this paper will provide greater insight into the complicated interactions occurring among genes in real biological systems. In addition, we can anticipate rapid progress in this field as state space models can greatly benefit from recent simulation techniques such as particle-filtering and Sequential Monte Carlo (Del Moral et al., 2010) which make more flexible versions of these models (non-linearity, discrete states) computationally feasible (Toni and Stumpf, 2010).

5 Acknowledgements

We gratefully acknowledge helpful discussions with Alan Qi, Bruce Craig, Jayanta Ghosh and Gayla Olbricht. We also thank My Truong and Doug Crabill for their computing expertise, the RWD research group for their support, as well as the Department of Statistics at Kansas State University and the Women in Science Program at Purdue University for providing funding to AR. This work is supported by a NSF Plant Genome grant (DBI-0733857) in part to RWD, and by funding from INRA AgroParisTech, Department of Animal Genetics and Integrative Biology to AR.

References

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Genetics Reviews 8*, 450–461.

Arbeitman, M. N., E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science 297*, 2270–2275.

Beal, M. J., F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics 21*(3), 349–356.

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin, and C. P. Robert (2009). Adaptivity for ABC algorithms: the ABC-PMC. *Biometrika 96*(4), 983–990.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*, 2025–2035.

Brandman, O. and T. Meyer (2008). Feedback loops shape cellular signals in space and time. *Science 322*, 390–395.

Bremer, M. and R. W. Doerge (2009). The KM-algorithm identifies regulated genes in time series expression data. *Advances in Bioinformatics 2009*, 284251.

Bremer, M. M. (2006). *Identifying Regulated Genes Through the Correlation Structure of Time Dependent Microarray Data*. Ph.D. dissertation, Purdue University, West Lafayette, IN USA.

Carlin, B. P. and T. A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.). Chapman and Hall/CRC.

Chawade, A., Bräutigam, A. Lindlöf, O. Olsson, and B. Olsson (2007). Putative cold acclimation pathways in *Arabidopsis thaliana* identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors. *BMC Genomics 8*(304).

Del Moral, P., A. Doucet, and A. Jasra (2010). Sequential Monte Carlo for Bayesian computation. *Bayesian Statistics 8*, 1–34.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological) 39*(1), 1–38.

Enders, W. (2004). *Applied Econometric Time Series Analysis*. Wiley.

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

Gifford, M. L., R. A. Gutiérrez, and G. M. Coruzzi (2006). *A Companion to Plant Physiology*, Chapter "Modeling the virutal plant: a systems approach to nitrogen-regulatory gene networks", pp. Essay 12.2. Sinauer.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Husmeier, D., R. Dybowski, and S. Roberts (Eds.) (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering 82*, 35–45.

Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology 4*(213).

Leuenberger, C. and D. Wegmann (2009). Bayesian computation and model selection without likelihoods. *Genetics 183*, 1–10.

Lipschutz, R. J., S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart (1999). High density synthetic oligonucleotide arrays. *Nature Genetics 21*, 20–24.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Reviews of Genomics and Human Genetics 9*, 387–402.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain Monte Carlo without likelihoods. *PNAS 100*(26), 15324–15328.

Opgen-Rhein, R. and K. Strimmer (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics 8(Suppl 2)*.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezann, and M. W. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*, 1791–1798.

Ratmann, O., O. Jorgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Computational Biology 3*(11), 2266–2278.

Rau, A. (2010). *Reverse Engineering Gene Networks Using Genomic Time-Course Data*. Ph.D. dissertation, Purdue University, West Lafayette, IN USA.

Rau, A., F. Jaffrézic, J.-L. Foulley, and R. W. Doerge (2010). An empirical Bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology 9*(9), 1–28.

Ronen, M., R. Rosenberg, B. I. Shraiman, and U. Alon (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *PNAS 99*(16), 10555–10560.

Schena, M., D. Schalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*(5234), 467–470.

Schena, M., S. J. Smith, and P. O. Brown (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research 6*(7), 639–645.

Schlitt, T. and A. Brazma (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics 8(Suppl 6)*(S9), 1–22.

Shinozaki, K. and K. Yamaguchi-Shinozaki (2006). Gene networks involved in drought stress response and tolerance. *Journal of Experimental Biology 58*(2), 221–227.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007). Sequential Monte Carlo without likelihoods. *PNAS 104*, 1760–1765.

Smith, S. M., D. C. Fulton, T. Chia, D. Thorneycroft, A. Chapple, H. Dunstan, C. Hylton, S. C. Zeeman, and A. M. Smith (2004). Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiology 136*(1), 2687–2699.

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

Tegnér, J., M. K. S. Yeung, J. Hasty, and J. J. Collins (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamcal modeling. *Genetics 100*(10), 59445949.

Toni, T. and M. P. H. Stumpf (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics 26*(1), 104–110.

Welch, S. M., J. L. Roe, and Z. Dong (2003). A genetic neural network model of flowering time control in *Arabidopsis thaliana*. *Agronomy Journal 95*, 71–81.

Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics 10*, 122–133.