CHAPTER 27

BIOLOGICAL NETWORK INFERENCE AT MULTIPLE SCALES: FROM GENE REGULATION TO SPECIES INTERACTIONS

Andrej Aderhold[1], V Anne Smith[1], and Dirk Husmeier[2]

[1]*School of Biology, University of St Andrews, St Andrews, UK*
[2]*School of Mathematics and Statistics, University of Glasgow, Glasgow, UK*

27.1 INTRODUCTION

Mathematics is transforming biology in the same way it shaped physics in the previous centuries [11]. The underlying paradigm shift that distinguishes modern quantitative systems biology from more traditional nonquantitative approaches is based on a conceptualization of molecular reactions in the cell, the elementary building block of life, as a complex network of interactions. The approach is intrinsically holistic, aiming to understand the properties of cells, tissues, and organisms functioning as a complex system. Besides aiming for a deeper theoretical understanding of molecular processes and their emergent properties, modern systems biology sees a huge range of potential applications, ranging from the targeted genetic modifications of plants for improved resistance, yield, and a variety of agronomically desired traits [54], to unraveling the causes of neurodegenerative diseases, cancer, and ultimately aging [40, 58].

Pattern Recognition in Computational Molecular Biology: Techniques and Approaches,
First Edition. Edited by Mourad Elloumi, Costas S. Iliopoulos, Jason T. L. Wang, and Albert Y. Zomaya.

The new challenge for systems biology is to encompass all the facets of biology. For instance, from the molecular scale of gene regulatory networks to organ-level fluid pressure and velocity fields, or from chemotaxis of cancer cells during invasion and metastasis, to mass movement and migration patterns in locusts and wildebeest. The ultimate quest is the elucidation of the common principles spanning several spatial and temporal scales. The following are important questions to be addressed: Which common mechanisms determine both aberrant behavior in groups of migrating animals and the movement of cancer cells in the human body? Which organizational principles are common to the response of eukaryotic gene regulatory networks to environmental stress and the response of trophic species interaction networks to climate change? How can mathematical models of biochemical signal transduction pathways in plants link cell-level regulation to metabolism, biomass production, and properties at whole organism level?

The emphasis of the present chapter is to focus on the modeling commonalities between systems biology at the molecular and species levels by investigating the application of computational methods from molecular systems biology to whole ecosystems. As with molecular regulatory networks, the complexity of ecological networks is staggering, with hundreds or thousands of species interacting in multiple ways, from competition and predation to commensalism (whereby one species profits from the presence of another) and mutualism (where two species exist in a relationship in which each benefits from the other). Understanding the networks that form ecosystems is of growing importance, for example, to understand how a change in one population can lead to dramatic effects on others [12], drive them to alternative stable states [6], or even cause catastrophic failure [51]. Ecologists need to understand how populations and ecosystems as a whole respond to these changes. This is of enormous importance during a period of rapid climate change [53] that affects not only land use and agriculture [39], but can also cause a reduction in biodiversity in terrestrial and aquatic ecosystems [47].

Inferring the interactions in complex ecosystems is not a straightforward task to accomplish. Direct observation requires minute observations and detailed fieldwork, and is capable of measuring only certain types of species interactions, such as those between predators and their prey, or between pollinators and their host plants. The majority of interactions are not directly observable. This includes competition for resources, commensalism, or mutualism. This restriction calls for the development of novel computational inference techniques to learn networks of species interactions directly from observed species concentrations.

Network inference can be generally defined as the process of "reverse engineering," or learning, the interactions between components of a system given its observed data. Interactions play such an important role because a system's behavior and that of its parts is defined not only by external factors but also by internal influences represented by the components of the system itself. In the context of gene regulation networks, this involves the control of gene expression through regulation factors, such as proteins or microRNAs [31]. However, these underlying processes are often hidden

from direct observation, causing poor understanding or complete lack of knowledge of how the components interact. The challenge arises with the fact that observable quantities of a system have to be sufficient to be used as a guide to identify the driving architecture, that is, the interaction networks that cause its behavior and observed characteristics. Previous knowledge about the system can help to improve network inference. However, in many instances little or no knowledge is available and patterns have to be inferred directly from the observed data. One such pattern is the gene regulation network, which is responsible for the control of the majority of molecular processes essential for growth and survival of any organism on earth. This can involve the control of organism development [5], response of the immune system to pathogens [46], or the adaptation to changing environmental conditions through stress responses [50].

In the last decade, molecular biology has been the driving force for the development of inference methods that help to identify such patterns. This becomes more important in light of the growing amount of biomolecular data from high-throughput techniques, such as DNA microarray experiments [49]. Ecological modeling could benefit from methods used in systems biology. Several commonalities between molecular and ecological systems exist: Genes are the basic building blocks of gene regulation networks and can be compared to organisms as principal participants in the formation of ecological networks; gene profile measurements from DNA or mRNA assays can be compared to population data gathered in field surveys; expression profiles of genes or proteins can be matched with population densities or species coverage; gene regulation compares to species interactions, and different conditions compare to different environments. It seems natural to apply the same methodology with certain modification to both the molecular and ecological data.

In this chapter, we show how established methods from systems biology (Section 27.2) can be modified to infer ecological networks (Section 27.3). The structure is in the following form: Section 27.4 introduces the mathematical notation (Section 27.4.1), state-of-the art regression methods (Sections 27.4.2, and 27.4.3), and a scoring metric to measure reconstruction accuracy (Section 27.4.4). Section 27.5 demonstrates an application to a circadian clock regulation system in the plant *Arabidopsis thaliana* involving simulated data and *Real-Time Polymerase Chain Reaction* (rtPCR) gene expression data (Sections 27.5.3 and 27.5.4), and the modification of a basic regression method to enable the handling of dynamic gene regulatory changes over time. Section 27.6 describes the method modifications that allow us to apply the previously defined methods to an ecological problem setting. This is realized with the expansion of the data domain from one-dimensional time to two dimensions of space. In addition, methods that can learn the spatial segmentation on a global scale (Section 27.6.2) and local scale using the *Mondrian process* (Section 27.6.3) are described. The learning performance of these methods is compared on synthetic data and realistic ecological simulated data (Sections 27.6.4 and 27.6.6). Finally, network reconstruction is demonstrated on a real-world case using plant coverage and environmental data (Section 27.6.6).

27.2 MOLECULAR SYSTEMS

The inference of molecular regulatory networks from postgenomic data has been a central topic in computational systems biology for over a decade. Following up on the seminal paper by Friedman et al. [21], a variety of methods have been proposed [59], and several procedures have been pursued to objectively assess the network reconstruction accuracy [32, 59, 60], for example, of the RAF kinase pathway, which is a cellular signaling network in human immune system cells [46]. It has been demonstrated that machine learning techniques can not only serve to broaden our biological knowledge [14, 19] but also handle the increasing amount of data from high-throughput measurements in a more efficient way than was previously attempted [29]. We describe four state-of-the art methods that are widely used for the inference of gene regulation networks. The accuracy of inference is assessed with a benchmark of a realistic gene and protein regulation system in the context of circadian regulation. The data are simulated from a recently published regulatory network of the circadian clock in *A. thaliana*, in which protein and gene interactions are described by a Markov jump process based on Michaelis–Menten kinetics. We closely follow recent experimental protocols, including the entrainment of seedlings to different *Light–Dark* (LD) cycles and the knockout of various key regulatory genes. Our study provides relative assessment scores for the comparison of the presented methods and investigates the influence of systematically missing values related to unknown protein concentrations and mRNA transcription rates.

27.3 ECOLOGICAL SYSTEMS

While interaction networks at the molecular level have been the forefront of modern biology, due to the ever increasing amount of available postgenomic data, interaction networks at other scales are drawing attention. This concerns, in particular, ecological networks, owing to their connection with climate change and biodiversity, which poses new challenges and opportunities for machine learning and computational statistics. Similar to molecular systems, ecological systems are complex dynamical systems with interconnected networks of interactions among species and abiotic factors of the environment. This interconnectedness can lead to seemingly unpredictable behavior: changing numbers of one species can influence unexpected changes in others [30]; the whole system can transit between different stable states [6]. Perturbations from features such as climate change and invasive species can affect both biodiversity and the stability of ecosystems [48]. Being able to make predictions on these scales requires an understanding of the ecological networks underlying the system [15].

The challenges for computational inference specific to ecological systems are that, first, the interactions take place in a spatially explicit environment which must be taken into account, and second, the interactions can vary across this environment depending on the makeup of the elements (species and abiotic factors) present. Here, we meet these challenges by showing the necessary modifications to an inference

method from systems biology [35] for temporally explicit (one-dimensional) gene expression data to infer ecological interactions from spatially explicit species abundance data on a two-dimensional grid. We describe a nonhomogeneous Bayesian regression model based on the Bayesian hierarchical regression model of Andrieu and Doucet [4], using a multiple global *change-point* process as proposed in Reference [1]. We modify the latter method with a *Mondrian process* that implements a spatial partitioning at different levels of resolution [2]. We make further use of the spatially explicit nature of ecological data by correcting for spatial autocorrelation with a regulator node (in Bayesian network terminology) that explicitly represents the spatial neighborhood of a node. The performance of these methods is demonstrated on synthetic and realistic simulated data and we infer a network from a real world data set. The results show that ecological modeling could benefit from these types of methods, and that the required modifications do not conflict with, but extend the basic methodology used in systems biology.

27.4 MODELS AND EVALUATION

This section introduces the notations used throughout this chapter, two well-established sparse regression methods, and the basic framework of a homogeneous Bayesian regression model that is applied both to infer gene regulation networks and ecological networks. The naming *homogeneous* implies that the data are treated as a single monolithic block. This simplifies inference but insufficiently reflects the actual nature of underlying biological processes, which can change over time and space. For instance, morphological changes of an insect with the distinct phases of an embryo, larva, pupa, and adult [5] are matched by changes in gene expression profiles and interconnectedness, for example, regulation through transcription factors that are proteins. A heterogeneous model allows the partitioning of data to account for different phases and associated changes in the interaction networks and network parameters. Section 27.5.1 describes this model with the possibility to set fixed phase boundaries (so-called *change-points*) and learn model parameters for each of the phases. This technique is extended to a segmentation in two dimensions with an additional inference scheme for the *change-points* in the case that the segmentation of the data is not known (Section 27.6).

27.4.1 Notations

The following notations are used throughout this chapter for both gene regulation and species interaction. For the regression models, which we use to infer the network interactions, we have target variables y_n ($n = 1, \ldots, N$) that can each represent a temporal mRNA concentration gradient of a particular gene n or the abundance of a particular species also denoted as n. The realizations of each target variable y_n can then be written as a vector $\mathbf{y}_n = (y_{n,1}, \ldots, y_{n,T})^T$, where $y_{n,t}$ is the realization of y_n in observation t. The potential regulators are either gene or protein concentrations in the case of gene regulation networks or species abundances for species interaction

networks. The task is to infer a set of regulators π_n for each response variable y_n. The collective set of regulators $\{\pi_1, \ldots, \pi_N\}$ defines a regulatory interaction network, $\mathcal{G}$. In $\mathcal{G}$, the regulators and the target variables represent the nodes, and from each regulator in π_n, a directed interaction (or *edge*) points to the target node n, attributed with the interaction strength vector, or regression coefficients, $\mathbf{w}_n = (w_n^p)_{p\in\pi_n}$. The complete set of regulatory observations is contained in the design matrix $\mathbf{X}$. Realizations of the regulators in the set π_n are collected in $\mathbf{X}_{\pi_n}$, where the columns of $\mathbf{X}_{\pi_n}$ are the realizations of the regulators π_n. Design matrix $\mathbf{X}$ and $\mathbf{X}_{\pi_n}$ are extended by a constant element equal to 1 for the intercept.

27.4.2 Sparse Regression and the LASSO

A widely applied linear regression method that encourages network sparsity is the *Least Absolute Shrinkage and Selection Operator* (LASSO) introduced in Reference [55]. The LASSO optimizes the parameters of a linear model based on the residual sum of squares subject to an $L1$-norm penalty constraint on the regression parameters, $\|\mathbf{w}_n\|_1$, which excludes the intercept [20]:

$$\hat{\mathbf{w}}_n = \text{argmin}\left\{\|\mathbf{y}_n - \mathbf{X}^T\mathbf{w}_n\|_2^2 + \lambda_1\|\mathbf{w}_n\|_1\right\} \tag{27.1}$$

where λ_1 is a regularization parameter controlling the strength of shrinkage. Equation 27.1 constitutes a convex optimization problem, with a solution that tends to be sparse. Two disadvantages of the LASSO are arbitrary selection of single predictors from a group of highly correlated variables and saturation at T predictor variables. To avoid these problems, the *Elastic Net* method was proposed in Reference [63], which combines the LASSO penalty with a ridge regression penalty of the standard squared $L2$-norm $\|\mathbf{w}_n\|_2^2$ excluding the intercept:

$$\hat{\mathbf{w}}_n = \text{argmin}\left\{\|\mathbf{y}_n - \mathbf{X}^T\mathbf{w}_n\|_2^2 + \lambda_1\|\mathbf{w}_n\|_1 + \lambda_2\|\mathbf{w}_n\|_2^2\right\} \tag{27.2}$$

Similar to Equation 27.1, Equation 27.2 constitutes a convex optimization problem, which we solve with cyclical coordinate descent [20] implemented in the R software package *glmnet*. The regularization parameters λ_1 and λ_2 were optimized by 10-fold cross-validation.

27.4.3 Bayesian Regression

We follow Reference [25] in the definition of the Bayesian regression model and will further refer to it as "homogBR". It is assumed to be a linear regression model for the targets:

$$\mathbf{y}_n|(\mathbf{w}_n, \sigma_n, \pi_n) \sim \mathcal{N}(\mathbf{X}_{\pi_n}^{\mathrm{T}}\mathbf{w}_n, \sigma_n^2\mathbf{I}) \tag{27.3}$$

where σ_n^2 is the noise variance, and $\mathbf{w}_n$ is the vector of regression parameters, for which we impose a Gaussian prior:

$$\mathbf{w}_n|(\sigma_n, \delta_n, \pi_n) \sim \mathcal{N}(\mathbf{0}, \delta_n\sigma_n^2\mathbf{I}) \tag{27.4}$$

δ_n can be interpreted as the *Signal-to-Noise Ratios* (SNR) [25]. For the posterior distribution, we get

$$\mathbf{w}_n|(\sigma_n, \delta_n, \pi_n, \mathbf{y}_n) \sim \mathcal{N}(\mathbf{\Sigma}_n \mathbf{X}_{\pi_n} \mathbf{y}_n, \sigma_n^2 \mathbf{\Sigma}_n) \tag{27.5}$$

where $\mathbf{\Sigma}_n^{-1} = \delta_n^{-1}\mathbf{I} + \mathbf{X}_{\pi_n}\mathbf{X}_{\pi_n}^{\mathrm{T}}$, and the marginal likelihood can be obtained by application of standard results for Gaussian integrals [7]:

$$\mathbf{y}_n|(\sigma_n, \delta_n, \pi_n) \sim \mathcal{N}(\mathbf{0}, \sigma_n^2(\mathbf{I} + \delta_n \mathbf{X}_{\pi_n}^{\mathrm{T}} \mathbf{X}_{\pi_n})) \tag{27.6}$$

For σ_n^{-2} and δ_n^{-2} we impose conjugate gamma priors, $\sigma_n^{-2} \sim \mathrm{Gam}(\nu/2, \nu/2)$, and $\delta_n^{-1} \sim \mathrm{Gam}(\alpha_\delta, \beta_\delta)$.[1] The integral resulting from the marginalization over σ_n^{-2},

$$P(\mathbf{y}_n|\pi_n, \delta_n) = \int_0^\infty P(\mathbf{y}_n|\sigma_n, \delta_n, \pi_n) P(\sigma_n^{-2}|\nu) d\sigma_n^{-2}$$

is then a multivariate Student t-distribution with a closed-from solution [7, 25]. Given the data for the potential regulators of y_n, symbolically y, the objective is to infer the set of regulators π_n from the marginal posterior distribution:

$$P(\pi_n|y, \mathbf{y}_n, \delta_n) = \frac{P(\pi_n)P(\mathbf{y}_n|\pi_n, \delta_n)}{\sum_{\pi_n^\star} P(\pi_n^\star)P(\mathbf{y}_n|\pi_n^\star, \delta_n)} \tag{27.7}$$

where the sum is over all valid regulator sets $\pi_n^\star$, $P(\pi_n)$ is a uniform distribution over all regulator sets subject to a maximal cardinality, $|\pi_n| \leq 3$, and δ_n is a nuisance parameter, which can be marginalized over. We sample sets of regulators π_n, signal-to-noise hyperparameter δ_n, and noise variances σ_n^2 from the joint posterior distribution with *Markov Chain Monte Carlo* (MCMC), following a Metropolis–Hastings within a partially collapsed Gibbs scheme [25].

27.4.4 Evaluation Metric

The previously described methods provide a means by which interactions can be ranked in terms of their significance or influence. If the true network is known, this ranking defines the *Receiver Operating Characteristic* (ROC) curve as shown in Figure 27.1, where for all possible threshold values, the *sensitivity* or recall is plotted against the complementary *specificity*.[2] By numerical integration we then obtain the *Area Under the ROC* curve (AUROC) as a global measure of network reconstruction accuracy, where larger values indicate a better performance, starting from AUROC = 0.5 to indicate random expectation, to AUROC = 1 for perfect network reconstruction. There have been suggestions that the *Area Under Precision-RECall*

[1] We set: $\nu = 0.01$, $\alpha_\delta = 2$, and $\beta_\delta = 0.2$, as in Reference [25].

[2] The *sensitivity* is the proportion of true interactions that have been detected, the *specificity* is the proportion of noninteractions that have been avoided.

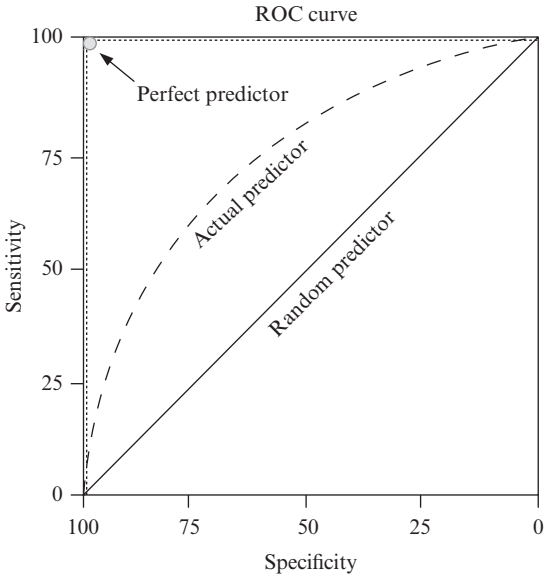

Figure 27.1 *Receiver Operating Characteristic* (ROC). An ROC curve for a perfect predictor, random expectation, and a typical predictor between these two extremes is shown. The *Area Under the ROC* curve (AUROC) is used as scoring metric.

curves (AUPREC) indicate differences in network reconstruction performance more clearly than AUROC curves [13]. While this is true for large, genome-wide networks, a study in Reference [26] has indicated that in networks with a low number of nodes, as with the studied networks in Figure 27.3, the difference between the two scoring schemes should be negligible. We therefore evaluate the performance of all methods with AUROC scores, due to their more straightforward statistical interpretation [28].

27.5 LEARNING GENE REGULATION NETWORKS

A typical feature of molecular regulation networks is the variability of interactions among genes and proteins that can occur over time, caused by changing internal and external conditions. The homogeneous Bayesian regression model described in Section 27.4.3 is unable to pick up such time-varying relationships because it treats all observations as coming from the same underlying regulatory process. This section describes a nonhomogeneous Bayesian regression model with a fixed *change-point* model that is able to approximate such nonhomogeneity. In the species interaction study (Section 27.6) we show that this model can be adapted to a spatial *change-point* process in two dimensions. In addition, protein–gene interactions affect transcription rates, but both these rates as well as protein concentrations might not be available from the wetlab assays. In such situations, mRNA concentrations have to be taken as

proxy for protein concentrations, and rates have to be approximated by finite difference quotient or analytic derivation (Section 27.5.2). We compare these setups on a simulated data set (Section 27.5.3) and select the most realistic setup with the best performing method for a real-world data application (Section 27.5.4). The content in this chapter is in part based on a recent study [3] on circadian regulation.

27.5.1 Nonhomogeneous Bayesian Regression

Underlying regulatory relationships can vary over time, as they are dependent on external and internal influences, for example, changing environmental conditions such as light and dark (e.g., circadian regulation in plants) or different stages of development (e.g., from embryo to adult). This implies nonhomogeneous dynamics over time that can be modeled by introducing so-called *change-point*s into the regression model. *Change-points* partition the observations on the timeline into individual segments that are combined to a piecewise linear model that approximates nonhomogeneity. We follow Reference [25] and extend the Bayesian regression model from Section 27.4.3 with a multiple *change-point* process and refer to it with the name "nonhomogBR". The *change-point* process imposes a set of $H_n - 1$ *change-points*, $\{\tau_{n,h}\}_{1\leq h\leq (H_n-1)}$ with $\tau_{n,h} < \tau_{n,h+1}$, to divide the temporal observations of a variable into H_n disjunct segments. With the two *pseudo-change-points* $\tau_{n,0} := 1$ and $\tau_{n,H_n} := T$ each segment $h \in \{1, \ldots, H_n\}$ is defined by two demarcating *change-points*, $\tau_{n,h}$ and $\tau_{n,h+1}$. The vector of the target variable realizations, $\mathbf{y}_n = (y_{n,1}, \ldots, y_{n,T})^{\mathrm{T}}$, is thus divided into H_n subvectors, $\{\mathbf{y}_{n,h}\}_{h=1,\ldots,H_n}$, where each subvector corresponds to a temporal segment: $\mathbf{y}_{n,h} = (y_{n,(\tau_{n,h}+1)}, \ldots, y_{n,\tau_{n,h+1}})^{\mathrm{T}}$. Following Reference [25], the distances between two successive *change-points*, $T_{n,h} = \tau_{n,h+1} - \tau_{n,h}$, are assumed to have a negative binomial distribution, symbolically $T_{n,h} \sim NBIN(p, k)$.

We keep the regulator set, π_n, fixed among the H_n segments[3] and we apply the linear Gaussian regression model, defined in Equation 27.3, to each segment h:

$$\mathbf{y}_{n,h}|(\mathbf{X}_{\pi_n,h}, \mathbf{w}_{n,h}, \sigma_n^2) \sim \mathcal{N}(\mathbf{X}_{\pi_n,h}^T \mathbf{w}_{n,h}, \sigma_n^2 \mathbf{I}) \tag{27.8}$$

where $\mathbf{X}_{\pi_n,h}$ is the segment-specific design matrix, which can be built from the realizations of the regulator set π_n in segment h, and $\mathbf{w}_{n,h}$ is the vector of the segment-specific regression parameters for segment h. As in Section 27.4.3, we impose an inverse Gamma prior on σ_n^2, symbolically $\sigma_n^{-2} \sim \mathrm{Gam}(\nu/2, \nu/2)$. For the segment-specific regression parameters, $\mathbf{w}_{n,h}$ $(h = 1, \ldots, H_n)$, we assume Gaussian priors:

$$\mathbf{w}_{n,h}|(\sigma_n, \delta_n, \mathbf{X}_{\pi_n,h}) \sim \mathcal{N}(0, \delta_n \sigma_n^2 \mathbf{I}) \tag{27.9}$$

with the hyperprior $\delta_n^{-1} \sim \mathrm{Gam}(A_\delta, B_\delta)$.

[3]The regulator set, that is, network structure, can differ between segments as shown in Reference [35]. However, here we limit the model assumption to a single network to decrease model complexity and based on the fact that interactions are more likely to change than to disappear.

As with the previously defined homogeneous Bayesian regression model (Section 27.4.3), posterior inference is again carried out with the Metropolis–Hastings within partially collapsed Gibbs sampling scheme [25]. The marginal likelihood in Equation 27.7 has to be replaced by

$$P(\pi_n|\mathbf{X}, \delta_n, \{\tau_{n,h}\}_{1\leq h\leq (H_n-1)}) \propto P(\pi_n) \prod_{h=1}^{H_n} P(\mathbf{y}_{n,h}|\mathbf{X}_{\pi_n,h}, \delta_n) \tag{27.10}$$

where $P(\mathbf{y}_{n,h}|\mathbf{X}_{\pi_n,h}, \delta_n)$ $(h = 1, \ldots, H_n)$ can be computed in closed-form; see Reference [25] for a mathematical derivation. The full conditional distribution of $\mathbf{w}_{n,h}$ is now given by [25]:

$$\mathbf{w}_{n,h}|(\mathbf{y}_{n,h}, \mathbf{X}_{\pi_n,h}, \sigma_n^2, \delta_n) \sim \mathcal{N}(\tilde{\mathbf{m}}_{n,h}, \sigma_n^2\boldsymbol{\Sigma}_{n,h}) \tag{27.11}$$

with $\boldsymbol{\Sigma}_{n,h}^{-1} = \delta_n^{-1}\mathbf{I} + \mathbf{X}_{\pi_n,h}\mathbf{X}_{\pi_n,h}^T$, and estimated mean $\tilde{\mathbf{m}}_{n,h} = \boldsymbol{\Sigma}_{n,h}\mathbf{X}_{\pi_n,h}\mathbf{y}_{n,h}$.

27.5.2 Gradient Estimation

The machine learning and statistical models applied in our study predict the rate of gene transcription from the concentrations of the putative regulators. With *de novo* mRNA profiling assays, the rate of transcription could in principle be measured, but these data are often not available. We therefore applied a numerical and analytical procedure to obtain the transcription rates. Appreciating that the transcription rate is just the time derivative of the mRNA concentration $c(t)$, the first approach is to approximate it by a difference quotient:

$$\frac{dc}{dt} \approx \frac{c(t+\delta t) - c(t-\delta t)}{2\delta t} \tag{27.12}$$

Although this is a straightforward procedure, it is well known that differencing noisy time series leads to noise amplification. As an additional procedure, an approach based on smooth interpolation with Gaussian processes is used. We follow Reference [52] and exploit the fact that the derivative of a Gaussian process is a Gaussian process again; hence analytic expressions for the mean and the standard deviation of the derivative are available (see Reference [52]). For the covariance of the Gaussian process, we used the squared exponential kernel, which is the standard default setting in the R package *gptk* [33].

27.5.3 Simulated Bio-PEPA Data

We generated data from the central circadian gene regulatory network in *A. thaliana*, as proposed in Reference [27] and depicted in Figure 27.3(a). Following Reference [61], the regulatory processes of transcriptional regulation and posttranslational protein modification were described with a Markov jump process based on Michaelis–Menten kinetics, which defines how mRNA and protein concentrations

change in dependence on the concentrations of other interacting components in the system (see Appendix of Reference [27] for detailed equations). We simulated mRNA and protein concentration time courses with the Gillespie algorithm [23], using the Bio-PEPA modeling framework [10]. We created 11 interventions in consistency with standard biological protocols [16]. These include knockouts of proteins GI, LHY, PRR7/PRR9, TOC1, and varying photo-periods of 4, 6, 8, 12, or 18 h of light in a 24-h LD cycle. For each intervention, we simulated protein and mRNA concentration time courses over 6 days. The first 5 days served as entrainment to the indicated LD cycles. This was followed by a day of persistent *Darkness* (DD) or *light* (LL), during which concentrations of mRNAs and proteins were measured in 2-h intervals. Combining 13 observations for each intervention yielded 143 observations in total for each network type. All concentrations were standardized to unit standard deviation. The temporal mRNA concentration gradient was approximated by the two alternative schemes described in Section 27.5.2: a difference quotient (Eq. 27.12) of mRNA concentrations with -2 and $+2$ h[4] (termed *coarse gradient*), and smooth interpolation with a Gaussian process yielding a derivative (gradient) at each observation time point (termed *interp gradient*), followed by z-score standardization.

When trying to reconstruct the regulatory network from the simulated data, we ruled out self-loops, such as from LHY (modified) protein to LHY mRNA, and adjusted for mRNA degradation by enforcing mRNA self-loops, such as from the LHY mRNA back to itself. Protein ZTL was included in the stochastic simulations, but excluded from structure learning because it has no direct effect on transcription. We carried out two different network reconstruction tasks. The first was based on complete observation, including both protein and mRNA concentration time series. The second was based on incomplete observation, where only mRNA concentrations were available, but protein concentrations were systematically missing. All network reconstructions were repeated on five independent data sets.

27.5.4 Real mRNA Expression Profile Data

In addition to the realistic data simulated from a faithful mathematical description of the molecular interaction processes, as described earlier, we used real transcription profiles for the key circadian regulatory genes in the model plant *A. thaliana*. The objective is to infer putative gene regulatory networks with the statistical methods described in Section 27.4, and then to compare these predictions with network models of the circadian clock from the biological literature [38, 41–43], of which two are displayed in Figure 27.3(a) and (b). It is important to note that, as opposed to the realistic data described in the previous subsection, we do not have a proper biological understanding. Besides the fact that the models in References [38, 41–43] show noticeable variations, they were not obtained on the basis of proper statistical

[4]Two-hour intervals are a realistic sampling frequency for plants in the wetlab, although, smaller intervals would be favorable.

model selection, as described, for example, in Reference [57]. Nevertheless, a qualitative comparison will reveal to what extent the postulated interaction features and structural network characteristics from the literature are consistent with those inferred from the data.

The data used in our study come from the EU project TiMet [56], whose objective is the elucidation of the interaction between circadian regulation and metabolism in plants. The data consist of transcription profiles for the core clock genes from the leaves of various genetic variants of *A. thaliana*, measured with rtPCR. The study encompasses two wildtypes of the strains Columbia (Col-0) and Wasilewski (WS) and four clock mutants, namely, a double knockout LHY/CCA1 in the WS strain, a single knockout of GI and TOC1 in the strain Col-0, and a double knockout PRR7/PRR9 in strain Col-0. The plants were grown in the following three light conditions: a diurnal cycle with 12 h light and 12 h darkness (12L/12D), an extended night with full darkness for 24 h (DD), and an extended light with constant light (LL) for 48 h. Samples were taken every 2 h to measure mRNA concentrations. Further information on the data and the experimental protocols is available from Reference [18]. The mRNA profiles for the genes LHY, CCA1, PRR5, PRR7, PRR9, TOC1, ELF3, ELF4, LUX, and GI were extracted from the TiMet data [18], yielding a total of 288 samples per gene. We used the log mean copy number of mRNA per cell and applied a gene-wise z-score transformation for data standardization. An additional binary light indicator variable with 0 for darkness and 1 for light was included to indicate the status of the experimentally controlled light condition.

27.5.5 Method Evaluation and Learned Networks

We evaluate the reconstruction accuracy with AUROC scores (Section 27.4.4) of LASSO, *Elastic Net* (Section 27.4.2), the homogeneous Bayesian regression model (homogBR, Section 27.4.3), and the nonhomogeneous Bayesian regression model (nonhomogBR, Section 27.5.1). The method with the highest score was applied to the real-world data from the previous section. Since light may have a substantial effect on the regulatory relationships of the circadian clock, we set the *change-points* of the nonhomogBR method according to the day/night transitions yielding two segments, $h = 1$ (light) and $h = 2$ (darkness). This reflects the nature of the laboratory experiments, where *A. thaliana* seedlings are grown in an artificial light chamber whose light is switched on or off. It would be straightforward to generalize this approach to more than two segments to allow for extended dawn and dusk periods in natural light. Given that the light phase is known, we consider the segmentation as fixed.

27.5.5.1 Simulated Data. The marginal posterior probabilities of all potential interactions are computed for the Bayesian regression methods. For LASSO and *Elastic Net*, we record the absolute values of nonzero regression parameters. Both measures provide a means by which interactions between genes and proteins can be ranked in terms of their significance or influence. The AUROC scores given these values are calculated with the reference network *P2010* shown in Figure 27.3(a).

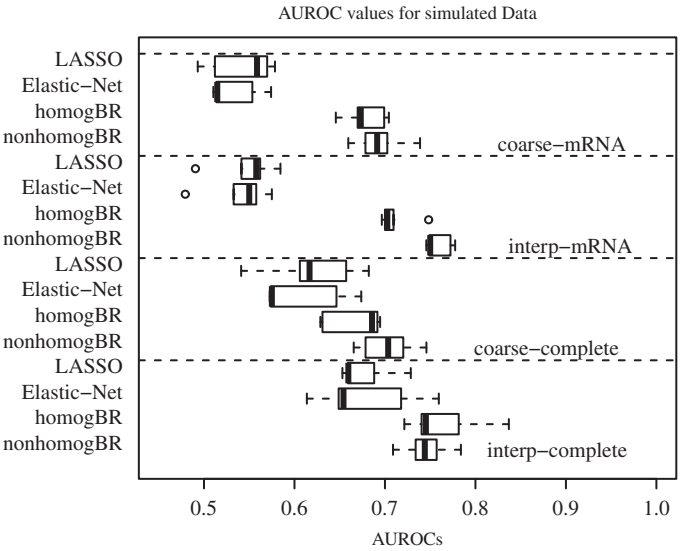

Figure 27.2 AUROC scores obtained for different reconstruction methods, and different experimental settings. Boxplots of AUROC scores obtained from LASSO, *Elastic Net* (both in Section 27.4.2), homogBR (homogeneous Bayesian regression, Section 27.4.3), and nonhomogBR (nonhomogeneous Bayesian regression, Section 27.5.1). The latter utilizes light- induced partitioning of the observations. The subpanels are *coarse-mRNA*: incomplete data, only with mRNA concentrations and coarse gradient, *interp-mRNA*: incomplete data with interpolated gradient, *coarse-complete*: complete data with protein and mRNA concentrations, and *interp-complete*: complete data with interpolated gradient. The *coarse* gradients are computed from Equation 27.12 from 4-h intervals, and the interpolated gradients (*interp*) are derived from a Gaussian process as described in Section 27.5.2.

They can assume values from around AUROC = 0.5 (random expectation) to AUROC = 1 for perfect learning accuracy (see Figure 27.1). The result of this study is shown in Figure 27.2 and includes four different experimental settings for the choice of predictor variables (mRNA or proteins) and gradient calculation (coarse or interpolated gradient). The *coarse-mRNA* setting is the most realistic experimental condition that is ordinarily met in wetlabs working with plants because it only involves mRNA sequencing and samples taken in rather coarse time intervals of 2 h.[5] However, we can use the mRNA profiles as proxies for missing protein data to predict mRNA target gradients. Protein profiles (as used in the *coarse-complete* and *interp-complete* setup) are in contrast less abundant but more suitable to predict target mRNA activities because they typically act as transcription factors. The *interp-mRNA* and *interp-complete* setup derive the mRNA gradient for the response from the derivative of a Gaussian process that interpolates the mRNA concentration (Section 27.5.2). A comparison of the AUROC scores in Figure 27.2 reveals that interpolated gradients (*interp*) improve inference over *coarse* gradients. Among

[5]The time interval can vary and depends on the time needed to extract the samples.

the compared inference methods, both Bayesian models, homogeneous Bayesian regression (homogBR) and nonhomogeneous Bayesian regression (nonhomogBR), display a significant improvement over the sparse regression models LASSO and *Elastic Net*. The nonhomogBR method shows slightly better AUROC scores for the *interp-mRNA* and *coarse-complete* setups and will be used in the following real data study.

27.5.5.2 Real Data. The nonhomogeneous Bayesian regression method (nonhomogBR) was applied to the real mRNA profile data set from the TiMet project (Section 27.5.4). Since this data lacks protein profiles and we previously discovered that an interpolated response gradient is superior to a coarse interval gradient, we use the *interp-mRNA* setup to learn the network. The observations were divided into day and night segments to allow for separate linear regression models for these two different conditions. MCMC simulations were run for 50,000 iterations (keeping every 10th sampled parameter configuration), with a burn-in of 40,000, after which standard convergence diagnostics based on the Rubin–Gelman scale reduction factors [22] were met.

Figure 27.3(c) shows the gene regulatory network learned from the TiMet data and Figures 27.3(a) and (b) show two hypothetical networks published in References [42] and [43], respectively. The networks contain two groups of genes: the morning genes LHY/CCA1, PRR5, PRR7, and PRR9 (which, as the name suggests, are expressed in the morning) and the evening genes GI, TOC1, LUX, ELF3, and ELF4 (which are expressed in the evening). A node in the graph represents both the gene as well as the corresponding protein. The subscript "mod" indicates a modified protein isoform. Solid lines represent transcriptional regulation and dashed lines represent protein complex formation. The latter cannot be learned from transcriptional data and are thus systematically missing. This explains, for instance, why the protein complexes EC and ZTL are detached from the remaining network. Various features of the inferred network are consistent with the published networks, such as the activation of PRR9 by LHY/CCA1, the inhibition of the evening genes by the morning genes, and the activation of PRR7 by PRR9. Similar to the network from Reference [43], LHY/CCA1 is inhibited by the evening genes, although the details of the inhibition are slightly different: the inhibition is caused by GI rather than TOC1. Consistent with both published network structures, PRR9 is the target of an inhibition. The regulating node is different from the two publications, however, which is indicative of the current level of uncertainty. Note that the two publications do not agree on this regulation pattern: according to Reference [42], PRR9 is inhibited by TOC1, whereas according to Reference [43], PRR9 is inhibited by EC, and according to the inferred network, PRR9 is inhibited by PRR5. Some interactions are correctly learned, but with opposite signs. For instance, the inferred network predicts that LHY/CCA1 is regulated by PRR9, which is consistent with References [42] and [43]. In these publications, PRR9 is an inhibitor of LHY/CCA1; however, in the inferred network, it is an activator. This ambiguity points to the intrinsic difficulty of the inference task due to the flexibility of the model, as a result of which similar gene expression profiles can be obtained with different interaction and sign configurations. Striking evidence

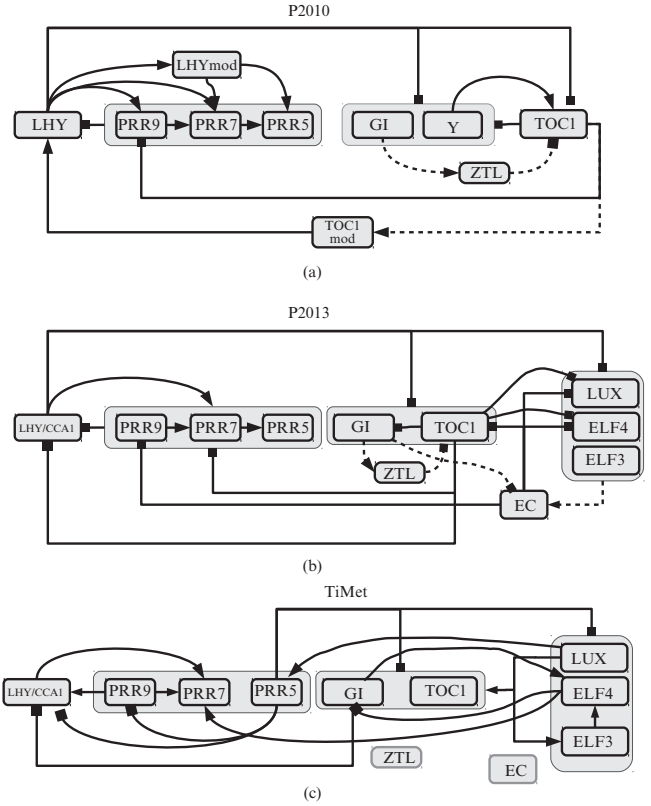

Figure 27.3 Hypothetical circadian clock networks from the literature and that inferred from the TiMet gene expression data. The panels *P2010* (a) and *P2013* (b) constitute hypothetical networks from the literature [42, 43]. The *TiMet* network (c) displays the reconstructed network from the TiMet data, described in Section 27.5.4, using the hierarchical Bayesian regression model from Section 27.5.1. Gene interactions are shown by black lines with arrowhead; protein interactions are shown by dashed lines. The interactions in the reconstructed network were obtained from their estimated posterior probabilities. Those above a selected threshold were included in the interaction network; those below were discarded. The choice of the cutoff threshold is, in principle, arbitrary. For optimal comparability, we selected the cutoff such that the average number of interactions from the published networks was matched (0.6 for molecular interactions).

can be found in the regulatory influence of TOC1 on LHY/CCA1: according to Reference [42], TOC1 is an activator, while according to Reference [43], TOC1 is an inhibitor. This inconsistency between the published networks underlines the fact that the true gene interactions of the circadian clock in *A. thaliana* are still unknown, and that different models focus on different aspects of the data. It is, thus, encouraging to note that several features of the published networks have been inferred with our statistical/machine learning model, using the data alone without any assumption of biological prior knowledge. This finding suggests that the statistical/machine learning models discussed in the present chapter, while too limited to uncover the

whole ground truth, still provide powerful tools for the generation of biological hypothesis.

27.6 LEARNING SPECIES INTERACTION NETWORKS

Ecological systems share several commonalities with molecular systems as was already pointed out in Section 27.1. One distinctive feature of species networks, however, is the so-called autocorrelation effect and dispersion of species, which can impact population densities in neighboring regions. We account for this effect by introducing a virtual node into the regression model as described in Section 27.6.1. Another major difference rests on the fact that observations in ecological field surveys spread over a two-dimensional space. Although it would be straightforward to process the two-dimensional data in the same way as a series of molecular observations, one has to consider that ecological processes can be highly nonhomogeneous in space. The main driving forces in this respect are changes in the environment and population densities that produce direct or indirect effects on species interaction dynamics, for example, leading to the formation of ecological niches. Hence, a challenge in ecological modeling is the partitioning of space into local neighborhoods with similar population dynamics. Prediction methods using this knowledge can improve their model accuracy. In addition, it can be beneficial to learn the partitioning directly from the data and in this way gain knowledge about potential neighborhoods. To this end, we extend the one-dimensional *change-point* model described in Section 27.5.1 to two dimensions, and extend it with an inference mechanism that attempts to learn the local neighborhoods from the observed population densities. The multiple *change-point* model introduced in Reference [1], uses a global partitioning scheme and is described in Section 27.6.2. The method is complemented by an improved version that uses a local partitioning scheme based on a *Mondrian process* (Section 27.6.3) introduced in Reference [2]. These methods are evaluated together with the previously defined homogeneous Bayesian regression and sparse regression methods on synthetic data (Section 27.6.4) and simulated data based on a realistic ecological niche model (Section 27.6.5.3). The best performing method is applied to a real data set involving plant coverage and soil attribute measurements (Section 27.6.6).

27.6.1 Regression Model of Species interactions

For all species n, the random variable $y_n(x_1, x_2)$ refers to the abundance of species n at location (x_1, x_2). Within any partition h, this abundance depends on the abundance levels of the species in the regulator set of species n, π_n. The regulator set π_n is defined to be the same in all partitions $h \in \{1, \ldots, H\}$ to rule out fundamental changes to the network, as network changes among partitions are less likely to occur than a change in interaction strength. We model the partition-specific linear regression model with the set of parameters $\{(w^p_{n,h})_{p\in\pi_n}, \sigma_{n,h}\}$, where $w^p_{n,h} \in \mathbb{R}$ is a regression coefficient and $\sigma^2_{n,h} > 0$ is the noise variance for each segment h and target species n. For all species

n and all locations (x_1, x_2) in segment h, the response species $y_n(x_1, x_2)$ depends on the abundance variable of the predictor species $\{y_p(x_1, x_2)\}_{p \in \pi_n}$ according to

$$y_n(x_1, x_2) = w^0_{n,h} + \sum_{p \in \pi_n} w^p_{n,h} y_p(x_1, x_2) + \epsilon_n(x_1, x_2) + w^A_{n,h} A_n(x_1, x_2) \tag{27.13}$$

where $\epsilon_n(x_1, x_2)$ is assumed to be white Gaussian noise with mean 0 and variance $\sigma^2_{n,h}$, $\epsilon_n(x_1, x_2) \sim N(0, \sigma^2_{n,h})$. We define $\mathbf{w}_{n,h} = (w^0_{n,h}, \{w^p_{n,h}\}_{p \in \pi_n}, w^A_{n,h})$ to denote the vector of all regression parameters of species n in partition h. This includes the parameters defining the strength of interactions with other species p, $w^p_{n,h}$, as well as a species-specific offset term, that is, the intercept or bias $w^0_{n,h}$. Spatial autocorrelation effects are represented with $A_n(x_1, x_2)$ weighted by an additional edge $w^A_{n,h}$. They reflect the influence of neighboring cells that can have a strong effect on model performance [36]. $A_n(x_1, x_2)$ denotes the average densities in the vicinity of (x_1, x_2), weighted inversely proportional to the distance of the neighbors:

$$A_n(x_1, x_2) = \frac{\sum_{(\tilde{x}_1, \tilde{x}_2) \in \mathcal{N}(x_1, x_2)} d^{-1}[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)] Y_n(\tilde{x}_1, \tilde{x}_2)}{\sum_{(\tilde{x}_1, \tilde{x}_2) \in \mathcal{N}(x_1, x_2)} d^{-1}[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)]} \tag{27.14}$$

where $\mathcal{N}(x_1, x_2)$ is the spatial neighborhood of location (x_1, x_2) (e.g., the four nearest neighbors), and $d[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)]$ is the Euclidean distance between (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$.

27.6.2 Multiple Global Change-Points

We assume a two-dimensional grid with observations sampled at the locations (x_1, x_2) with $x_1 \in \{1, \ldots, T^1\}$ and $x_2 \in \{1, \ldots, T^2\}$, where T^i specifies the number of observations in the horizontal direction with $i = 1$ and vertical direction with $i = 2$. The regulatory relationships among the species may be influenced by latent variables, which are represented by spatial *change-points*. We assume that latent effects in close spatial proximity are likely to be similar, but locations where spatially close areas are not similar are separated by *change-points*. They are modeled with two *a priori* independent multiple *change-point* processes for each target species n along the two orthogonal spatial directions: $\boldsymbol{\tau}^i_n = (\tau^i_{n,1}, \ldots, \tau^i_{n,k^i})$, with $i \in \{1, 2\}$ and k^i_n defining the number of *change-points* along the horizontal $(i = 1)$ and vertical $(i = 2)$ direction. The *pseudo-change-points* $\tau^i_{n,0} := 1$ and $\tau^i_{n,k^i+1} := T^i$ define the boundaries of the two-dimensional grid (see Figure 27.4). The *change-point* vectors $\boldsymbol{\tau}^1_n$ and $\boldsymbol{\tau}^2_n$ can be set to fixed values but we attempt to learn the partitioning and, thus, $\boldsymbol{\tau}^i_n$ contains an *a priori* unknown number of k^i_n *change-points*. The vectors $\boldsymbol{\tau}^i_n$ and *pseudo-change-points* $\tau^i_{n,0} := 1, \tau^i_{n,k^i+1} := T^i$ partition the space into $H_n = (k^1_n + 1)(k^2_n + 1)$ nonoverlapping segments. We denote the latent variable associated with a segment by $h_n \in \{1, \ldots, H_n\}$. Figure 27.4 illustrates an example partitioning of a two-dimensional grid.

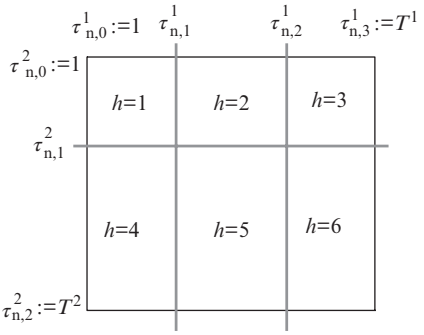

Figure 27.4 Multiple global *change-point* example. Partitioning with a horizontal *change-point* vector $\boldsymbol{\tau}_n^{i=1} = (\boldsymbol{\tau}_{n,1}^1, \boldsymbol{\tau}_{n,2}^1)$ and vertical vector $\boldsymbol{\tau}_n^{i=2} = (\boldsymbol{\tau}_{n,1}^2)$. The *pseudo-change-points* $\boldsymbol{\tau}_{n,0}^1 = \boldsymbol{\tau}_{n,0}^2 := 1$ define the left and upper boundaries; $\boldsymbol{\tau}_3^1 = T^1$ and $\boldsymbol{\tau}_2^2 = T^2$ define the lower and right boundaries, where T^1 and T^2 are the number of locations along the horizontal and vertical directions, respectively. The number of *change-points* is $k_n^1 = 2$, $k_n^2 = 1$ and the number of segments $H_n = 6$.

The priors for the parameter $\mathbf{w}_{n,h}$, regulator set π_n, variance σ_n^2, and signal-to-noise hyperparameter δ_n^2, as well as the likelihood, are defined in a similar way as in the nonhomogeneous Bayesian regression model in Section 27.5.1 with the modification that considers the different partitioning scheme. In addition, the design matrix $\mathbf{X}_{n,h}$ is extended by the spatial autocorrelation variable A_n, and $\mathbf{w}_{n,h}$ includes an additional fixed edge $w_{n,h}^A$. The number of *change-points*, k_n^1, k_n^2, and locations, $\boldsymbol{\tau}_n^1$, $\boldsymbol{\tau}_n^2$, for a species n is sampled from the marginal posterior distribution

$$P(\boldsymbol{\tau}_n^i, k_n^i | \mathbf{X}, \delta_n) \propto \prod_{i=1}^{2} P(\boldsymbol{\tau}_n^i | k_n^i) P(k_n^i | \lambda_n) \prod_{h=1}^{H_n} P(\mathbf{y}_{n,h} | \mathbf{X}_{\pi_n,h}, \delta_n) \tag{27.15}$$

using an RJMCMC scheme [24], conditional on the following *change-point* priors: for both spatial directions $i \in \{1, 2\}$, the $k_n^i + 1$ intervals are delimited by k_n^i *change-points* and boundary *change-points* $\boldsymbol{\tau}_{n,0}^i$ and $\boldsymbol{\tau}_{n,k_n^i+1}^i$, where k_n^i is distributed *a priori* as a truncated Poisson random variable with mean λ_n and maximum $\overline{k^i} = T^i - 1$: $P(k_n^i | \lambda_n) \propto \frac{\lambda_n^{k_n^i}}{k_n^i!} \mathbb{1}_{\{k_n^i \leq \overline{k^i}\}}$. Conditional on k_n^i *change-points*, the *change-point* position vector $\boldsymbol{\tau}_n^i = (\boldsymbol{\tau}_{n,1}^i, \ldots, \boldsymbol{\tau}_{n,k^i}^i)$ takes nonoverlapping integer values, which we take to be uniformly distributed *a priori*. There are $(T^i - 1)$ possible positions for the k_n^i *change-points*, thus vector $\boldsymbol{\tau}_n^i$ has the prior density $P(\boldsymbol{\tau}_n^i | k_n^i) = 1 / \binom{T^i - 1}{k_n^i}$. Reference [1] provides a more detailed description of the model.

27.6.3 Mondrian Process Change-Points

The global *change-point* process described in the previous section lacks the capability to create segmentations with spatially varying length scales and different local

fineness and coarseness characteristics. In fact, introducing global *change-points* that might improve segmentation in one region can introduce artifacts in the form of undesired partitioning in other regions. A local approach to partitioning was proposed in Reference [2] using the so-called *Mondrian process* [45]. It is a generative recursive process for self-consistently partitioning the two-dimensional domain in the following way. A hyperparameter λ (the so-called *budget*) determines the average number of cuts in the partition. At each stage of the recursion, a Mondrian sample can either define a trivial partition $\Theta_1 \times \Theta_2$, that is, a segment, or a cut that creates two subprocesses $m_<$ and $m_>$: $m = \langle d, \chi, \lambda', m_<, m_> \rangle$, where $d \in \{0, 1\}$ is a binary indicator for the horizontal and vertical directions and χ the position of the cut. The direction d and position χ are drawn from a binomial and a uniform distribution, respectively, both depending on Θ_1 and Θ_2. The process of cutting a segment is limited by the budget λ associated with each segment and the cost E of a cut. Conditional on the half-perimeter $\boldsymbol{\tau} = |\Theta_1| + |\Theta_2|$, a cut is introduced yielding $m_<$ and $m_>$ if the cost $E \sim \exp(\boldsymbol{\tau})$ does not exceed the budget λ, that is, satisfies $\lambda' = \lambda - E > 0$ with λ' defining the budget that is assigned to the subprocesses $m_<$ and $m_>$. The process is recursively repeated on $m_<$ and $m_>$ until the budgets are exhausted. This creates a binary tree with the initial Mondrian sample $m_{k=1}$ as the root node spanning the unit square $[0; 1]^2$ and subnodes representing Mondrian samples $m_{1<k\leq K}$, $k \in \{1, \ldots, K\}$ where K is the total number of nodes in the tree, for example, $K = 15$ in Figure 27.5. The leaf nodes present nonoverlapping segments and are associated each with a latent variable $h(k)$ labeled with $m^{h(k)}$ (Figure 27.5). These latent variables determine the interactions among species, as described in Section 27.6.1. We denote by H the number of uncut segments, for example, $H = 8$ in Figure 27.5(a), and $h(k) \in \{1, \ldots, H\}$.

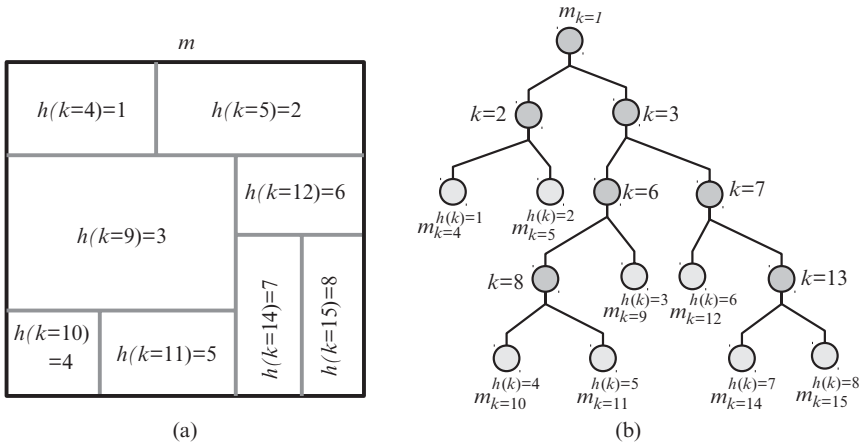

Figure 27.5 *Mondrian process* example. (a) An example partitioning with a *Mondrian process*. (b) The associated tree with labels of the latent variable $h(k)$ identifying each nonoverlapping segment with leaf nodes (light gray) designated as $m_k^{h(k)}$, where k indexes all tree nodes.

An essential step of the inference procedure is to sample a new *Mondrian process* segment m or remove existing ones. This is realized with an RJMCMC scheme described in detail in Section 2.8 of Reference [2]. The inference of model parameters $\mathbf{w}_n$ and structure π are the same as with the previously defined nonhomogeneous Bayesian regression model.

27.6.4 Synthetic Data

For an objective evaluation of network inference, we test the ability of the previously described methods to recover the true network structure from test data generated from a piecewise linear regression model following Equation 27.13. The data was partitioned by two-dimensional *change-points* that resemble a *Mondrian process* following Reference [2], that is, the data grid is iteratively subdivided into local segments (e.g., as shown in Figure 27.5a). The number of observations was selected to be 15 in each direction. The number of nodes n was set to 10 and the number of regulators for each node was sampled from a Poisson distribution. The regression coefficients $\mathbf{w}_{n,h}$ together with the intercept $w^0_{n,h}$ of each segment h were sampled from a uniform distribution in the interval of $[-1; -0.5]$ and $[0.5, 1.0]$. The noise ϵ_n was sampled from a normal distribution. Nodes without an incoming edge were initialized to a Gaussian random number. The values of the remaining nodes were calculated at each grid cell following Equation 27.13.

27.6.5 Simulated Population Dynamics

For a realistic evaluation, we followed Reference [17] and generated data from an ecological simulation that combines a niche model [62] with a stochastic population model [34] in a two-dimensional lattice.

27.6.5.1 Niche Model and Species Interactions. The niche model defines the structure of the trophic network and has two parameters: the number of species N and the connectivity (or network density) defined as L/N^2 where L is the number of interactions (edges) in the network. Each species n is assigned a niche value x_n, drawn uniformly from $[0, 1]$. This gives an ordering of the species, where higher values mean that species are higher up in the food chain. For each species, a niche range R_n is drawn from a beta distribution with expected value $2C$ (where C is the desired connectivity), and species n consumes all species falling in a range R_n that is placed by uniformly drawing the centre c_n of the range from $[R_n/2, x_n]$ as illustrated in Figure 27.6 and introduced in Reference [62]. Despite its simplicity, it was shown there that the resulting networks share many characteristics with real food webs.

27.6.5.2 Stochastic Population Dynamics. The population model is defined by a stochastic differential equation where the dynamics of the log abundance $X_n(t)$ of species n at time t can be expressed as

$$\frac{dX_n(t)}{dt} = r_n + \frac{\sigma_d}{\sqrt{e^{X_n(t)}}}\frac{dA_n(t)}{dt} + \sigma_e\frac{dB_n(t)}{dt} - \gamma X_n(t) - \Omega(X) + \sigma_E\frac{dE(t)}{dt} \qquad (27.16)$$

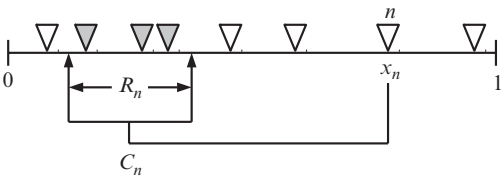

Figure 27.6 Diagram of the niche model. Species are indicated by triangles. A species n is placed with a niche value x_n into the interval $[0, 1]$. A value c_n is uniformly drawn that defines the centre of the range R_n. All species with a value x inside this interval, that is, $c_n - \frac{R_n}{2} \leq x \leq c_n + \frac{R_n}{2}$, as indicated by the gray triangles, are consumed ("eaten") by species n. Diagram adapted from Reference [62].

where X is the set of all $X_N(t)$, r_n is the growth rate of species n, σ_d is the standard deviation of the demographic effect, $A_n(t)$ is the species-specific demographic effect, σ_e is the standard deviation of the species-specific environmental effect, $B_n(t)$ is the species-specific environmental effect, γ is the intraspecific density dependence, Ω is the effect of competition for common resources, σ_E is the standard deviation of the general environmental effect, and $E(t)$ is the general community environment. The growth rates r_n are location dependent (depending on the cell of a rectangular grid), with a spatial pattern that is generated by noise with spectral density f^β (with $\beta < 0$, and f denoting the spatial frequency at which the noise is measured). An illustration is given in Figure 27.7. To model species migration, we included an exponential dispersal model, where the probability of a species moving from one location to another is determined by the Euclidean distance between the locations. To incorporate the niche model, we modified the term Ω in (27.16) to include predator–prey interactions in the Lotka–Volterra form. A detailed description is available in Reference [17].

27.6.5.3 Simulation. We applied this model to 10 species living in a 25×25 rectangular grid. We simulated the dynamics of this model for 3000 steps and then recorded species abundance levels in all grid cells at the final step; this corresponds to an ecological survey carried out at a fixed moment in time. For each grid cell, we

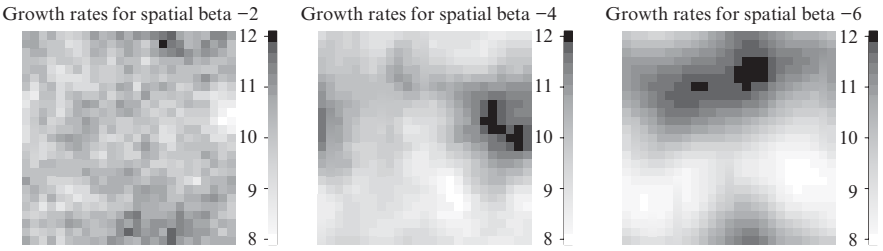

Figure 27.7 Spatial distribution. Shown are the spatial distributions of growth rates r_n entering Equation 27.16 as the spatial β parameter (Section 27.6.1) decreases from −2 to −6. A value of 0 corresponds to uniformly random noise, and −2 is Brownian noise.

counted the number of species that went extinct. These counts were added up over all cells, yielding the total number of extinctions. A simulation was rejected if these extinctions exceeded the value 50. For each of the spatial β parameters displayed in Figure 27.9, 30 surveys were collected by running the simulation repeatedly with different networks and parameter initializations.

27.6.6 Real World Plant Data

We have applied the method to real-world data from Reference [37], including 106 vascular plants and 12 environmental variables collected from a 200 m × 2162 m Machair vegetation land stripe at the western shore of the Outer Hebrides. Samples were taken at 217 locations, each 1 m ×1 m in size, equally distributed with a 50 m spacing. Plant samples were measured as ground coverage in percentage and physical samples as absolute values (such as moisture, pH value, organic matter, and slope). The data was log-normal transformed after observing substantial skewness in the distributions. Each sample point was mapped into a two-dimensional grid ignoring locations with no sample data available. The spatial autocorrelation value for each plant and location was calculated from neighbors inside a radius of 70 m. Since we are interested only in plant interactions, we defined each plant to have all 12 physical soil variables as fixed input, that is, permanent predictor variables.

27.6.7 Method Evaluation and Learned Networks

27.6.7.1 Synthetic Data. To provide a fair comparison between the methods, we disabled the spatial autocorrelation variables for all methods because no dispersion effect was simulated in the synthetic data model (Section 27.6.4). Figure 27.8 shows the AUROC scores for BRAMP (Bayesian regression with *Mondrian process change-points*, Section 27.6.3), BRAM (Bayesian regression with global *change-points*, Section 27.6.2), homogBR (homogeneous Bayesian regression without *change-points*, Section 27.4.3), LASSO and *Elastic Net* (Section 27.4.2). BRAMP outperforms all other schemes, which is not surprising, in that the data have been generated from a process that is consistent with the modeling assumptions of BRAMP. Notable is also the improvement of BRAMP over BRAM, which indicates that BRAMP is more flexible in terms of identifying local segments. It is reassuring that both nonhomogeneous Bayesian regression schemes (BRAMP and BRAM) can handle the increased model complexity, and improve network reconstruction accuracy compared to the competing methods homogBR, LASSO, and *Elastic Net*.

27.6.7.2 Simulated Population Dynamics. For a fair comparison, additional spatial autocorrelation variables (Section 27.6.1) were added to all methods, enabling them to account for dispersion effects intrinsic to this data set. The *Elastic Net* method was excluded from comparison because of the very similar results to LASSO. Figure 27.9 shows the AUROC scores for four different settings of the spatial β parameter that controls the heterogeneity of species distributions as

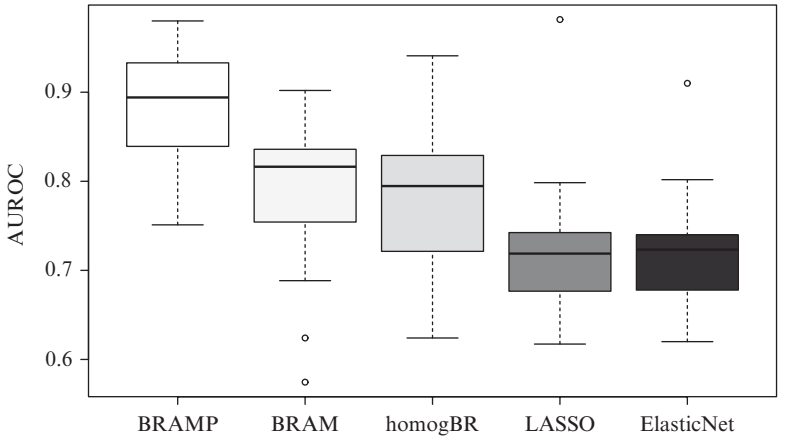

Figure 27.8 Comparison on synthetic data. Boxplots of AUROC scores obtained with five methods on the synthetic data described in Section 27.5: A *Bayesian regression model with* Mondrian process *change-points* (BRAMP, Section 27.6.4), a *Bayesian regression model with global change-points* (BRAM, Section 27.6.2), a Bayesian linear regression model without *change-points* (homogBR, Section 27.4.3), L1-penalized sparse regression (LASSO, Section 27.4.2), and the sparse regression *Elastic Net* method (Section 27.4.2). The boxplots show the distributions of the scores for 30 independent data sets with higher scores indicating better learning performance.

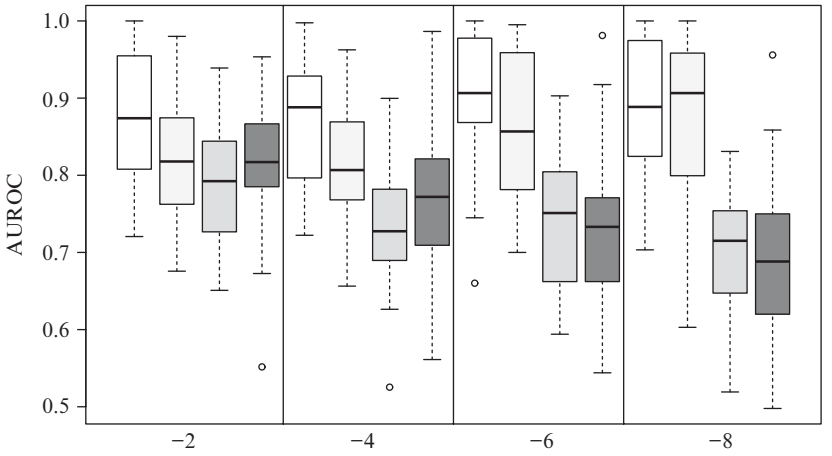

Figure 27.9 Comparative evaluation of four network reconstruction methods for the stochastic population dynamics data. Boxplots of AUROC scores obtained on the realistic simulated data described in Section 27.6.5.3 for different settings of the spatial β parameter, with lower values causing stronger heterogeneity in the data. Box color scheme: BRAMP (white), BRAM (light gray), homogBR (gray), and LASSO (dark gray).

illustrated in Figure 27.7. Lower values of β lead to the formation of clusters or "neighborhoods" of similar species concentrations, making it more difficult to learn underlying structures when averaging over the whole data domain. By treating these neighborhoods separately with a partitioning scheme, we would expect improved inference. In fact, Figure 27.9 shows that methods with a *change-point* process, BRAMP and BRAM, produce better AUROC scores than the competing methods homogBR and LASSO, which lack this feature. BRAMP, the Bayesian regression model with local partitioning, consistently outperforms the other methods, as displayed in Figure 27.9 with the single exception of BRAM and a spatial $\beta = -8$. Table 27.1 summarizes the corresponding p-values of paired Wilcoxon tests for the AUROC scores comparing BRAMP against BRAM, homogBR, and LASSO. The low p-values indicate a significant performance gain of BRAMP and suggest

Table 27.1 Improvement of the Bayesian regression model with *Mondrian process change-points* (BRAMP) on the stochastic population dynamics data

Spatial β:	-2	-4	-6	-8
BRAM	2.2e−04	1.9e−04	6.4e−03	0.14
homogBR	1.2e−06	2.9e−07	1.0e−07	1.9e−09
LASSO	6.1e−04	7.2e−04	1.3e−08	9.3e−10

p-Values for paired one-sided Wilcoxon tests for the difference of AUROC scores between BRAMP and the competing methods (BRAM, homogBR, LASSO) for several spatial β values. The alternative hypothesis states that BRAMP scores are greater than the competing methods with low p-values < 0.05 indicating significant performance gain of BRAMP.

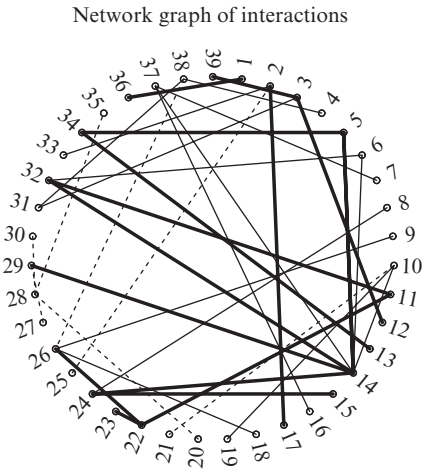

Figure 27.10 Species interaction network. Species interactions as inferred with BRAMP (Section 27.6.3), with an inferred marginal posterior probability of 0.5 (thick lines) and 0.1 (thin lines). Solid lines are positive interactions (e.g., mutualism, commensalism) and dashed are negative interactions (e.g., resource competition). Species are represented by numbers and have been ordered phylogenetically as displayed in Table 27.2.

Table 27.2 Indices with full scientific names as appearing in Figure 27.10

ID	Name	ID	Name
1	*Anagallis tenella*	21	*Aira praecox*
2	*Calluna vulgaris*	22	*Anthoxanthum odoratum*
3	*Drosera rotundifolia*	23	*Cynosurus cristatus*
4	*Epilobium palustre*	24	*Festuca rubra*
5	*Galium verum*	25	*Festuca vivipara*
6	*Hypochaeris radicata*	26	*Holcus lanatus*
7	*Leontodon autumnalis*	27	*Koeleria macrantha*
8	*Lychnis flos-cuculi*	28	*Molinia caerulea*
9	*Odontites verna*	29	*Poa pratensis*
10	*Plantago lanceolata*	30	*Juncus effusus*
11	*Potentilla erecta*	31	*Juncus kochii*
12	*Potentilla palustris*	32	*Luzula campestris*
13	*Prunella vulgaris*	33	*Luzula pilosa*
14	*Ranunculus bulbosus*	34	*Carex arenaria*
15	*Ranunculus repens*	35	*Carex demissa*
16	*Sagina procumbens*	36	*Carex dioica*
17	*Succisa pratensis*	37	*Carex flacca*
18	*Trifolium repens*	38	*Carex nigra*
19	*Viola riviniana*	39	*Eriophorum angustifolium*
20	*Agrostis capillaris*		

These plants can be assigned to four taxonomies of forbs (1–19), grasses (20–29), rushes (30–33), and sedges (34–39).

that the *Mondrian process* better captures spatial heterogeneity. In fact, both nonhomogeneous Bayesian regression models, BRAMP and BRAM, achieve high AUROC scores for the data simulated with low spatial β values, that is, high data heterogeneity. In contrast, the performance of homogBR and LASSO deteriorates as expected with higher data heterogeneity (i.e., lower spatial β).

27.6.7.3 Real-World Plant Data. We have applied BRAMP to the plant abundance data from the ecological survey described in Section 27.6.6. We sampled interaction network structures from the posterior distribution with MCMC and computed the marginal posterior probabilities of the individual potential species interactions. We kept all species interactions with a marginal posterior probability above a certain threshold, as shown in Figure 27.10, resulting in 39 out of 106 species with relevant interaction in the reconstructed network shown in Figure 27.10. Since we had defined the 12 soil attributes as fixed predictors to each plant, the interactions in this network represent plant–plant interactions not mediated by similar soil preferences. This network can lead to the formation of new ecological hypotheses. For instance, *Ranunculus bulbosus* (species 14) is densely connected with five interspecific links above the threshold. Can that be related to its tolerance for nutrient-poor soil and its preferred occurrence in species-rich patches? There is a noticeable imbalance between positive and negative interactions. An initial consultation with ecologists indicates the

fact that our analysis tends to find more positive than negative links is interesting as it points to a dominance of commensalism over competition. The importance of commensalism was emphasized in Reference [9]. Ecologists also suggest that positive interactions may be more characteristic for harsh environments (e.g., in Reference [8]) as is found in the Machair vegetation. These results demonstrate that the proposed method provides a useful tool for exploratory data analysis in ecology with respect to both species interactions and spatial heterogeneity.

27.7 CONCLUSION

We have addressed the problem of reconstructing gene regulation networks from molecular expression data and modified the corresponding methods to reconstruct species interaction networks from species abundance data. To this end, we have described and applied two sparse regression methods, LASSO and *Elastic Net*, and a Bayesian regression method. The latter was extended from a homogeneous model (homogBR) to a nonhomogeneous model that can approximate nonhomogeneity with multiple *change-points*. This nonhomogeneous Bayesian regression method (nonhomogBR) was successfully applied to the inference of a circadian regulation network and benefited in terms of better performance from light-induced day and night phases. We have seen that several commonalities exist between molecular and ecological systems. They can be exploited to adapt established inference methods from systems biology to ecological applications. This was demonstrated on the nonhomogeneous Bayesian regression method in two ways: the partitioning of time was modified to segment a two-dimensional spatial domain, and an additional variable that reflects species dispersion effects was introduced. As a result, a global *change-point* model (BRAM) and a spatially varying, local partitioning scheme (BRAMP) were applied to realistic simulated and real data. We found that both approaches greatly benefit from learning the spatial segmentation in terms of the quality of inferring species interactions. They showed superior performance compared to LASSO and homogBR.

These results are encouraging and suggest that ecological modeling could greatly benefit from the methodologies established in systems biology during the last decade. This is of particular relevance given the great complexity of ecological systems and the increasing amount of data that has to be analyzed. A substantial improvement of the methodology could help to better understand the ecological processes, such as the impact of global warming on biodiversity [47], land use and agriculture [39], or the stability of ecological systems that are threatened by a radical extinction of plant and animal life, as occurring in the Amazonian rain forest [44].

REFERENCES

1. Aderhold A, Husmeier D, Lennon J, Beale C, Smith V. Hierarchical Bayesian models in ecology: reconstructing species interaction networks from non-homogeneous species abundance data. Ecol Inform 2012;11:55–64.

2. Aderhold A, Husmeier D, Smith V. Reconstructing ecological networks with hierarchical Bayesian regression and Mondrian processes. Proceedings of the 16th International Conference on Artificial Intelligence and Statistics; Phoenix, AZ; 2013. p 75–84.
3. Aderhold A, Husmeier D, Smith V, Millar A, Grzegorczyk M. Assessment of regression methods for inference of regulatory networks involved in circadian regulation. 10th International Workshop on Computational Systems Biology (WCSB); Tampere, Finland; 2013. p 30–42.
4. Andrieu C, Doucet A. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. IEEE Trans Signal Process 1999.;47(10):2667–2676.
5. Arbeitman M, Furlong E, Imam F, Johnson E, Null B, Baker B, Krasnow M, Scott M, Davis R, White K. Gene expression during the life cycle of *Drosophila melanogaster*. Science 2002;297(5590):2270–2275.
6. Beisner B, Haydon D, Cuddington K. Alternative stable states in ecology. Front Ecol Environ 2003;1(7):376–382.
7. Bishop C. *Pattern Recognition and Machine Learning*. Singapore: Springer-Verlag; 2006.
8. Brooker R, Callaghan T. The balance between positive and negative plant interactions and its relationship to environmental gradients: a model. Oikos 1998;81(1):196–207.
9. Bruno J, Stachowicz J, Bertness M. Inclusion of facilitation into ecological theory. Evolution 2003;18(3):119–125.
10. Ciocchetta F, Hillston J. Bio-PEPA: A framework for the modelling and analysis of biological systems. Theor Comput Sci 2009;410(33):3065–3084.
11. Cohen J. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. PLoS Biol 2004;2(12):e439.
12. Cohen J, Schoenly K, Heong K, Justo H, Arida G, Barrion A, Litsinger J. A food web approach to evaluating the effect of insecticide spraying on insect pest population dynamics in a Philippine irrigated rice ecosystem. J Appl Ecol 1994;31:747–763.
13. Davies J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning; Pittsburgh, PA; 2006. p 233–240.
14. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000;16(8):707–726.
15. Dunne J, Williams R, Martinez N. Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecol Lett 2002;5:558–567.
16. Edwards KD, Akman OE, Knox K, Lumsden PJ, Thomson AW, Brown PE, Pokhilko A, Kozma-Bognar L, Nagy F, Rand DA, Millar AJ. Quantitative analysis of regulatory flexibility under changing environmental conditions. Mol Syst Biol 2010;6:424.
17. Faisal A, Dondelinger F, Husmeier D, Beale C. Inferring species interaction networks from species abundance data: a comparative evaluation of various statistical and machine learning methods. Ecol Inform 2010;5(6):451–464.
18. Flis A, Fernandez P, Zielinski T, Sulpice R, Pokhilko A, McWatters H, Millar A, Stitt M, Halliday K. Biological regulation identified by sharing timeseries data outside the 'omics; 2013. Submitted.
19. Fogelberg C, Palade V. Machine learning and genetic regulatory networks: a review and a roadmap. In: *Foundations of Computational Intelligence*. Volume 1. Berlin, Germany: Springer-Verlag; 2009. p 3–34.

20. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1–22. [Online]. Available at http://www.jstatsoft.org/v33/i01/.
21. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7:601–620.
22. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457–511.
23. Gillespie D. Exact stochastic simulation of coupled chemical reactions. J Phys Chem 1977;81(25):2340–2361.
24. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995;82:711–732.
25. Grzegorczyk M, Husmeier D. A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology. Stat Appl Genet Mol Biol (SAGMB) 2012;11(4):Article 7.
26. Grzegorczyk M, Husmeier D. Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. Mach Learn 2013;91(1):105–154.
27. Guerriero M, Pokhilko A, Fernández A, Halliday K, Millar A, Hillston J. Stochastic properties of the plant circadian clock. J R Soc Interface 2012;9(69):744–756.
28. Hanely J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
29. Hayete B, Gardner T, Collins J. Size matters: network inference tackles the genome scale. Mol Syst Biol 2007;3:77.
30. Henneman M, Memmott J. Infiltration of a Hawaiian community by introduced biological control agents. Science 2001;293(5533):1314–1316.
31. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. BMC Genomics 2006;7(1):25.
32. Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 2003;19:2271–2282.
33. Kalaitzis A, Honkela A, Gao P, Lawrence N. gptk: Gaussian Processes Tool-Kit; 2013, r package version 1.06. [Online]. Available at http://CRAN.R-project.org/package=gptk. Accessed 2015 Aug 8.
34. Lande R, Engen S, Saether B. *Stochastic Population Dynamics in Ecology and Conservation*. New York: Oxford University Press; 2003.
35. Lèbre S, Becq J, Devaux F, Stumpf M, Lelandais G. Statistical inference of the time-varying structure of gene-regulation networks. BMC Syst Biol 2010;4:130.
36. Lennon J. Red-shifts and red herrings in geographical ecology. Ecography 2000;23:101–113.
37. Lennon J, Beale C, Reid C, Kent M, Pakeman R. Are richness patterns of common and rare species equally well explained by environmental variables. Ecography 2011;34:529–539.
38. Locke JC, Southern MM, Kozma-Bognar L, Hibberd V, Brown PE, Turner MS, Millar AJ. Extension of a genetic network model by iterative experimentation and mathematical analysis. Mol Syst Biol 2005;1:2005.0013.

39. Mendelsohn R, Nordhaus W, Shaw D. The impact of global warming on agriculture: a ricardian analysis. Am Econ Rev 1994;84:753–771.

40. Passos J, Nelson G, Wang C, Richter T, Simillion C, Proctor CJ, Miwa S, Olijslagers S, Hallinan J, Wipat A, Saretzki G, Rudolph KL, Kirkwood TB, von Zglinicki T. Feedback between p21 and reactive oxygen production is necessary for cell senescence. Mol Syst Biol 2010;6:347.

41. Pokhilko A, Fernández A, Edwards K, Southern M, Halliday K, Millar A. The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. Mol Syst Biol 2012;8:574.

42. Pokhilko A, Hodge SK, Stratford K, Knox K, Edwards KD, Thomson AW, Mizuno T, Millar AJ. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. Mol Syst Biol 2010;6:416.

43. Pokhilko A, Mas P, Millar A. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. BMC Syst Biol 2013;7(1):1–12.

44. Rammig A, Thonicke K, Jupp T, Ostberg S, Heinke J, Lucht W, Cramer W, Cox P. Estimating Amazonian rainforest stability and the likelihood for large-scale forest dieback. EGU General Assembly Conference Abstracts, Volume 12; 2010. p 14289.

45. Roy DM. Computability, inference and modeling in probabilistic programming [PhD dissertation]. Massachusetts Institute of Technology; 2011.

46. Sachs K, Perez O, Pe'er D, Lauffenburger D, Nolan G. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308(5721):523–529.

47. Sala OE, Chapin FS, Armesto JJ, Berlow E, Bloomfield J, Dirzo R, Huber-Sanwald E, Huenneke LF, Jackson RB, Kinzig A, Leemans R, Lodge DM, Mooney HA, Oesterheld M, Poff NL, Sykes MT, Walker BH, Walker M, Wall DH. Global biodiversity scenarios for the year 2100. Science 2000;287(5459):1770–1774.

48. Scheffer M, Carpenter S, Foley JA, Folke C, Walker B. Catastrophic shifts in ecosystems. Nature 2001;413(6856):591–596.

49. Schena M, Shalon D, Davis R, Brown P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467–470.

50. Shinozaki K, Yamaguchi-Shinozaki K, Seki M. Regulatory network of gene expression in the drought and cold stress responses. Curr Opin Plant Biol 2003;6(5):410–417.

51. Sinclair A, Byrom A. Understanding ecosystem dynamics for conservation of biota. J Anim Ecol 2006;75(1):64–79.

52. Solak E, Murray-Smith R, Leithead W, Leith D, Rasmussen C. Derivative observations in Gaussian process models of dynamic systems. In: *Proceedings of Neural Information Processing Systems*. Vancouver, Canada: MIT Press; 2002.

53. Solomon S. *Climate Change 2007 - The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Volume 4. Cambridge, United Kingdom: Cambridge University Press; 2007.

54. Tait R. The application of molecular biology. Curr Issues Mol Biol 1999;1(1):1–12.

55. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodological) 1995;58(1):267–288. [Online]. Available at http://www.jstor.org/stable/2346178.

56. TiMet-Consortium. The TiMet Project - Linking the clock to metabolism; 2012. [Online] Available at http://timing-metabolism.eu. Accessed 2015 Aug 8.

57. Vyshemirsky V, Girolami M. Bayesian ranking of biochemical system models. Bioinformatics 2008;24:833–839.

58. Wang G, Zhu X, Hood L, Ao P. From Phage lambda to human cancer: endogenous molecular-cellular network hypothesis. Quant Biol 2013;1(1):1–18.

59. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, DREAM5 Consortium, Bussemaker HJ, Morris QD, Bulyk ML, Stolovitzky G, Hughes TR. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol 2013;31(2):126–134.

60. Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics 2006;22:2523–2531.

61. Wilkinson D. *Stochastic Modelling for Systems Biology*. Volume 44. Boca Raton, FL: CRC Press; 2011.

62. Williams R, Martinez N. Simple rules yield complex food webs. Nature 2000;404(6774): 180–183.

63. Zou H, Hastie T. Regularization and variable selection via the Elastic net. J R Stat Soc Ser B (Stat Methodol) 2005;67(2):301–320. [Online]. DOI: 10.1111/j.1467-9868.2005 .00503.x.