Chapter 7
Introduction to Network Inference in Genomics

Ernst C. Wit

Abstract The genome is the archetypical complex system: it is a finely tuned whole whose many parts, such as DNA, RNA and proteins, interact at various levels to execute intricate functions, such as repair, replication and adapting to the external environment. One particularly effective way of conceptualizing this complex system is by means of a network, in which the vertices describe the genomic components and the edges describe their physical or functional interactions. With the advent of modern high-throughput genomic measuring devices, such as microarrays, RNA-seq and other next generation sequencing tools, it has become possible to measure the vertices of the genomic system in real time. One central question is whether from these measurements it is possible to reconstruct the edges of the genomic network. This essay describes three modelling and inference strategies to answer this central biological question.

7.1 Introduction

Networks have become an important paradigm to describe genomic systems: from describing the physical, molecular interactions between proteins to the abstract interactions between functional genetic units, the vocabulary of networks has been adopted eagerly by biologists tasked with studying complex biological systems. For example, Corominas et al. (2014) define the concept of spliceform networks for translating genetic knowledge into a better understanding of human diseases, whereas Costanzo et al. (2016) argues that a global genetic interaction network highlights the functional organization of a cell and provides a resource for predicting gene and pathway function.

Within biostatistics, mathematical biology and, more recently, bioinformatics, there have been a number of modelling and inference procedures proposed to capture genetic networks. Traditionally, metabolic pathway analysis has been using ordinary

E. C. Wit (✉)
Institute of Computational Science, Università Della Svizzera Italiana,
Lugano, Switzerland
e-mail: wite@usi.ch

F. Biagini et al. (eds.), *Network Science,* https://doi.org/10.1007/978-3-030-26814-5_7

differential equation models, or simplifications thereof, such as flux balance analyses (Papoutsakis 1984). This involved typically small network representations of a number of intertwined genetic pathways. With the advent of high-throughput genomic analysis, Boolean network representations of the transcription process became popular (Akutsu et al. 1999). More recently, stochastic differential equation models (Purutçuoğlu and Wit 2008; Wilkinson 2006), graphical models (Vinciotti et al. 2016), Bayesian networks (Grzegorczyk and Husmeier 2011) and vector autoregressive models (Abegaz and Wit 2013) have entered the scene.

In this chapter, we aim to introduce the reader to the way networks are being used in the analysis of biological systems. In Sect. 7.2, we describe a number of ways on how to think about various genomic systems *as* networks. In Sect. 7.3, we connect those systems with mathematical network models *and* high-throughput genomic data by showing what kinds of inference strategies are available for analysing those processes.

7.2 What Are Genomic Networks?

The language of genomic networks can be used in various ways, although roughly speaking biologists use "genes" as the nodes, connected by edges, which stands for some type of "genetic interactions". This may seem obvious, but the devil is in the details and there are various ways in which this can be made precise. Below we will consider a number of genomic networks, that each uses the concept of network in a somewhat different way.

Mechanic genomic networks
First, and perhaps, the most basal form of a genomic network is the molecular interactions between DNA, RNA and proteins. The interactions in this view are the physical binding of proteins to each other and to DNA, whereby the molecular functionality of the resulting molecule changes and leads to further downstream changes. This cascade of molecular interactions is typically initiated by outside forces, such as sunlight in the case of a circadian clock, the lack of water leading to a stress response in plants or the intake of food leading to production of energy by our mitochondria.

Figure 7.1 is an example of this first type of genomic network. It shows a simplified version of the MAPK-Erk pathway, which is a chain of proteins that via physical interactions carries the signal from a receptor on the surface of a cell to the DNA in the nucleus. It is a ubiquitous pathway and appears in the cell of many organisms. Malfunctioning of the MAPK-Erk pathway in humans has been linked to uncontrolled cell growth, and therefore cancer (Downward 2003). Understanding the activation, inhibition and feedback mechanisms in this network is, therefore, an important goal, which has already led to various drug targets (Hilger et al. 2002).

This mechanistic view of a genomic network is highly localized. The interactions described are individual binding events within a cell. Because of this, the event boundary of the network is typically the cell wall.

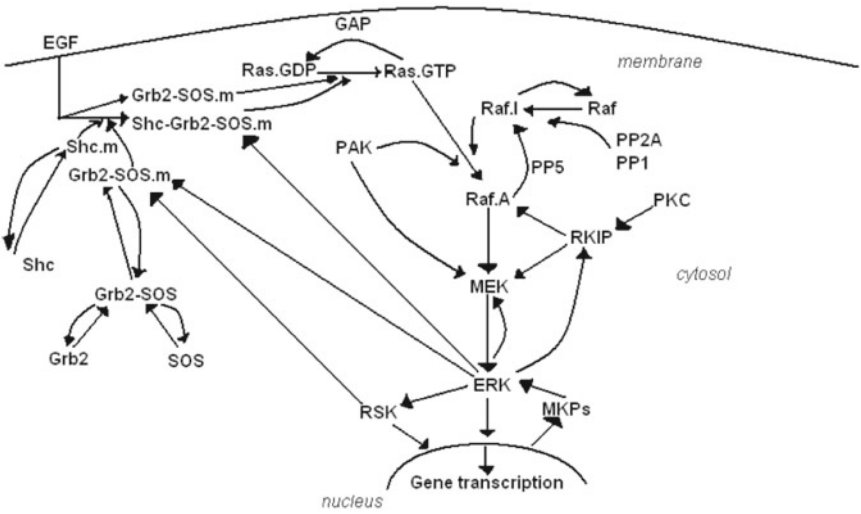

Fig. 7.1 Representation of the single-cell dynamics of the MAPK-Erk Network

Functional genomic networks

In contrast to the mechanistic description of a genomic network is a functional description. Although the nodes of this network can again be proteins or RNA, it is not uncommon that the nodes in this network are abstractly described as "genes". Typically, the focus is on larger systems than a single cell, such as organs or other biological subsystems. Interactions do typically not refer to specific mechanistic binding events, but rather to functional relationships. Often these networks are referred to as *gene regulatory networks*.

Just like the mechanistic genomic networks, the functional genomic network is most naturally interpreted as a dynamic process. However, whereas the changes in the mechanistic network are typically discrete, referring to a particular binding event, the functional network is more naturally seen as continuous, also referred to as a *flow network*.

Evolutionary networks

There are other genomic processes that can be described as a network, for example, how genes get passed on from generation to generation in the presence of genetic variability and selection. Most studied organisms are *diploid*, i.e. organisms that carry two copies of each gene. These copies can be the same, in which case we refer to them as *homozygous*, or different, in which case we refer to them as *heterozygous*. Mendel suggested that offspring receive a randomly selected version of each gene from either parent. Clearly, if the genetic make-up for a particular gene of the parents is the same and homozygous, then the offspring will be homozygous for that gene too. However, for many genetic loci within all species there is genetic variation, which means that offspring displays a "random" mosaic of the genetic make-up of their parents. Various constellations of this mosaic may lead to genetic advantage or

disadvantage for the organism. This will boost or suppress the presence of particular genotype combinations, which can be detected as *dependence*, or in the language of networks: an interaction, between pairs of genes. This network of evolutionary "interactions" defines an evolutionary network.

7.3 Stochastic Models for Genomic Networks

Although the aim of this section is not to give a *comprehensive* overview of network models in genomics, it does aim to provide an introduction to the type of models that are suited to various types of modern genomic data. In fact, we argue that the sampling scheme and design of a genomic experiment should match the type of model that is used for analysing it. In this chapter, we outline three modelling strategies, that are useful in various aspects of this enterprise. We start in Subsect. 7.3.1 with a system of stochastic differential equations to describe single-cell interactions, which takes into account the underlying stochasticity of genomic particle interactions. Often, however, genomic data is collected at either a more agglomerated level or across a number of cells that are destructively sampled. In those cases, temporal models are more appropriately described by means of ordinary differential equations, described in Subsect. 7.3.2.1. In large genomic systems, both SDE and ODE descriptions can be unstable or computationally prohibitive. In such cases, vector autoregressive models, described in Subsect. 7.3.2.2, are useful. All these models are inherently dynamic. Nevertheless, the genotype is, at ordinary time-scales, a non-dynamic process, in which case it is more appropriate to describe these genomic interactions by means of a static network. This and other final considerations are described in Subsect. 7.3.3.

7.3.1 Modelling Mechanistic Genomic Networks

A cell is a natural unit of biology, whose state varies according to external influences and to internal regulation. The process of carrying over a signal, i.e. information, in the cell's environment is regulated by various signal transduction pathways. This signalling process is typically started by an external stimulus of the pathway leading to a binding of the signal to a receptor, i.e. hormones or growth factors, and ends by binding of a target protein. All cellular decisions such as cell proliferation, differentiation, or apoptosis are directed by different levels of transductions (Hornberg 2005). Deregulation of a single "renegade" cell can lead to diseases such as cancers, neurological disorders and developmental disorders (Macaulay et al. 2017).

Sequencing technologies now permit profiling the genome (Gawad et al. 2016), epigenome (Schwartzman and Tanay 2015), transcriptome (Stegle et al. 2015), or proteome (Wu and Singh 2012) of single cells sampled from heterogeneous cell types and cellular states. This allows us to study biological processes, such as disease development, at the cellular level. The technology is subject to measurement noise,

but more importantly, the single-cellular process itself contains intrinsic stochasticity: the cellular system, characterized by its external environment and its internal protein levels, started at the same state may develop in different ways, merely by chance.

It is our aim to describe on the one hand the structured interactions between molecular particles and on the other hand the stochasticity involved in this process. We do this by means of a collection of random reaction equations. A general single-cellular, biochemical reaction can be defined as

$$k_1 Q_1 + k_2 Q_2 + \ldots + k_l Q_l \xrightarrow{\theta} s_1 P_1 + s_2 P_2 + \ldots + s_p P_p, \tag{7.1}$$

where the terms on the left side, denoted as Q, are called the *reactants* and the ones on the right side, denoted as P, are named the *products*. The coefficients k_i $(i = 1, \ldots, l)$ and s_j $(j = 1, \ldots, p)$ represent the *stoichiometric coefficients* associated with the ith reactant Q_i and the jth product P_j, respectively. The quantity l refers the number of required reactants and p stands for the number of resulting products. So the chemical interpretation of this equation is that while molecules move around randomly in a cellular environment k_1 molecules of type Q_1, k_2 molecules of type Q_2, etc., "collide" with each other and produce s_1 molecules of type P_1, s_2 molecules of type P_2, etc. (Wilkinson 2006). Therefore under thermal equilibrium and fixed volume, a biochemical reaction shows which species and in what proportions react together and what they produce (Bower and Bolouri 2001).

For a set of r reactions and d species, accordingly, we can show the molecular transfer from reactant to product species as a net change of $V = S - K$ where V is called the $d \times r$ dimensional *net-effect* matrix when S denotes the $d \times r$ dimensional matrix of stoichiometry of products and K is the $d \times r$ dimensional matrix of stoichiometry of reactants. A reaction corresponds to a directed edge between the nodes $(Q_1, \ldots, Q_l)$ on the one hand and the nodes $(P_1, \ldots, P_p)$ on the other. The collection of r reactions, therefore, corresponds to a network with r directed edges between the d species or nodes of the network. This set of reactions can also contain uncertain, hypothesized reactions or even competing hypotheses, as shown in Fig. 7.2. This network is a representation of the potential stoichiometry between three proteins. The inference procedure with sufficient amount of data will eventually assign a zero reaction rate θ to reactions that are not part of the true underlying system. For example, if the reaction rate θ_1 associated with reaction 1 is inferred to be zero, then the resulting network would only involve the two reactions that are part of the second pathway. An over-parameterized system is, therefore, not a problem a priori and could be a modelling strategy to learn not only the kinetic parameters of the genomic system, but also the structure of the system.

We collect the amount of d reactants and products at time t in the vector U_t. They are put together in the same vector because products of one reaction are the reactants of another. There is therefore no fundamental difference between reactants and products. In the genomic context, they are typically proteins, protein complexes, enzymes, RNA and DNA. The aim is to define a probabilistic model for the evolution of the temporal process $\{U_t\}_t$. This is done by means of the master equation.

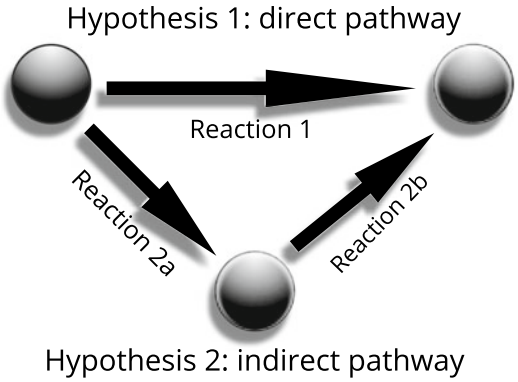

Fig. 7.2 The stochastic differential equation models could include competing hypotheses. The data would eventually weed out the links for which there is no evidence

The master equation is defined as a differential equation for the process transition probability and is written as:

$$\frac{dP(\boldsymbol{U};t)}{dt} = \sum_{k=1}^{r} \{h_k(\boldsymbol{U} - \boldsymbol{V_{\cdot k}}, \boldsymbol{\theta}) P(\boldsymbol{U} - \boldsymbol{V_{\cdot k}}, t) - h_k(\boldsymbol{U}, \boldsymbol{\theta}) P(\boldsymbol{U}, t)\}. \quad (7.2)$$

In other words, the probability of being in state U_t is positively related to the tendency of the r available reactions to transit to state U_t and negatively related to these same reactions to leave state U_t. The hazard h_k is a deterministic function of the state and the reaction rate θ_k. For example, the reaction

$$2H + O \xrightarrow{\theta} H_2O,$$

in a volume with 5 hydrogen molecules H, 4 oxygen molecules and a rate of $\theta = 2$ reactions per time unit would lead to a hazard $h((5,4),2) = \binom{5}{2}\binom{4}{1}2 = 80$. By means of a multivariate Taylor expansion, it is possible to derive an equivalent and alternative formulation of any master equation, named the Kramers–Moyal expansion (Van Kampen 1981):

$$\frac{dP(\boldsymbol{U};t)}{dt} = \sum_{m=1}^{\infty} \frac{(-1)^m}{m!} \sum_{j_1,\ldots,j_m=1}^{N} \frac{d^m}{dU_{j_1},\ldots,dU_{j_m}} [\boldsymbol{a_m}(\boldsymbol{U},\boldsymbol{\theta}) P(\boldsymbol{U},t)], \quad (7.3)$$

where $\boldsymbol{a_m}(\boldsymbol{U})$ are m-order symmetric tensors commonly called *jump moments* (Moyal 1949) or *propagator moment functions* (Gillespie 1992).

Various approximations to the process are possible. We can expand the distribution $P(U,t)$ by a second-order Taylor expansion and use a Fokker–Planck approach for the change of each state (Bower and Bolouri 2001; Van Kampen 1981). This

stochastic expression is solved via Itô or Stratonovich integrals (Gillespie 1996; Golightly and Wilkinson 2005; Risken 1984; Van Kampen 1981) to obtain the following diffusion approximation

$$dU(t) = \mu(U_t, \theta)dt + \beta^{\frac{1}{2}}(U_t, \theta)dW(t), \tag{7.4}$$

where

$$\mu(U, \theta) = V'h(U, \theta),$$
$$\beta(U, \theta) = V'\text{diag}\{h(U, \theta)\}V$$

are the *drift* and *diffusion* matrices, respectively, both explicitly depending on state $U_t = (U_{t1}, \ldots, U_{td})$ at time t, the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_r)'$ and the net-effect matrix V. The expression $dW(t)$ represents the change of a Brownian motion during the time interval dt and $dU(t)$ shows the change in state U over time dt. This is effectively a large volume approximation that follows from the central limit theorem, whereby the reactions follow a Poisson process with rate $h(U, \theta)$ and the states changes therefore have mean $V'h(U, \theta)$ and variance $V'\text{diag}\{h(U, \theta)\}V$.

Due to the difficulties of inference of continuous-time multivariate diffusions, a further discrete Euler–Maruyama approximation is possible,

$$\Delta U_t = \mu(U_t, \theta)\Delta t + \beta^{\frac{1}{2}}(U_t, \theta)\Delta W_t \tag{7.5}$$

where ΔU_t is the change of state U over small time interval $[t, t + \Delta t]$ and ΔW_t is a d-dimensional independent identically distributed Gaussian random vector $\Delta W_t \sim N(0, I\Delta t)$ (Eraker 2001).

Data

The genomic interactions described above form a continuous-time process $\{U_t\}_t$ of gene activities on top of a genomic network. At best, we will be able to see snapshots X_t from this process. We will assume that we will have discrete observations $\{X_t\}_t$ from a single-cell genomic system $\{U_t\}_t$. For simplicity of presentation, we assume that the observations are equally spaced at regular time intervals of steps of size $\Delta t = 1$. This is merely for notational simplicity and not important for the inferential methods we use. There may be two types of missing values: first of all, several substrates may not be observed. It is quite common that due to technological limitations or experimental errors, it is not possible to measure the activity of all genomic species of interest. Various experimental techniques, such as microarrays, Chip-Seq analysis or mass-spectroscopy, have limitations to what they can measure. Furthermore, as most current technologies are capable of only discrete snapshots, the non-observed time points can also be considered missing.

Inference

There are various approaches possible for inference in such systems. The main issue the methods need to deal with is that the rate of change of the process is typically

faster than the observation rate, which leads to nonlinearities between the observation times. Frequentist approaches typically rely on the conditional nonlinear first and second moments of the process to propose a method of moments estimator for the reaction rates. To define a method of moment estimator via a generalized least squares objective function that can be minimized in order to estimate the unknown parameters vector $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left(\boldsymbol{X}_{1:T} - \boldsymbol{m}(\boldsymbol{\theta})\right)' \boldsymbol{W}^{-1} \left(\boldsymbol{X}_{1:T} - \boldsymbol{m}(\boldsymbol{\theta})\right) \qquad s.t.\, \boldsymbol{\theta} \geq \mathbf{0}_r$$

where

$$\boldsymbol{X}_{1:T} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_T \end{bmatrix} \text{ and } \boldsymbol{m}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{m}(1;\boldsymbol{\theta}) \\ \boldsymbol{m}(2;\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{m}(T;\boldsymbol{\theta}) \end{bmatrix}$$

are dT-dimensional column vectors with the observed cell-type count data and predicted mean evolutions, respectively. (Sotiropoulos and Kaznessis 2011) provide a general schema to derive analytical expressions for jump moments for any Markov process. Furthermore,

$$\boldsymbol{W} = \begin{bmatrix} b(\boldsymbol{X}_0;\boldsymbol{\theta}) & 0 & \dots & 0 \\ 0 & b(\boldsymbol{X}_1;\boldsymbol{\theta}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b(\boldsymbol{X}_{T-1};\boldsymbol{\theta}) \end{bmatrix}.$$

is a $dT \times dT$ block diagonal matrix, in which blocks correspond to expected variance-covariance matrices and zeros reflect the independence among measurements belonging to different time points.

An alternative way to deal with partially observed process is defining an augmented state space in combination with Bayesian inference. By inserting intermediate, unobserved states, the process can be linearized in the augmented, latent space. In a Bayesian approach to infer the kinetic parameters θ of the stochastic differential equation, one can use MCMC inference for calculating the posterior of the Euler–Maruyama system described in (7.5). Typically, Gibbs sampling can be difficult, because of the above-described data sparsity. In principle, it is possible to augment the data X with "missing" observations Z. A large number of augmented states in the Bayesian method increases the precision of the Euler–Maruyama approximation, but deteriorates the mixing of the Markov chain. Additional details about this problem and suggested solutions can be found in (Roberts and Stramer 2001) and (Golightly and Wilkinson 2008). In order to deal with these types of missingness, one can use a *Metropolis-within-Gibbs* step (Carlin and Louis 2000), whereby a Metropolis-Hastings step is implemented at each Gibbs step of the update. Therefore, the augmented process $U = \{U_t\}_{t=1}^T$ is a combination of X and Z, i.e. $U = (X, Z)$.

This method has been applied to estimating the MAPK-Erk pathway, consisting of 35 measured proteins and 16 unmeasured proteins across 77 time points that are involved in 66 reactions. Part of the inferred system is shown in Fig. 7.1. (Purutçuoğlu and Wit 2008) describe the biological interpretation of the results.

7.3.2 Modelling Functional Genomic Networks

Single-cell data, especially longitudinal single-cell data, are not very common. In fact, more often time-course genomic data are measured across a collection of cells. Moreover, not infrequently the measurements at different time points are on physically different samples. For example, various petri dishes with cells from some cell line are treated at a nominal time zero, and at various time points, the various dishes, one by one, are measured on the expression of their genomic constituents. As in many cases sampling tends to be destructive, each petri dish can be only measured once. This can be seen as cross-sectional sampling, where time is considered the factor of interest.

7.3.2.1 Ordinary Differential Equation Models

In such cases, it is not sensible to consider the stochastic relatedness between the various time points. However, it can still be interesting to consider the average dynamic behaviour of a genomic system. In fact, consider a simple reversible reaction,

$$A + B \overset{\theta_b, \theta_f}{\longleftrightarrow} C,$$

where proteins A and B bind with forward rate θ_f into protein complex C, and, reversely, protein C breaks apart into constituents A and B with backward rate θ_b. According to the *Law of Mass Action* (Érdi and Tóth 1989), the average change in the amount of substrate A at time t_0 is negative proportional to the number of times forward reactions can happen, i.e. $a \times b$, and positively proportional to the number of times backward reactions can happen, i.e. c, where $A_t = a$, $B_t = b$ and $C_t = c$. This leads to the simple expression for the average change in A,

$$\frac{dm_A(t,\theta)}{dt} = c\theta_b - ab\theta_f.$$

Similarly, for B and C we have,

$$\frac{dm_B(t,\theta)}{dt} = c\theta_b - ab\theta_f,$$
$$\frac{dm_C(t,\theta)}{dt} = ab\theta_f + c\theta_b.$$

However, whereas the *Law of Mass Action* suggests a linear increase in the production rate of the product with an increase of the underlying substrate, in practice the increase will saturate. One reason is that there is only a finite amount of enzymes available, which are crucial auxiliary components in the genomic transcription system. (Michaelis and Menten 1913) introduced an intermediate substrate–enzyme complex, $C = SE$, in the transcriptional system,

$$S + E \overset{\theta_b, \theta_{f_1}}{\longleftrightarrow} C \overset{\theta_{f_2}}{\longleftrightarrow} P + E.$$

Combining the assumption of a finite amount of enzyme, $C + E =$ constant, with a *mass action* equilibrium $(\theta_b + \theta_{f_2})C_t = \theta_{f_1} S_t E_t$, they derived the so-called nonlinear Michaelis–Menten kinetics,

$$\frac{dm_P(t, \theta)}{dt} = \frac{\theta_{f_2} s}{\frac{\theta_b + \theta_{f_2}}{\theta_{f_1}} + s}.$$

Fig. 7.3 shows the typical saturation effect of the Michaelis–Menten production rate. This shows that for realistic descriptions of genomic interactions, we may have to consider a wider class of functions beyond *mass action* kinetics.

For the purposes of this overview, we will focus on a class of nonlinear ODEs that are linear in the rate parameters. Any of the models satisfying the *Law of Mass Action* satisfy also this requirement, but the class is larger than that and can accommodate saturation effects. Consider the gene regulatory or signalling network, described by a system of ordinary differential equations of the form

$$\begin{cases} z'(t) = \boldsymbol{g}(z(t))\boldsymbol{\theta} \text{ for } t \in [0, T], \\ z(0) = \boldsymbol{\xi}, \end{cases} \tag{7.6}$$

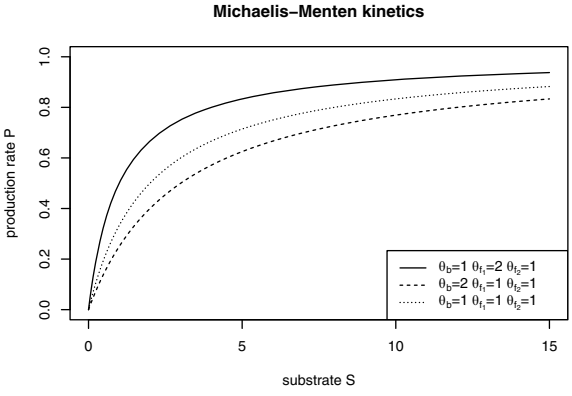

Fig. 7.3 Typical saturation of the Michaelis–Menten production rate for various choices of the kinetic parameters

where $\boldsymbol{x}(t)$ takes values in $\mathbb{R}^d$, with, possibly unknown, initial values $\xi \in \mathbb{R}^d$, and with the parameters of interest unknown $\boldsymbol{\theta} \in \mathbb{R}^p$. We assume that $\boldsymbol{g} = (g_1, \ldots, g_d)'$ is a known function, whose components $g_j : \mathbb{R}^d \to \mathbb{R}^p$. In particular, we consider a special case in which we want to model the change of each substrate by a saturating function of all the other substrates, i.e.

$$\begin{cases} z_j'(t) = \sum_{k=1}^{d} \theta_{kj} \log(z_k(t) + 1), \\ z_j(0) = \xi_j \end{cases} \quad j = 1, \ldots, d. \tag{7.7}$$

This model defines a network between the d substrates in that if $\theta_{jk} \neq 0$, then substrate k affects the change in substrate j. The logarithmic function is chosen to deal with natural saturation effects. Moreover, by its very definition $z_k(t) \geq 0$ and the leading Taylor term of $\log(z + 1)$ near zero is z, similar to the Michaelis–Menten production term. The solution $z(\cdot, \theta, \xi)$ implied by the ODE (7.7) — or more generally (7.6) — is assumed to be the mean of the observations taken from the system. In particular, we assume that at time points $t_i \in [0, T]$, $i = 1, \ldots, n$, we observe

$$X_j(t_i) = z_j(t_i, \boldsymbol{\theta}, \xi) + \varepsilon_j(t_i), \quad j = 1, \ldots, d_1; i = 1, \ldots, n, \tag{7.8}$$

where $0 \leq t_1 < \cdots < t_n = T < \infty$ and $\varepsilon_i(t_j)$ is the measurement error for x_i at time t_j. The problem is to estimate $\boldsymbol{\theta}$, and thereby the underlying gene regulatory network, from the data $\{X_j(t_i)\}_{ij}$.

Inference of ODE networks

Inference of parameters in ODEs is not straightforward due to the possibly computationally prohibitive calculation of ODE solution $z(\cdot, \theta, \xi)$ for lots of values of θ and ξ. Regularization-based approaches, which make use of properties of differential operators, have been proposed to avoid numerical integration of the system of differential equations (González et al. 2013, 2014; Steinke and Schölkopf 2008). In most cases, the main computational bottleneck lies in the optimization of a nonlinear objective function. Alternatively, the idea of smoothing can be used as a way to avoid numerical integration of the system of differential equations and is referred to as the *collocation* estimation method; for example, there are *two-step* methods (Bellman and Roth 1971; Brunel 2008; Dattner and Klaassen 2013; Fang et al. 2011; Gugushvili and Klaassen 2012; Gugushvili and Spreij 2012; Liang and Wu 2008; Varah 1982) and *generalized profiling* methods (Ramsay et al.2007; Qi and Zhao 2010; Xun et al. 2011; Hooker et al. 2013).

The method we present here is a special case of generalized Tikhonov regularization (Vujačić et al. 2016) and without penalization has been shown to be $\sqrt{n}$-consistent (Vujačić et al. 2015). We consider estimators of the parameters θ and ξ that are obtained by minimizing the integral equation derived from (7.6),

$$L(\xi, \theta) = \int_0^T \left\| z(t) - \xi - \int_0^t g(z(s))\, \mathrm{d}s\, \theta \right\|^2 \mathrm{d}t, \tag{7.9}$$

with respect to ξ and θ, where $z(t) = (t; \theta, \xi)$ will be replaced by a suitable estimator. We divide the interval $[0, T]$ in $\lfloor\sqrt{n}\rfloor$ subintervals, so that in every interval, we have at least $\lfloor\sqrt{n}\rfloor$ observations in it. Let $S_i = [a_{i-1}, a_i)$ be the ith subinterval $i = 1, \ldots, \lfloor\sqrt{n}\rfloor - 1$ and $S_{\lfloor\sqrt{n}\rfloor} = [a_{\lfloor\sqrt{n}\rfloor-1}, a_{\lfloor\sqrt{n}\rfloor}]$ and let $S(t)$ denote the subinterval to which t belongs. The piecewise constant *window estimator* of z is defined as

$$\hat{z}(t) = \frac{1}{|S(t)|} \sum_{t_j \in S(t)} X(t_j), \qquad t \in S(t). \tag{7.10}$$

This estimator $\hat{z}(t)$ estimates $z(t)$ as the mean of the observations that belong to interval $S(t)$. This allows us to estimate the inner integral in (7.9),

$$\begin{aligned} G(t) &= \int_0^t g(\hat{z}(s))ds \\ &= \sum_{m=1}^{i-1} g(\hat{z}(S_m))(a_m - a_{m-1}) + g(\hat{z}(S_i))(t - a_{i-1}), \qquad \text{where } t \in S_i. \end{aligned}$$

Throughout the paper, we adhere to the convention that the sums of the form $\sum_{m=1}^{i-1} f_m$ are equal to zero for $i = 1$. Minimizing the criterion function (7.9) with respect to $\omega = (\xi; \theta)'$ yields explicit formulas for the estimators of the parameters. Indeed, the objective function L can be written as a quadratic function of the parameters,

$$L(\omega) = \omega' \int_0^T F(t)'F(t)\mathrm{d}t\omega - 2\omega' \int_0^T F(t)'\hat{z}(t)\mathrm{d}t + \int_0^T \|\hat{z}(t)\|^2\mathrm{d}t,$$

where $F(t) = (TI_d; G(t))$. The minimizer of this quadratic expression is given by

$$\hat{\omega} = \Big(\int_0^T F(t)'F(t)\mathrm{d}t \Big)^{-1} \int_0^T F(t)'\hat{z}(t)\mathrm{d}t$$

which has an explicit form by means of finite sums as shown in (Vujačić et al. 2015). It can be shown that this estimator is $\sqrt{n}$-consistent.

Example 7.1 *Circadian clock in Arabidopsis*

Consider the previously introduced, heavily parameterized ODE describing the change of each substrate in the gene regulatory network by a slowly saturating function of all the other substrates, i.e.

$$\begin{cases} z_j'(t) = \sum_{k=1}^d \theta_{kj} \log(z_k(t) + 1), \\ z_j(0) = \xi_j \end{cases} \qquad j = 1, \ldots, d, \tag{7.11}$$

or using some other production terms, such as $g(z) = \sqrt{z}$ or simply $g(z) = z$. This relatively simple gene regulatory network contains d^2 interaction parameters $\theta =$

$\{\theta_{kj}\}$. Many of these parameters can be expected to be zero as only a few genes will be responsible for activating other genes.

To enforce sparsity, we will add a L_1 regularization term on the objective function (7.9),

$$L_\lambda(\theta, \xi) = L(\theta, \xi) + \lambda||\theta||_1.$$

The estimator of θ and ξ will depend on the tuning parameter λ. In fact, the path estimator $(\hat{\xi}_\lambda, \hat{\theta}_\lambda)$ will correspond to the original lasso estimator $\hat{\beta}_\lambda$ for a quadratic problem (Tibshirani 1996),

$$\hat{\beta}_\lambda = \arg\min_\beta (y - X\beta)'(y - X\beta) + \lambda||\beta||_1,$$

whereby $X'X = \int_0^T F(t)'F(t)\mathrm{d}t$ and $X'y = \int_0^T F(t)'\hat{z}(t)\mathrm{d}t$, whereby the first d parameters, corresponding to ξ, will not be penalized and always included in the solution path.

We illustrate our proposed approach by applying it to a time-course gene expression dataset related to the study of circadian regulation in plants. The data used in our study come from the EU project TiMet (FP7-245143, 2014), whose objective is the elucidation of the interaction between circadian regulation and metabolism in plants.

The data consist of transcription profiles for 9 core clock genes from the leafs of various genetic variants of *Arabidopsis thaliana*. The plants were grown in 3 light conditions: a diurnal cycle with 12-hour light and 12-hour darkness (12L/12D), an extended night with full darkness for 24 hours, and an extended light with constant light for 24 hours. Samples were taken every 2 hours to measure mRNA concentrations. In total, there are 51 measurements across time. The nine genes are known to be involved in circadian regulation (Grzegorczyk et al. 2008; Aderhold et al. 2014). They consist of two groups of genes: "Morning genes", which are LHY, CCA1, PRR9 and PRR5, whose expression peaks in the morning, and "Evening genes", including TOC1, ELF4, ELF3, GI and PRR3, whose expression peaks in the evening. The expressions for all the genes are strictly positive.

Figure 7.4 shows the resulting sparse ODE network inferred with three different functions g, two of which deal explicitly with possible saturation effects, such as $g(x) = \log(x + 1)$ and $g(x) = \sqrt{x}$ and the naive linear production function $g(x) = x$. The results are quite robust, but suggest that it is worth considering possible saturation effects.

7.3.2.2 Vector Autoregressive Models

Both SDE and ODE models are in principle generative models for the underlying process of interest. Their aim is to describe the intrinsic relationship between the genomic substrates, typically on the basis of the *Law of Mass Action* or extensions thereof. Often, part of the model is inspired by biological knowledge. In this section,

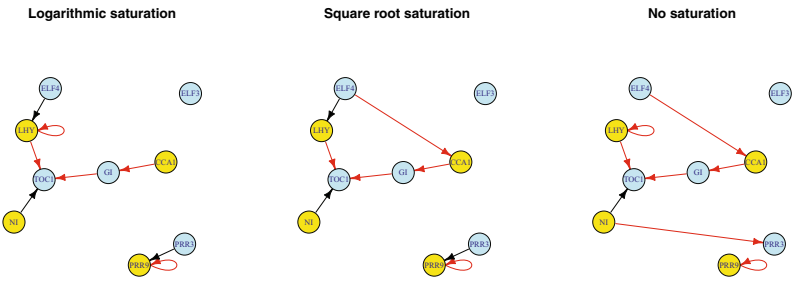

Fig. 7.4 Circadian clock network in *Arabidopsis thaliana*: red arrows represent suppression, whereas black arrow suggests activation. The ODE network inference results are quite robust, whether one considers a saturation model, whereby the effect on the production term depends on $g(z) = \log(z+1)$ or $g(z) = \sqrt{z}$, or one that does not saturate, in which case $g(z) = z$. The yellow genes are the morning genes, whereas the blue genes are the evening genes

we describe a method fundamentally aimed at a more exploratory approach of high-dimensional genomic time series data. The idea is to explore potential temporal interactions between substrates, without focusing on the details of the kinetics. For this, we will use vector autoregressive models (VARs), which have been studied more in detail in the econometric literature (Dahlhaus and Eichler 2003). The details of the method described in this section can be found in (Abegaz and Wit 2013).

Within a vector autoregressive model, the time-course gene–gene interactions evolve according to Markovian dynamics, rather than an explicit functional form as in the ODE approach. Specifically, within a VAR(1) model the vector of gene expressions at time t relates only to those at time $t-1$; extensions to a Markovian lag dependence greater than 1 are straightforward. Let X_t be a d-dimensional random vector associated with the expression of the d genes at time t. According to the first-order Markov property, the joint probability density of $X_0, \ldots, X_T$ can be decomposed as:

$$f(X_0, \ldots, X_T) = f(X_0) f(X_1 \mid X_0) \times \cdots \times f(X_T \mid X_{T-1}). \tag{7.12}$$

We focus only on the conditional distributions in (7.12) and ignore the initial term $f(X_0)$. Furthermore, we assume a time-homogeneous dynamic network structure for the conditional distribution $f(X_t \mid X_{t-1})$ that can be approximated via a multivariate Gaussian,

$$X_t \mid X_{t-1} \sim N(\Gamma X_{t-1}, \Sigma). \tag{7.13}$$

This vector autoregressive process of order one can also be expressed as

$$X_t = \Gamma X_{t-1} + \epsilon_t, \tag{7.14}$$

where $\epsilon_t \sim N(0, \Sigma)$. The parameter elements in the matrices Γ and in the inverse of Σ represent directed and undirected links in the Markovian conditional independence

graph, respectively. In particular, a nonzero element in Γ, say $\Gamma_{ij} \neq 0$, corresponds to a directed edge in the conditional independence graph between gene j at the previous time point and gene i at the current one. This edge is given the name *Granger causality* and reflects a delayed interaction between two genes, which can be cautiously given a semi-causal interpretation (Granger 1988). Given Σ and the corresponding precision matrix $\Theta = \Sigma^{-1}$ undirected edges relate to nonzero elements in the precision matrix Θ. If $\Theta_{ij} \neq 0$, then after adjusting for the past and present effects of other genes, there is an instantaneous interaction, or dependence, between genes i and j. A cartoon representation of the model formulation is given in Fig. 7.5.

Data

Suppose that we have n replications of a T time point longitudinal microarray study across p genes. The data, then, can be summarized as an $n \times p \times T$ array $X = (X_1, \ldots, X_n)'$ whose ith submatrix X_i has columns such that $X_{i.t} = (X_{i1t}, \ldots, X_{ipt})'$ which correspond to the expression levels of p genes measured at time t. That is, X_{ijt} is the jth gene expression level at time t for the ith replicate.

Sparse VAR network inference

The inference aim is to reconstruct the dynamic and contemporaneous genomic networks. Time-course genomic data typically consist of hundreds or thousands of genes measured on a comparatively small number of replications (typically 3) of microarray experiments across a few time steps (often not more than 10). The model formulation in (7.14) is in a standard vector autoregressive form with correlated errors and estimation approach for high-dimensional time-course genomic data is challenging. (Abegaz and Wit 2013) proposes a penalized maximum likelihood estimation methods for the analysis of the high-dimensional time-course gene expression data. The

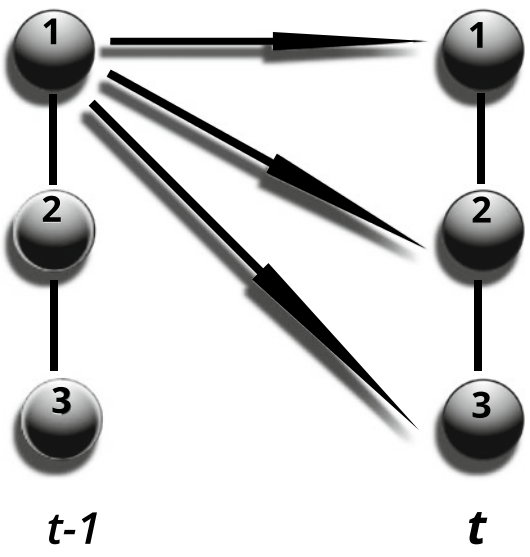

Fig. 7.5 The dynamic network encoded in Γ shows that gene 1 is an important regulator. The instantaneous network Θ shows a central role of gene 2, but because the genomic interaction times are faster than the sampling times δt, it is not possible to say whether gene 2 regulates the other or the other way around

proposed approach provides sparse estimates of the autoregressive coefficient matrix Γ and the precision matrix Θ in (7.14), which are used to reconstruct the genomic network.

Under the Gaussian assumption described in (7.13), the conditional density of the tth observation is given by

$$f_c\left(X_t \mid X_{t-1}; \Gamma, \Theta\right) = (2\pi)^{p/2}|\Theta|^{1/2} \exp\left[-\frac{1}{2}\left(X_t - \Gamma X_{t-1}\right)' \Theta \left(X_t - \Gamma X_{t-1}\right)\right].$$

Then the conditional log-likelihood for n replicates each at T time steps becomes

$$\begin{aligned}\ell(\Gamma, \Theta) &= \sum_{i=1}^{n}\sum_{t=1}^{T} \log f_c\left(X_{it} \mid X_{i,t-1}; \Gamma, \Theta\right) \\ &= -\frac{npT}{2}\log(2\pi) + \frac{nT}{2}\log|\Theta| - \frac{nT}{2} tr(S_\Gamma \Theta), \end{aligned} \tag{7.15}$$

where

$$S_\Gamma = (1/nT)\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X_{it} - \Gamma X_{i,t-1}\right)\left(X_{it} - \Gamma X_{i,t-1}\right)'.$$

We consider a penalized likelihood framework, where the objective function based on (7.15) is defined as

$$\ell_{pen}(\Gamma, \Theta) = \log|\Theta| - tr(S_\Gamma \Theta) - \sum_{i\neq j}^{p} P_\lambda(|\theta_{ij}|) - \sum_{i\neq j}^{p} P_\rho(|\gamma_{ij}|), \tag{7.16}$$

where θ_{ij} and γ_{ij} are the (i, j)-elements of the matrix Θ and Γ and λ and ρ are the corresponding tuning parameters of the penalty functions $P_\lambda(\cdot)$ and $P_\rho(\cdot)$ corresponding to Θ and Γ. Various penalty functions have been proposed in the literature. We consider the L_1 penalty function, which is convex and given by

$$P_\lambda(\theta) = \lambda|\theta|, \quad P_\rho(\gamma) = \rho|\gamma|. \tag{7.17}$$

This leads to a desirable convex optimization problem. To obtain the L_1 penalized likelihood we substitute the penalty function in (7.17) into the objective function (7.16). Then, the optimization problem that gives sparse estimates of Γ and Θ is the solution of

$$(\widehat{\Theta}, \widehat{\Gamma})_{\lambda,\rho} = \arg\max_{\Theta,\Gamma}\left\{\log|\Theta| - tr(S_\Gamma\Theta) - \lambda\sum_{i\neq j}^{p}|\theta_{ij}| - \rho\sum_{i,j}^{p}|\gamma_{ij}|\right\}. \tag{7.18}$$

Model selection

Under the penalized maximum likelihood framework for time series chain graphical models, the sparsity of the estimated precision matrix Θ and the autoregressive coefficient matrix Γ are controlled by the tuning parameters λ and ρ. The Bayesian information criterion can be used for selecting parsimonious parameter representations (Yin and Li 2011). The BIC is defined as

$$BIC(\lambda, \rho) = -nT \left\{ \log |\widehat{\Theta}_\lambda| - tr(S_{\widehat{\Gamma}_\rho} \widehat{\Theta}_\lambda) \right\} + \log(nT)(a_n/2 + b_n + p), \quad (7.19)$$

where p is the number of variables, a_n is the number of nonzero off-diagonal elements of $\hat{\Theta}_\lambda$ and b_n is the number of nonzero elements of $\hat{\Gamma}_\rho$. Thus, we select the values of λ and ρ that minimizes the criterion in (7.19). Here the minimization of $BIC(\lambda, \rho)$ with respect to λ and ρ is achieved by a grid search.

Example 7.2 *Mammary gland gene expression network*

We illustrate the proposed approach on the analysis of mammary gland gene expression time-course data from (Stein et al. 2004). In the mammary gland expression experiment, there are 12,488 probe sets representing approximately 8,600 genes. These probe sets are measured over 54 arrays of 3 replicates on each of 18 time points. We identified 30 genes that yield the best separation between the four developmental stages (virgin, pregnant, lactating, involution) using cluster analysis. We implemented the sparse VAR procedure in the `R` package `SparseTSCGM`. We apply the proposed VAR model to study the interaction between these crucial genes that trigger the transitions to the main developmental events in the mammary gland of mice. Fig. 7.6(a) shows the undirected links associated with Θ, related to instantaneous interactions among the genes and Fig. 7.6(b) displays the directed links that indicate Granger causality relations among the genes.

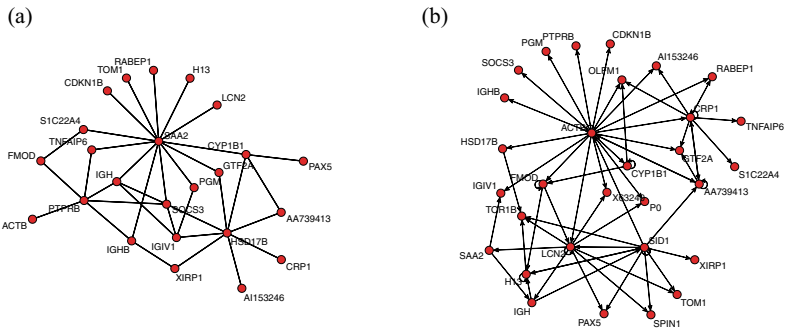

Fig. 7.6 Undirected (left) and directed (right) time series chain graphical model network inferred from the mammary gland time-course expression data with a VAR(1) model

7.3.3 Other Genomic Network Models

The models we have considered so far have all been dynamic network models. The main reason is that these models capture the dynamic nature of the genetic process. Depending on external stimuli and the internal state of a cell, The main reason is that these models capture the dynamic nature of the genetic process: at each moment the then relevant genes are transcribed, translated and broken down again in an intricate, interdependent process. Nevertheless, the three models that we have discussed are not the only ones that can be used. Some people might have noticed that we did not explicitly deal with Bayesian network models. Although they are closely related to vector autoregressive processes, the biostatistics and bioinformatics literature has seen many fine examples of such models applied to gene regulatory systems (Grzegorczyk and Husmeier 2011).

At the same time, there are also certain biological processes that can be modelled very elegantly by means of static network models. Genome-wide association studies (GWAS) are aimed at uncovering associations between genotype and phenotype. At the same time, certain genotype combinations might be evolutionary very advantageous or, more likely, detrimental. That is why such GWAS data can also be used to study epistasis by inferring the conditional independence graph: if there is no epistasis, the conditional independence graph will show the chromosomal backbone, whereas, if there is some epistasis, then we will find additional links between regions of the genome that are possibly on different chromosomes. Figure 7.7 shows such an example in *Arabidopsis thaliana*, which has been found by means of L_1 penalized Gaussian copula graphical modelling (Behrouzi and Wit 2017).

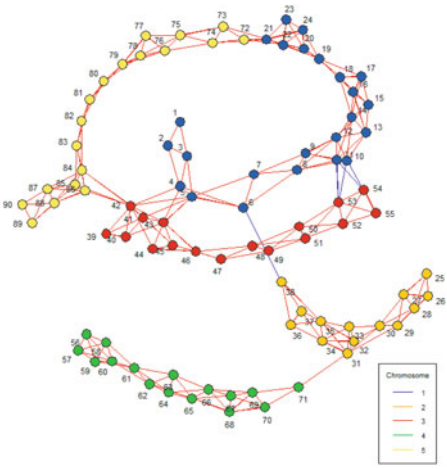

Fig. 7.7 Epistatic effects in a genotype study involving an *Arabidopsis thaliana* recombinant inbred line. The sparse Gaussian copula graphical model clearly shows the chromosomal backbone in the conditional dependency graph as a result of the meiosis process

7.4 Discussion

In this chapter, we have looked at modelling dynamic biological networks. Unlike in social networks, this typically does not involve random graph models. The reason is that the biological phenomena of interest, such as gene transcription, pertain to the nodes of the network, rather than the edges. In other words, the random process of interest lives on the vertices of the graph. For this reason, the network models we have considered in this chapter are more closely connected to engineering networks used to describe flows.

Although networks have become an important modelling paradigm in genomics, there is currently no single network model to describe all the genomic interaction structures. In fact, it will be unlikely that there will ever be one. As the underlying generative model in biology is extremely complicated, we will always rely on convenient parameterizations to answer specific questions that arise in system biology. We have considered three types of models, namely stochastic differential equation models, ordinary differential equation models and vector autoregressive models and each of these modelling frameworks was selected depending on the underlying sampling design ("Are the measurements from a single cell or average over many cells?") and on the question of interest ("Do we want to describe the kinetics of the interactions or get an idea of the overall interaction structure of the genome?"). As George Box is said to have once said "all models are wrong, but some are useful" (Wit et al. 2012), and very useful indeed.

References

Abegaz, F. & Wit, E. (2013), 'Sparse time series chain graphical models for reconstructing genetic networks', *Biostatistics* **14**(3), 586–599.

Aderhold, A., Husmeier, D. & Grzegorczyk, M. (2014), 'Statistical inference of regulatory networks for circadian regulation', *Statistical applications in genetics and molecular biology* **13**(3), 227–273.

Akutsu, T., Miyano, S. & Kuhara, S. (1999), 'Identification of genetic networks from a small number of gene expression patterns under the Boolean network model', *Pacific Symposium on Biocomputing* pp. 17–28.

Behrouzi, P. & Wit, E. C. (2017), 'Detecting epistatic selection with partially observed genotype data by using copula graphical models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Bellman, R. & Roth, R. S. (1971), 'The use of splines with unknown end points in the identification of systems', *Journal of Mathematical Analysis and Applications* **34**(1), 26–33.

Bower, J. M. & Bolouri, H. (2001), *Computational Modelling of Genetic and Biochemical Networks*, 2nd edn, Massachusetts Institute of Technology.

Brunel, N. J-B (2008), 'Parameter estimation of ODE's via nonparametric estimators', *Electronic Journal of Statistics* **2**, 1242–1267.

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

Corominas, R., Yang, X., Lin, G. N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S. A. et al. (2014) , 'Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism', *Nature communications* **5**.

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D. et al. (2016), 'A global genetic interaction network maps a wiring diagram of cellular function', *Science* **353**(6306), aaf1420.

Dahlhaus, R. & Eichler, M. (2003), Causality and graphical models in time series analysis, *in* R. S., ed., 'Highly Structured Stochastic Systems', Oxford University Press, pp. 115–137.

Dattner, I. & Klaassen, C. A. (2013), 'Estimation in systems of ordinary differential equations linear in the parameters', arXiv:1305.4126 .

Downward, J. (2003), 'Targeting RAS signalling pathways in cancer therapy', *Nature Reviews Cancer* **3**(1), 11.

Eraker, B. (2001), '$MCMC$ analysis of diffusion models with application to finance', *Journal of Business and Economic Statistics* **19**(2), 177–191.

Érdi, P. & Tóth, J. (1989), *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*, Manchester University Press.

Fang, Y., Wu, H. & Zhu, L.-X. (2011), 'A two-stage estimation method for random coefficient differential equation models with application to longitudinal HIV dynamic data', *Statistica Sinica* **21**(3), 1145–1170.

Gawad, C., Koh, W. & Quake, S. R. (2016), 'Single-cell genome sequencing: current state of the science', *Nature reviews. Genetics* **17**(3), 175.

Gillespie, D. (1992), *Markov processes: An introduction for physical scientists.*, Academic Press.

Gillespie, D. T. (1996), 'The multivariate Langevin and Fokker-Planck equations', *American Journal of Physics* **64**(10), 1246–1257.

Golightly, A. & Wilkinson, D. J. (2005), 'Bayesian inference for stochastic kinetic models using a diffusion approximation', *Biometrics* **61**(3), 781–788.

Golightly, A. & Wilkinson, D. J. (2008), 'Bayesian inference for nonlinear multivariate diffusion models observed with error', *Computational Statistics and Data Analysis* **52**(3), 1674–1693.

González, J., Vujačić, I. & Wit, E. (2013), 'Inferring latent gene regulatory network kinetics', *Statistical applications in genetics and molecular biology* **12**(1), 109–127.

González, J., Vujačić, I. & Wit, E. (2014), 'Reproducing kernel Hilbert space based estimation of systems of ordinary differential equations', *Pattern Recognition Letters* **45**, 26–32.

Granger, C. W. (1988), 'Causality, cointegration, and control', *Journal of Economic Dynamics and Control* **12**(2-3), 551–559.

Grzegorczyk, M. & Husmeier, D. (2011), 'Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes', *Bioinformatics* **27**(5), 693–699.

Grzegorczyk, M., Husmeier, D., Edwards, K. D., Ghazal, P. & Millar, A. J. (2008), 'Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler', *Bioinformatics* **24**(18), 2071–2078.

Gugushvili, S. & Klaassen, C. A. J. (2012), '$\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing', *Bernoulli* **18**, 1061–1098.

Gugushvili, S. & Spreij, P. (2012), 'Parametric inference for stochastic differential equations: a smooth and match approach', *ALEA* **9**(2), 609–635.

Hilger, R., Scheulen, M. & Strumberg, D. (2002), 'The Ras-Raf-MEK-ERK pathway in the treatment of cancer', *Oncology Research and Treatment* **25**(6), 511–518.

Hooker, G., Ellner, S., Earn, D. et al. (2011), 'Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in ontario.', *Journal of the Royal Society, Interface* **8**(60), 961–974.

Hornberg, J. J. (2005), Towards integrative tumor cell biology control of MAP kinase signalling, PhD thesis, Vrije Universiteit, Amsterdam.

Liang, H. & Wu, H. (2008), 'Parameter estimation for differential equation models using a framework of measurement error in regression models', *Journal of the American Statistical Association* **103**(484), 1570–1583.

Macaulay, I. C., Ponting, C. P. & Voet, T. (2017), 'Single-cell multiomics: multiple measurements from single cells', *Trends in Genetics*.

Michaelis, L. & Menten, M. L. (1913), 'The kinetics of the inversion effect', *Biochem. Z* **49**, 333–369.

Moyal, J. (1949), 'Stochastic processes and statistical physics.', *Journal of the Royal Statistical Society. Series B* **11**, 150–210.

Papoutsakis, E. T. (1984), 'Equations and calculations for fermentations of butyric acid bacteria', *Biotechnology and bioengineering* **26**(2), 174–187.

Purutçuoğlu, V. & Wit, E. (2008), 'Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters', *Bayesian Analysis* **3**(4), 851–886.

Qi, X. & Zhao, H. (2010), 'Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations', *The Annals of Statistics* **38**(1), 435–481.

Ramsay, J. O., Hooker, G., Campbell, D. & Cao, J. (2007), 'Parameter estimation for differential equations: a generalized smoothing approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(5), 741–796.

Risken, H. (1984), *The Fokker-Planck Equation: Methods of Solution and Applications*, Springer-Verlag.

Roberts, G. O. & Stramer, O. (2001), 'On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm', *Biometrika* **88**(3), 603–621.

Schwartzman, O. & Tanay, A. (2015), 'Single-cell epigenomics: techniques and emerging applications', *Nature reviews. Genetics* **16**(12), 716.

Sotiropoulos, V. & Kaznessis, Y. (2011), 'Analytical derivation of moment equations in stochastic chemical kinetics.', *Chemical engineering science* **66**(3), 268–277.

Stegle, O., Teichmann, S. A. & Marioni, J. C. (2015), 'Computational and analytical challenges in single-cell transcriptomics', *Nature reviews. Genetics* **16**(3), 133.

Stein, T., Morris, J. S., Davies, C. R., Weber-Hall, S. J., Duffy, M.-A., Heath, V. J., Bell, A. K., Ferrier, R. K., Sandilands, G. P. & Gusterson, B. A. (2004), 'Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving lbp, cd14 and stat3', *Breast Cancer Res* **6**, R75–R91.

Steinke, F. & Schölkopf, B. (2008), 'Kernels, regularization and differential equations', *Pattern Recognition* **41**(11), 3271–3286.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Van Kampen, N. G. (1981), *Stochastic Processes in Physics and Chemistry*, North-Holland, Amsterdam.

Varah, J. (1982), 'A spline least squares method for numerical parameter estimation in differential equations', *SIAM Journal on Scientific and Statistical Computing* **3**(1), 28–46.

Vinciotti, V., Augugliaro, L., Abbruzzo, A. & Wit, E. C. (2016), 'Model selection for factorial Gaussian graphical models with an application to dynamic regulatory networks', *Statistical applications in genetics and molecular biology* **15**(3), 193–212.

Vujačić, I., Dattner, I., González, J. & Wit, E. (2015), 'Time-course window estimator for ordinary differential equations linear in the parameters', *Statistics and Computing* **25**(6), 1057–1070.

Vujačić, I., Mahmoudi, S. M. & Wit, E. (2016), 'Generalized Tikhonov regularization in estimation of ordinary differential equations models', *Stat* **5**(1), 132–143.

Wilkinson, D. J. (2006), *Stochastic Modelling for Systems Biology*, Chapman and Hall/CRC.

Wit, E., Heuvel, E. v. d. & Romeijn, J.-W. (2012), "All models are wrong...': an introduction to model uncertainty', *Statistica Neerlandica* **66**(3), 217–236.

Wu, M. & Singh, A. K. (2012), 'Single-cell protein analysis', *Current Opinion in Biotechnology* **23**(1), 83–88.

Xun, X., Cao, J., Mallick, B., Maity, A. & Carroll, R. J. (2013), 'Parameter estimation of partial differential equation models', *Journal of the American Statistical Association* **108**(503), 1009–1020.

Yin, J. & Li, H. (2011), 'A sparse conditional Gaussian graphical model for analysis of genetical genomics data', *The Annals of Applied Statistics* **5**(4), 2630.