Bayesian Analysis (2008) **3**, Number 4, pp. 851–886

Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters

Vilda Purutçuoğlu* and Ernst Wit†

Abstract. The MAPK/ERK pathway is one of the major signal transduction systems which regulates the cellular growth control of all eukaryotes like the cell proliferation and the apoptosis. Because of its importance in cellular lifecycle, it has been studied intensively, resulting in a number of qualitative descriptions of this regulatory mechanism. In this study we describe the MAPK/ERK pathway as an explicit set of reactions by combining different sources. Our reaction set takes into account the localization and different binding sites of the molecules in the cell by implementing the multiple parametrization. Then we estimate the model parameters of the network in a Bayesian setting via MCMC and data augmentation schemes. In the estimation we apply the Euler approximation, which is the discretized version of the diffusion technique. Additionally in inference of such a realistic and complex system we consider all possible kinds of dependencies coming from distinct stages of updates. To test the inference method we use the simulated data generated by the Gillespie algorithm. From the analysis it is clear that the sampler mixes well and partially is able to identify the dynamics of the MAPK/ERK pathway.

Keywords: MCMC, MAPK/ERK pathway, diffusion approximation, data augmentation, dependency in diffusion matrix

1 Introduction

The biochemical reaction is the discrete event which is occurred by molecular collisions in continuous time. A set of reactions which builds a system can be mathematically modelled in different ways (Orton et al. 2005). There are mainly three types of techniques to model biochemical reactions. These are the Boolean, differential equations, and stochastic methods (Bower and Bolouri 2001). Among these main approaches the random nature of microscopic molecular collisions is taken into account by stochastic methods. These methods constitute a probabilistic model of the reaction kinetics, thereby capture the small and heterogenous environment of reactions (Turner et al. 2004). In stochastic methods the reaction rate constants (Section 2.2) have crucial importance in the analysis of a system.

The stochastic property of biochemical reactions can be generated by different exact simulation techniques
(Gillespie 1977, 1992; Gibson and Bruck 2000; Morton-Firth and Bray 1998). The Gille-

*Middle East Technical University, Ankara, Turkey, mailto:vpurutcu@metu.edu.tr
†Groningen University, Groningen, The Netherlands

DOI:10.1214/08-BA332

spie algorithm (Gillespie 1977) is an accurate and the most common simulator. However its performance in estimation is not computationally efficient (Gillespie 2001; Tian and Burrage 2004; Cao et al. 2005). The diffusion method is an approximation technique whose performance is not as accurate as Gillespie in simulation, whereas, is computationally less demanding in estimation (Wilkinson 2006; Golightly and Wilkinson 2005). Therefore in this study we implement it to infer the model parameters of a specific network structure. Our network of interest is the MAPK/ERK pathway.

In estimation of the model parameters, i.e. stochastic rate constants denoted as c_j, we typically face with both quite imprecise and scarce observations in addition to the associated levels of uncertainty. The underlying uncertainty is mainly caused by distinct sources of variations for measuring proteins (Wit and McClure 2004), and the limited knowledge of the system (Vyshemirsky et al. 2006; Endy and Brent 2001; Brent 2005). In order to make inference under the available information and to deal with the problem of missing data, we use the Bayesian methodology based on the diffusion approximation.

In the presentation of this study we use the following structure. Section 2 provides the details about the MAPK/ERK pathway and its modelling via the stochastic approach. Section 3 outlines the formulation of the diffusion approximation, the description of the hierarchical model, and details about the update of the system. In Section 4, we propose two algorithms for the MAPK/ERK pathway considering its realistic complexity. Then in Section 5 we describe the simulated data which we use in inference. We present our results from the application of this dataset in Section 6. Section 7 concludes and discusses possible extensions of the proposed algorithms.

2 Modelling the MAPK/ERK pathway

2.1 Features of the MAPK/ERK pathway

The cellular signal transduction is the process of carrying over of information (signal) in the cell's environment for taking an appropriate response (Lawrence 2005). 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 is ended up by a binding of a target protein. All cellular decisions such as the cell proliferation, which refers to a frequent and repeated reproduction of the cell, the differentiation, which is the development of the cell with specialized structure, or the apoptosis, which implies the cell death as a result of an intracellular suicide programme, are directed by different levels of transductions (Hornberg 2005). Because of their underlying importance in the cellular lifecycle, any malfunction in these structures has a direct influence on the expression or on the function of gene products which are components of these regulatory mechanisms. As a result it may lead to many illnesses such as heart diseases and cancer (Kolch 2000; Schoeberl et al. 2002). Therefore the knowledge about pathways can be very helpful for understanding the behaviour of distinct biological activations and for developing drugs, which target the proteins involved in associated illnesses (Hornberg

2005).

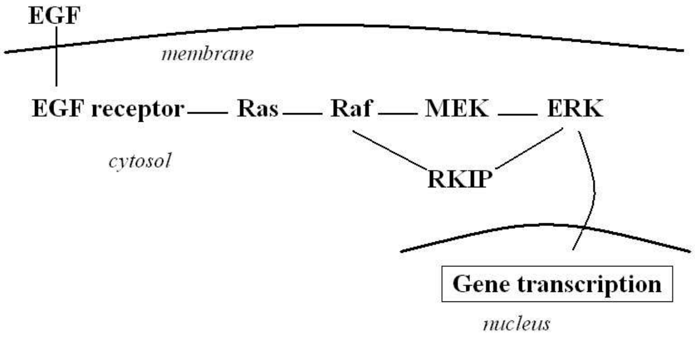


Figure 1: Main components of the MAPK/ERK pathway.

The MAPK (mitogen-activated protein kinase) or its synonymous ERK (extracellular signal regulated kinase) pathway is one of the major signal transduction systems which regulates the cellular growth control of all eukaryotes from the reproduction to the death of the cell. The major components of this mechanism involve Ras, Raf, MEK, and ERK proteins (Fig. 1). But apart from these substrates, there are a number of other species in the activation of the pathway (Fig. 2).

In the MAPK system phosphorylations at various locations in the cell, such as near the cell membrane or in the cytosol, typically enable the activation or inhibition of subsequent major proteins in the chain, resulting in a regulatory flow. An external stimulus of the EGF (growth factor) protein, which binds to the activated tyrosine receptor (EGFR), triggers the activation of the pathway. Then this signal is transferred within the cell until it arrives in the nucleus to produce the target protein c-Fos. During this process, most components in the system are regulated by directly ERK and RKIP proteins or ERK and RKIP with SOS, Raf or MEK proteins. Moreover negative and positive feedback loops are typically used (Kolch et al. 2005; Yeung et al. 1999, 2000).

The proteomic functionality in the MAPK pathway is stochastic in nature. Also the structure of the pathway is too complex for implementing a simple representation (Fig. 2) to explain its organizational behaviours. The characteristics of this system involve non-linear features like ultrasensitivities, bistabilities, periodic behaviours and existences of location-depended proteins (Kolch et al. 2005).

The MAPK pathway has been intensively studied, particularly, in cancer researches (Kolch 2000; Yeung et al. 1999; Chang and Karin 2001; Schoeberl et al. 2002). Thus there are lots of biological sources which describe this mechanism. However most of these sources give qualitative information about the structure and do not explain the system by an explicit set of reactions. In this study by combining the qualitative knowledge, we represent biochemical activations of the pathway as a list of reactions (Purutçuoğlu and Wit 2006). However we call our list a *quasi list* due to the fact

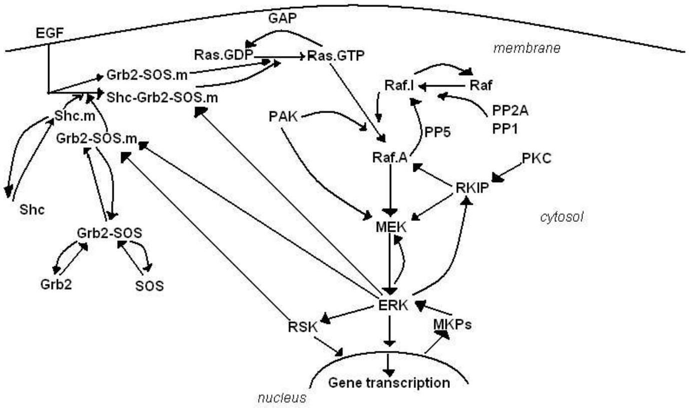


Figure 2: Simple representation of the structure of the MAPK/ERK pathway.

that it can explain the main topology of the pathway according to the current biological theories. While biochemists do not know yet the complete picture about all interactions, they continue to discover more details about the underlying system (Vyshemirsky et al. 2006). In this set of reactions we indicate each protein by a simple notation. For instance known theories accept that Raf protein becomes active when its inhibitory binding side is closed before it is recruited from the cytosol to the membrane by membrane resident GT-Pase Ras. PP2A, a protein located both at the cell membrane and in the cytosol, takes away the inhibitory phosphorylate of the inactive Raf at the S259 site, thereby enables Raf to be made active. (Kolch et al. 2005; Kolch 2000). So when we use this qualitative information, we summarize it as $\text{Raf} + \text{PP2A} \longrightarrow \text{Raf.I} + \text{PP2A}$ denoting that Raf indicates the inactive and non-phosphorylated Raf protein in the cytosol and Raf.I is the inactive Raf phosphorylated on the S259 binding site in the cytosol.

Furthermore as a novelty we implement multiple parametrization in order to express distinct localizations of the protein in the cell and to describe the protein using different binding sites and various phosphorylations. Indeed the importance of this kind of spatial localization for the substrate has been mentioned by earlier studies of Endy and Brent (2001) and it has been suggested that such an artificial construction of a real system would be the only way to understand the actual function of the pathway and to exhibit its complex dynamical behaviour (Endy and Brent 2001; Brent 2004). As implementations of the multiple parametrization, we use the abbreviation m to denote the translocation of the protein from the cytosol to the membrane. On the other hand different levels of the phosphorylation of the same protein are represented by the index p or p1 and p2 in which the first two notations indicate the mono phosphorylation and the latter shows the double phosphorylation. The set of reactions given in Section 2.2 is an example of this description for the MAPK system. The complete list of reactions, on the other side, is presented in the Appendix.

2.2 Stochastic approach for modelling the pathway

A general biochemical reaction can be defined as

$$k_1 R_1 + k_2 R_2 + \ldots + k_l R_l \xrightarrow{c} s_1 P_1 + s_2 P_2 + \ldots + s_p P_p \tag{1}$$

where the terms on the left side, denoted as R, are called the *reactants* and the ones on the right side, denoted as P, are named as the *products*. Accordingly l refers to the number of required reactants and p stands for the number of resulting products. k_i $(i = 1, \ldots, l)$ is the number of molecules of R_i consumed in a single reaction step, whereas s_j $(j = 1, \ldots, p)$ represents the number of molecules of P_j produced in a single reaction step. These k_i and s_j terms are also known as the *stoichiometric coefficients* associated with the ith reactant R_i and the jth product P_j, respectively. Finally c is the *stochastic reaction rate constant* which indicates the speed of the execution of the reaction and depends on physical properties of reactants and the temperature of the system (Gillespie 1977). So the chemical interpretation of this equation is that k_i molecules of the type R_i collide with each other and produce s_j molecules of the type P_j with speed c while molecules move around randomly by the Brownian motion (Wilkinson 2006). Therefore under a thermal equilibrium and fixed volume a biochemical reaction shows which species and in what proportions react together and what they produce in a particular speed (Bower and Bolouri 2001; Wilkinson 2006).

For a set of r reactions and n 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 $n \times r$ dimensional *net effect* matrix when S denotes the $n \times r$ dimensional matrix of stoichiometries of products and K is the $n \times r$ dimensional matrix of stoichiometries of reactants.

For instance, in the following set of equations, we demonstrate the activation of the MAPK pathway by EGF receptor (EGFR) with $r = 6$ reactions and $n = 10$ species.

(a) $\text{EGFR} + \text{Shc} \xrightarrow{c_1} \text{EGFR} + \text{Shc}_m$

(b) $\text{Grb2} + \text{SOS} \xrightarrow{c_2} \text{Grb2-SOS}$

(c) $\text{EGFR} + \text{Grb2-SOS} \xrightarrow{c_3} \text{EGFR} + \text{Grb2-SOS}_m$

(d) $\text{Shc}_m + \text{Grb2-SOS}_m \xrightarrow{c_4} \text{Shc-Grb2-SOS}_m$

(e) $\text{Shc-Grb2-SOS}_m + \text{Ras.GDP} \xrightarrow{c_5} \text{Shc-Grb2-SOS}_m + \text{Ras.GTP}$

(f) $\text{Grb2-SOS}_m + \text{Ras.GDP} \xrightarrow{c_6} \text{Grb2-SOS}_m + \text{Ras.GTP}$.

Here Grb2, SOS, Shc, Ras.GTP, and Ras.GDP are single proteins and Gr2-SOS, Shc-Grb2-SOS are protein complexes in the cytosol. From the implementation of the multiple parametrization we use Shc_m and Shc-Grb2-SOS_m for the single protein Shc and protein complex Shc-Grb2-SOS, respectively, near the cell membrane. Similarly MEK and MEK.p2 proteins describe the inactive and non-phosphorylated MEK and the double phosphorylated MEK (active MEK) on the S218 and S222 binding sites in the cytosol, in the order given. With respect to this explanation we present the 10×6 dimensional reactant K, product S, and net effect V matrix as follows (for the species EGFR, Shc, Shc$_m$, Grb2, SOS, Grb2-SOS, Grb2-SOS$_m$, Shc-Grb2-SOS$_m$, Ras.GDP, and Ras.GTP, respectively, and the reactions from (a) to (f)):

$$K = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, S = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

and

$$V = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

In these representations, for instance, the first column of K shows stoichiometric coefficients of reactants of the first reaction (Reaction (a)). In this equation one molecule of EGFR and Shc proteins are consumed in a single reaction step and the remaining proteins are not used. Therefore we put 1 in associated rows of EGFR and Shc species in the first column and set to 0 for the rest. The first column of S, on the other hand, displays stoichiometric coefficients of products of Reaction (a). So seeing that in each reaction step a single molecule of EGFR and Shc$_m$ proteins is produced, corresponding rows of EGFR and Shc$_m$ species in the first column of S are equated to 1, and the remaining rows set to 0. Finally in the first column of V, we summarize the net change of the system as a result of Reaction (a). As the net change is found by the substraction of molecules produced and used in the associated single reaction step, we take the difference between the first columns of S and K. This calculation corresponds to $(1, 0, 1, 0, 0, 0, 0, 0, 0, 0)' - (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)' = (0, -1, 1, 0, 0, 0, 0, 0, 0, 0)'$ which is written in the first column of V where $(')$ indicates the transpose of the selected vector.

In a biological sense, on the other side, this reaction set states that the activation of the MAPK pathway is triggered by an external stimulus of the growth factor EGF which binds to activated tyrosine kinase receptors. EGFR phosphorylates Shc, hereby, recruits it from the cytosol to the cell membrane (Reaction (a)). In the cytosol, SOS, whose function is an exchange factor, forms a complex with the adaptor protein Grb2 (Reaction (b)). This complex (Grb2-SOS) is then phosphorylated and recruited by EGFR (Reaction (c)). By this way SOS in the cytosol is translocated near the cell membrane (Grb2-SOS$_m$) where it enables to activate Ras. During these reactions Shc pathway (Reaction (a)) and Grb2 pathway (Reaction (a)-(b)) can run in parallel at the same time and can bind in the membrane to make a complex (Reaction (d)). Finally SOS, which forms a complex with either adaptor proteins Shc and Grb2 (Reaction (e)) or only adaptor protein Grb2 (Reaction (f)), activates Ras by promoting the exchange of GDP for GTP.

Similarly, we define our pathway by 51 species in which 34 of them are major proteins and 94 reactions where 65 of them represent changes in activities and translocations of species, and the rest indicates their degradations after dissociation. The full list of proteins and the quasi set of reactions are given in the Appendix.

When we model a biochemical system like the MAPK pathway, the randomness of molecular collisions is captured by stochastic approaches. The stochastic behaviour is included into the model via the *master equation* (Wilkinson 2006; Turner et al. 2004) which describes the stochasticity by

$$\frac{\partial P(Y,t)}{\partial t} = \sum_{j=1}^{r} \{ h_j(Y - v_j, \Theta) P(Y - v_j, t) - h_j(Y, \Theta) P(Y, t) \} \tag{2}$$

where the n-dimensional vector $Y = (Y_1, Y_2, \ldots, Y_n)$ represents the state of the system at time t, thereby $P(Y,t)$ is the probability distribution of states which is described by discrete number of molecules and continuous time t. $\Theta = (c_1, c_2, \ldots, c_r)$ stands for the r-dimensional vector of reaction rates, and v_j denotes the jth column of the net effect matrix V. n and r show the total number of substrates and the total number of reactions in the system, respectively. Accordingly $h_j(Y, \Theta)$ describes the hazard, also called the *rate law of reaction*, which is the product of the number of distinct molecular reactant combinations available in the state Y with the stochastic rate constant Θ for the reaction j, i.e. c_j. For instance, the hazard of Reaction (b) at time t is computed by $h_2(Y, \Theta) = c_2 \times \binom{[\mathrm{Grb2}]}{1} \times \binom{[\mathrm{SOS}]}{1}$ where $\binom{[\mathrm{A}]}{b}$ denotes the molecular combination and [A] stands for the number of present molecules of A. c_2, on the other hand, is the stochastic reaction rate constant of this reaction. As a result the term $h_j(Y - v_j, \Theta) P(Y - v_j, t)$ indicates the probability that the jth reaction occurs over time interval $[t, t + dt]$ moving the state from $Y - v_j$ to Y (Turner et al. 2004; Kampen 1981).

3 Inference of the system

Under the assumption of continuous number of molecules Y, the probability distribution of the number of molecules at time t, i.e. $P(Y,t)$, can be described via differential equation models. If the probability distribution $P(Y,t)$ is expanded by a second-order Taylor expansion and the change in states for each species at t is found by a Fokker-Planck approach (Kampen 1981; Bower and Bolouri 2001) and finally this stochastic expression is solved by Itô or Stratonovich integrals (Kampen 1981; Risken 1984; Gillespie 1996; Golightly and Wilkinson 2005), we get the following diffusion formulation of the system

$$dY(t) = \mu(Y,\Theta)dt + \beta^{\frac{1}{2}}(Y,\Theta)dW(t). \tag{3}$$

In a diffusion equation $\mu(Y,\Theta) = V'h(Y,\Theta)$ and $\beta(Y,\Theta) = V'\text{diag}\{h(Y,\Theta)\}V$ are *mean*, or *drift*, and *variance*, or *diffusion*, matrices, respectively, both explicitly depending on states $Y = (Y_1,\ldots,Y_n)$ at time t and the parameter vector $\Theta = (c_1,c_2,\ldots,c_r)'$. n and r are the total number of substrates and the total number of reactions in the system, in order as used beforehand. The notation $(')$ in Θ, on the other side, shows the transpose vector of reaction rates. $dW(t)$ represents the change of a Brownian motion during the time interval dt and $dY(t)$ shows the change in state Y over time dt. V is the net effect matrix, accordingly, each row of V, v_j (an n-dimensional vector whose components are v_{ij}, $i = 1,\ldots,n$), represents associated stoichiometric coefficients of the reaction j ($j = 1,\ldots,r$) and similarly V' is the transpose of this matrix. Finally $h(Y,\Theta)$ indicates the r-dimensional vector of hazards whose component $h_j(Y,\Theta)$ stands for the hazard of the jth reaction (Gillespie 2000; Golightly and Wilkinson 2005; Wilkinson 2006).

In a diffusion process, we need continuous time observations. But in practice we have discrete time measurements, thereby we have to employ the discretized version of the diffusion process, so called the Euler-Maruyama approximation represented by

$$\Delta Y_t = \mu(Y_t,\Theta)\Delta t + \beta^{\frac{1}{2}}(Y_t,\Theta)\Delta W_t \tag{4}$$

where ΔY_t is the change of the state Y over small time interval Δt and ΔW_t is an n-dimensional independent identically distributed Brownian random vector $\Delta W_t \sim N(0, I\Delta t)$ (Wilkinson 2006; Eraker 2001).

For the update of the system according to Equation 4, we cannot directly use a Gibbs sampling as we have a large number of missing values (Wilkinson 2006; Golightly and Wilkinson 2005). But we can perform a special type of the Metropolis sampling in which a Metropolis-Hastings step is implemented at each Gibbs step of the update. This algorithm is known as the *Metropolis-within-Gibbs* technique (Carlin and Louis 2000). Furthermore in order to get a more precise estimate from the Euler method, we use the data augmentation for non-observed states by putting latent states within time-course measurements. The right number of augmented states, which balances the high dependence between states and parameters on the one hand and the accuracy of the Euler approximation on the other, is a matter of discussion. Additional details about this problem and the suggested solution in a Bayesian methodology can be found in Roberts and Stramer (2001), and Golightly and Wilkinson (2008).

3.1 Model description

In the estimation of reaction rates of the MAPK pathway we define an observation matrix Y consisting of observed X and unobserved Z components at given time t ($t = t_0, t_1, \ldots, t_T$) with the dimension d_1 and d_2, respectively, without any augmented state. Thus Y can be described as

$$Y = \begin{bmatrix} X_1(t_0) & X_1(t_1) & X_1(t_2) & \ldots & X_1(t_T) \\ X_2(t_0) & X_2(t_1) & X_2(t_2) & \ldots & X_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{d_1}(t_0) & X_{d_1}(t_1) & X_{d_1}(t_2) & \ldots & X_{d_1}(t_T) \\ Z_1(t_0) & Z_1(t_1) & Z_1(t_2) & \ldots & Z_1(t_T) \\ Z_2(t_0) & Z_2(t_1) & Z_2(t_2) & \ldots & Z_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{d_2}(t_0) & Z_{d_2}(t_1) & Z_{d_2}(t_2) & \ldots & Z_{d_2}(t_T) \end{bmatrix}.$$

We increase the number of states, and accordingly the dimension of Y, by introducing augmented states at every Δt time interval between observed states. In our simulated data since the observations are taken at evenly spaced times, the number of augmented states m is the same in every pair of observed time points. Therefore we get the following matrix Y

$$Y = \begin{bmatrix} x_1(t_0) & X_1(t_1) & \ldots & x_1(t_m) & X_1(t_{m+1}) & \ldots & X_1(t_{mT-1}) & x_1(t_T) \\ x_2(t_0) & X_2(t_1) & \ldots & x_2(t_m) & X_2(t_{m+1}) & \ldots & X_2(t_{mT-1}) & x_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{d_1}(t_0) & X_{d_1}(t_1) & \ldots & x_{d_1}(t_m) & X_{d_1}(t_{m+1}) & \ldots & X_{d_1}(t_{mT-1}) & x_{d_1}(t_T) \\ Z_1(t_0) & Z_1(t_1) & \ldots & Z_1(t_m) & Z_1(t_{m+1}) & \ldots & Z_1(t_{mT-1}) & Z_1(t_T) \\ Z_2(t_0) & Z_2(t_1) & \ldots & Z_2(t_m) & Z_2(t_{m+1}) & \ldots & Z_2(t_{mT-1}) & Z_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{d_2}(t_0) & Z_{d_2}(t_1) & \ldots & Z_{d_2}(t_m) & Z_{d_2}(t_{m+1}) & \ldots & Z_{d_2}(t_{mT-1}) & Z_{d_2}(t_T) \end{bmatrix}$$

in which x_i denotes the observed data by observed components, whereas X_i stands for the augmented data by observed components. So each $Y_i \equiv (X_i, Z_i)'$ indicates the ith column of Y. In the estimation we use 20 data points and set $\Delta t = 0.25$, thereby add 3 states between each pair of observed t. In that way the number of columns of Y is enlarged from 20 to 77.

Although rate constants are positive by definition, there is no further information for choosing an appropriate prior distribution for those rates. In order not to favour one Θ over another, a noninformative or *Jeffreys* prior can be seen as an option. But the former may cause the problem of improper prior in inference and the latter is not practical for multivariate situations (Carlin and Louis 2000; Gelman et al. 2004). Therefore, we choose the exponential distribution as a heavy tailed (Exp(1)) prior information for our rates. So our joint posterior density is written by

$$\pi(Y, \Theta) = \frac{\pi(\Theta)\pi(Z_0)\prod_{i=1}^{T} f(Y_i|Y_{i-1}, \Theta)}{\int \pi(\Theta)\pi(Z_0)\prod_{i=1}^{T} f(Y_i|Y_{i-1}, \Theta)d\Theta} \tag{5}$$

where $\pi(Y, \Theta)$ and $\pi(\Theta)$ represent the likelihood and prior density of the vector of model parameters, i.e. stochastic reaction rates, respectively, $\pi(Z_0)$ shows the prior density of Z_0, Z_0 denotes the d_2-dimensional vector of the augmented data by observed components at time zero t_0, and f is the transition density which has the form

$$\begin{aligned} f(Y_i|Y_{i-1}, \Theta) \quad = \quad & |\beta(Y_{i-1}, \Theta)|^{-1/2} \times \exp\left\{ -\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \Theta)\Delta t)' \right. \\ & \left. (\beta(Y_{i-1}, \Theta)\Delta t)^{-1}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \Theta)\Delta t) \right\}. \end{aligned} \tag{6}$$

In Equation 5 the posterior density $\pi(Y, \Theta)$ is practically intractable. In this expression we denote the numerator of π by p which is the unnormalized π. In implementations we use p, rather than π, to infer conditional densities of parameters and latent states since $\pi(Y, \Theta) \propto p(Y, \Theta)$ directly.

3.2 Updates of the system

In the update of the MAPK pathway we use roughly the same sampling steps which are given in the study of Golightly and Wilkinson (2005) and are listed below. In their paper the method has been used for a simple model of the prokaryotic autoregulation which consists of 7 reactions and 5 species in which only one of them is linearly dependent on other species. From the analysis it has been shown that as the number of observations increases and the number of unobserved species decreases, the method works well for estimating parameters.

Step 1: We initialize the model parameters Θ and the augmented data whose components are observed X and unobserved Z values. Then we set the counter of the iteration g to 0.

Step 2: In the update of the model parameters we use the random walk algorithm by drawing the candidate values from the normal distribution.

Step 3: Apart from the final state Y_T, each ith column of Y is updated via the Metropolis-Hastings algorithm. We sample the augmented state, which is composed of missing X and Z terms, from the proposal density $q(Y_i|Y_{i-1}, Y_{i+1})$. Then we calculate the acceptance probability of the candidate state α by using the full conditional density of Y as presented in Equation 10. If the updated state includes observed components $X = x$, the sampling is done from the conditional distribution of Y, denoted as $q(Z_i|x_i, Y_{i-1}, Y_{i+1}, \Theta)$, given the previous and next state as well as x terms in the current state. Then the α is computed as before by using the Metropolis-Hastings step.

On the other hand in the final state of Y, since all previous states are already updated, we implement the Gibbs sampling for simulating new values of Y_T. In this case Y_T is directly generated from the conditional normal distribution.

Step 4: We increase the counter from g to $(g+1)$ and turn back Step 2 until we satisfy the convergence.

After initializing system parameters and missing values as stated in Step 1, we move to Step 2 which updates the model parameter Θ. To update Θ, we select the normal distribution as a candidate generator because of its simplicity. In every iteration of the MCMC run we use block updates to sample the parameter vector, hereby do not face with bad mixing as both the local and global update have. In our block scheme we divide the vector Θ into small and equally-sized groups with the dimension d, and then simulate each d-dimensional group sequentially according to the conditional posterior

$$\begin{aligned}\pi(\Theta|Y) &\propto \prod_{i=1}^{T} f(Y_i|Y_{i-1},\Theta)\pi(\Theta)\\ &\propto \prod_{i=1}^{T} \exp\left\{-\sum_{j=1}^{r}\Theta_j\right\} \times \exp\Bigg\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1},\Theta)\Delta t)'\\ &\qquad (\beta(Y_{i-1},\Theta)\Delta t)^{-1}(Y_i - Y_{i-1} - \mu(Y_{i-1},\Theta)\Delta t)\Bigg\} \times |\beta(Y_{i-1},\Theta)|^{-1/2} \quad (7)\end{aligned}$$

where T and r are the total number of time points and the total number of reactions, respectively. In the MAPK pathway, $T = 77$ and $r = 66$. The system accepts the candidate Θ^* with an acceptance probability

$$\alpha(\Theta,\Theta^*|Y) = \min\left\{1, \frac{\pi(\Theta^*|Y)}{\pi(\Theta|Y)}\right\} \tag{8}$$

in which a proposal Θ^* is proposed via $\Theta_j^* = \Theta_j + w_j$ $(j = 1,\ldots,r)$, where $w_j \sim N(0,\gamma_j)$. The variance γ_j in sampling w_j is a *tuning parameter* which has a significant effect on the mixing property of the algorithm. If γ_j is too small, the acceptance probability given in Equation 8 will be very high but won't explore the posterior efficiently. Whereas if γ_j is too big, then the acceptance probability can be very low. For good mixing in random walk chains although an acceptance rate of around 24% is optimal in univariate cases (Gamerman and Lopes 2006), it is known that with respect to the complexity of the network structure, i.e. the associated high dimensionality of variables, very low ratios such as 5% for some parameters can be tolerable when candidate values for particular reaction rates are hardly proposed. For the MAPK pathway while choosing a sensible tuning parameter for each c_j, we track acceptance rates, α's, separately and adjust them adaptively in the burn-in phase. The candidate proposal variance γ_j is multiplied by 1.1, if α is greater than 60%, and divided by 1.1, if α is lower than our cut-off 5%. On the other hand if the corresponding α is within these two critical values (0.05, 0.60), we keep the current γ_j. We do this adaptation periodically at every 100th iteration during burn-in of the first 20,000 MCMC runs. At the end of 20,000 runs the final γ_j's are kept fixed and used for the rest of the algorithm.

In the update of Θ, the acceptance probability $\alpha(\Theta, \Theta^*|Y)$ of each d-dimensional group is compared with a random value u from the uniform distribution $U(0,1)$, i.e. $u \sim U(0,1)$. The candidate reaction rates are accepted if $u < \alpha$, otherwise current rates are preserved at the $(g+1)$th iteration. After the renewal of those parameters, we move to the update of Y in Step 3.

The candidate generator of Y for the given Θ is sampled from the multivariate normal distribution and is updated column by column as described in Golightly and Wilkinson (2005). The full conditional density for each column of Y, Y_i, is given as

$$\pi(Y_i|Y_{i-1}, Y_{i+1}, \Theta) \propto p(Y_i|Y_{i-1}, Y_{i+1}, \Theta) \tag{9}$$

where

$$\begin{aligned} p(Y_i|Y_{i-1}, Y_{i+1}, \Theta) \quad = \quad & \exp\left\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu_{i-1}\Delta t)'(\beta_{i-1}\Delta t)^{-1}(Y_i - Y_{i-1} - \mu_{i-1}\Delta t)\right\} \\ & \times \exp\left\{-\frac{1}{2}(Y_{i+1} - Y_i - \mu_i\Delta t)'(\beta_i\Delta t)^{-1}(Y_{i+1} - Y_i - \mu_i\Delta t)\right\} \\ & \times |\beta_{i-1}|^{-1/2} \times |\beta_i|^{-1/2}. \end{aligned} \tag{10}$$

In Equation 10, $\mu_{i-1} = \mu(Y_{i-1}, \Theta)$, $\mu_i = \mu(Y_i, \Theta)$, $\beta_{i-1} = \beta(Y_{i-1}, \Theta)$, and finally $\beta_i = \beta(Y_i, \Theta)$.

If Y_i is completely composed of the augmented data, the candidate Y_i^*, $q(.|Y_{i-1}, Y_{i+1}, \Theta)$, which converges pointwise to $\pi(.|Y_{i-1}, Y_{i+1}, \Theta)$ when $\Delta t \to 0$, is generated from

$$Y_i^* \sim N\left(\frac{1}{2}(Y_{i-1} + Y_{i+1}), \frac{1}{2}\Delta t \beta(Y_{i-1}, \Theta)\right) \tag{11}$$

and is accepted with probability $\alpha(Y_i^*|Y_i) = \min\left\{1, \frac{p(Y_i^*|Y_{i-1}, Y_{i+1}, \Theta) q(Y_i|Y_{i-1}, Y_{i+1}, \Theta)}{p(Y_i|Y_{i-1}, Y_{i+1}, \Theta) q(Y_i^*|Y_{i-1}, Y_{i+1}, \Theta)}\right\}$.

If the column i $(i \neq 0, T)$ is partially observed, then we only sample a candidate value for Z_i conditional on $X_i = x_i$, $q(.|x_i, Y_{i-1}, Y_{i+1}, \Theta)$, which converges pointwise to $\pi(.|Y_{i-1}, Y_{i+1}, \Theta)$, via

$$Z_i^* \sim N(\eta_{Z_i^*}, \Sigma_{Z_i^*}) \tag{12}$$

with mean $\eta_{Z_i^*} = \frac{1}{2}(Z_{i-1} + Z_{i+1}) + \beta_{i-1}^{zx}(\beta_{i-1}^{xx})^{-1}(x_i - \frac{1}{2}[X_{i-1} + X_{i+1}])$ and variance $\Sigma_{Z_i^*} = \frac{1}{2}\Delta t(\beta_{i-1}^{zz} - \beta_{i-1}^{zx}(\beta_{i-1}^{xx})^{-1}\beta_{i-1}^{xz})$, then we decide on the acceptance or rejection of the step with probability

$$\alpha(Z_i^*|Z_i) = \min\left\{1, \frac{p(Z_i^*|x_i, Y_{i-1}, Y_{i+1}, \Theta) q(Z_i|x_i, Y_{i-1}, Y_{i+1}, \Theta)}{p(Z_i|x_i, Y_{i-1}, Y_{i+1}, \Theta) q(Z_i^*|x_i, Y_{i-1}, Y_{i+1}, \Theta)}\right\}. \tag{13}$$

Here $\beta^{xx}_{i-1} = \beta(Y^{xx}_{i-1}, \Theta)$ and has full rank, $\beta^{zz}_{i-1} = \beta(Y^{zz}_{i-1}, \Theta)$, $\beta^{zx}_{i-1} = \beta(Y^{zx}_{i-1}, \Theta)$, and $\beta^{xz}_{i-1} = \beta(Y^{xz}_{i-1}, \Theta)$.

On the other hand the proposals of the first ($i = 0$) and the last column ($i = T$) are generated from their associated conditional normal distributions and their acceptance probabilities are calculated similar to Equation 13. More details about candidate generators and corresponding formulations can be found in Golightly and Wilkinson (2005) and Eraker (2001). After the update of the last column $i = T$, we control the convergence of the chain as done in Step 4. If the chain converges, we stop the algorithm, otherwise we turn to Step 2 and move the iteration from g to $g + 1$.

4 MCMC algorithms for the MAPK/ERK pathway

Although we basically apply the MCMC methods described in Section 3.1 and 3.2 for the MAPK pathway, we extend the underlying plan as we face with challenges in the updates because of the complexity of our system. Therefore we suggest two MCMC sampling schemes. The first plan (Scheme 1) rejects any kind of linear dependence in the update. In other word it puts a zero prior probability on a singular diffusion matrix. The second plan (Scheme 2), on the other hand, considers all possible sources of the dependence appearing at distinct stages of updates of time states and model parameters. So when the proposal leads to the dependence, the algorithm calculates the acceptance probability under a nonsingular diffusion matrix which has lower dimension.

4.1 Scheme 1: MCMC runs under complete independence

1. Structural dependence. The *structural dependence* results from the rank of $V'V$ where V is the net effect matrix and V' is its transpose. So the substrates which are linearly dependent on other species and thereby decrease the rank of the diffusion matrix $\beta(Y, \Theta)$ are excluded at the beginning of the algorithm. This dependence causes zero value in the determinant of $\beta(Y, \Theta)$, resulting in problems in the calculation of the likelihood function. Moreover the underlying singularity also causes infeasible candidate generators due to the linear dependence of some state values. Therefore the dependent substrates are eliminated and the MCMC sampler is run only for the substrates that are linearly independent.

2. Incidental dependence. After omitting the dependent substrates at the beginning, all unknowns, assigned for either missing states or reaction rate constants are initialized. These initial values are checked whether they cause any new dependency in the system. We call this second type of the dependence the *incidental dependence* which is due to the rank of $V'\text{diag}\{h(Y, \Theta)\}V$ originating from the numeric value of $h(Y, \Theta)$. $h(Y, \Theta)$ denotes the hazard of the system for a given state $Y = (Y_1, \ldots, Y_n)$ and reaction rate $\Theta = (c_1, \ldots, c_r)$. During the update of Y or Θ, the hazard $h(Y, \Theta)$ can be equal or close to zero such that the product of V and a lower dimensional matrix $\text{diag}\{h(Y, \Theta)\}$ results in a singular matrix.

If the non-singularity of the system is still preserved after the initialization, the iteration counter g is set to 0. Otherwise new initial values are proposed in place of the dependent ones until the singularity completely disappears.

3. Block update. d deviance terms are sampled from the normal $N(0, \gamma_j)$ ($j = 1, \ldots, r$) to generate d candidate values for Θ where d is the number of reaction rates which are simultaneously updated by blocks and γ_j is the tuning parameter of the jth reaction (Section 3.2). The new Θ for each d-dimensional group is tested whether it causes dependency in Y. If the candidate Θ, Θ^*, maintains independence, it is taken as the proposal candidate value for the calculation of the acceptance probability α. Otherwise a new Θ^* is proposed till the underlying condition is satisfied.

4. Acceptance probability. The acceptance probability α of rate constants is evaluated according to Equation 8. If the move is accepted, $\Theta^{(g)} = \Theta^*$, if not, the chain does not move at the gth iteration.

5. Update latent states. After updating the rate constants, the algorithm moves to the update of the state Y by the Metropolis-within-Gibbs sampling. In each augmented or partially missing column, the corresponding candidates, Y_i^* or Z_i^* ($i = 0, 1, \ldots, T$), are checked for being positive definite and are tested beforehand whether their candidate diffusion matrices and $\beta_{i-1}^{xx} = \beta(Y_{i-1}^{xx}, \Theta)$ submatrices where necessary have full rank. As long as they maintain the positivity and the non-singularity, the associated α is calculated by using the associated expression of the given state. Otherwise the candidate state is rejected even before computing the acceptance probability.

6. Iterate. The counter moves from g to $(g + 1)$ and the algorithm is repeated from Step 2 until the chain converges to the stationary distribution.

4.2 Scheme 2: MCMC runs under dependence

Scheme 2 controls the singularity of the system in every MCMC run as already applied in Scheme 1. But, particularly, in the update of reaction rates and state values, different from Scheme 1, the dependencies of each d-dimensional group of reaction rates as well as dependent substrates of the current state Y_i, previously updated state Y_{i-1}, and candidate current state Y_i^* are considered separately. If any dependence is observed, then the calculation of α is merely based on the independent reaction rates and the substrates which are linearly independent of Y_i, Y_{i-1}, and Y_i^* seeing that the accepted new rate or state can have singular diffusion matrices.

The steps below briefly describe our second updating scheme:

1. Structural and incidental dependence. The substrates that indicate dependency with regard to the initial diffusion matrix are eliminated at the beginning of the algorithm as similarly applied in Scheme 1 (under Structural dependence). Then all unknown missing and reaction rate constants are initialized and the iteration counter g is set to 0. Finally the incidental dependency in the system is solved as explained in Scheme 1 (under Incidental dependence).

2. Block update. The candidate reaction rate Θ^* is generated by sequentially adding d-dimensional deviances to the current rate constant Θ. The calculation of likelihoods in the acceptance probability (Equation 8) is computed by using lower dimensional diffusion matrices if the associated diffusion terms have singularity. In other words the acceptance probability for each d-dimensional reaction rate is computed by taking into account merely linearly independent reaction rates as long as $\beta^{xx}_{i-1} = \beta(Y^{xx}_{i-1}, \Theta)$ of each state is non-singular if we accept candidate rates. By this implementation since the likelihood of the dependent rate given the remaining rates is 1, there is no lost of information after the highlighted exclusion. Moreover when the reaction rate is accepted, i.e. $\Theta = \Theta^*$, including both dependent and independent terms as long as its corresponding α is high, the update can preserve the dependent structure of the system without affecting the convergent feature of the algorithm. This dependent structure can be important in the estimation.

3. Update latent states. The algorithm updates the column of Y sequentially by the Metropolis-within-Gibbs sampling. If the corresponding $\beta^{xx}_{i-1} = \beta(Y^{xx}_{i-1}, \Theta)$ is non-singular, the diffusion matrices of Y_{i-1}, Y_i, and Y^*_i are controlled separately. Then if there is any substrate which destroys the nonsingularity, the underlying substrate(s) is/are excluded only from the calculation of the associated part of likelihoods, that is the corresponding computation of Y_i, Y_{i+1}, or Y^*_i. This calculation enables to preserve dependencies of the states without changing the convergence of the system as similarly implemented in the update of rates in Step 2. If the move is accepted with probability α, Y_i is set to Y^*_i, otherwise, the chain keeps current values at the gth iteration.

4. Iterate. The counter goes from g to $(g + 1)$ and the algorithm is repeated from the control of the incidental dependence in Step 1 until the convergence is satisfied.

5 Description of the simulated data

We simulate the MAPK pathway by using the Gillespie algorithm since this algorithm gives an exact result of the system (Gillespie 1977; Wilkinson 2006). In our simulation we choose 3 gradations of reaction time speeds, namely slow, normal, and fast. Then we assume that the initial hazards $h_j(Y, \Theta)$, $j = 1, \ldots, r$, are constant for each level by arbitrarily assigning $h_j(Y, \Theta) = 50, 100$, and 150 for slow, normal, and fast reactions, respectively. Then for the selected constant hazards, the reaction rate constants c's are calculated with respect to the order of reactions, the given hazards, and the number of molecules which is initialized at 100 for all substrates. For instance, if the reaction is the first-order reaction like $\mathrm{S}_1 \rightarrow \mathrm{S}_2$ and has normal speed, the reaction rate is taken as $c_j = h_j(Y, \Theta)/100 = 100/100 = 1.000$. If the reaction is the second-order reaction having normal speed, then c_j is calculated via $c_j = h_j(Y, \Theta)/100^2 = 100/100^2 = 0.010$. Similarly if the reaction is the second-order but also fast, then we compute c_j by $c_j = h_j(Y, \Theta)/100^2 = 150/100^2 = 0.015$. Considering the underlying distinction in speeds and orders of reactions, the reaction rates are set to 0.500 (for the first-order slow reaction), 0.005 (for the second-order slow reaction), 1.000 (for the first-order normal reaction), 0.010 (for the second-order normal reaction), or 0.015 (for the second-order fast reaction) where necessary. In our description since none of the reaction is of first-order and fast simultaneously, c_j does not equal to 1.500 for any case.

Moreover in simulation we run the algorithm merely including the EGFR degradation. Because in biochemical reactions, apart from the EGFR degradation, the reactions of degradation are much slower than the time periods during which biochemical activation and de-activation processes take place. Therefore ignoring these reactions in the MAPK pathway is realistic. However the effect of the EGF dissociation from its receptor is a direct result of the activation of the MAPK pathway via the internalization into vesicles of this receptor and is very fast with respect to any kind of degradations. The lists of reactions as well as substrates used in this model are given in the Appendix.

Finally in order to get a data matrix Y, the Gillespie algorithm is run until each protein achieves a convergent distribution. According to the plot of every species which shows the changes in activities through time, we set the total time interval t to 20 and observe the steady-state period from $t = 15$.

5.1 Data generation for inference

In inference considering the number of variables, the necessity of augmented states for the accuracy of estimated parameters, and the possible computational cost of the Bayesian inference for such a complex system, we generate a time-course dataset which has many more observed substrates than any real dataset has currently. Because in a real time-course dataset like a western blotting data, the number of observed substrates is less due to the technical limitations for measuring the protein levels (Vyshemirsky et al. 2006). The implementation of the method in a real western-blot data set can be found in Purutçuoğlu and Wit (2008).

In our simulated data we sample 20 time points of the selected proteins by moving 0.05 unit of time from $t = 19.05$ to $t = 20$. The observed substrates are chosen in such a way that most of them can be used in inference after the elimination of the structural dependent ones. We start our MCMC with 35 measured MAPK proteins (chosen as Ras.GDP, Ras.GTP, Raf, Raf.I, Raf.I-Ras.GTP$_m$, Raf.A$_m$, Raf.I-RKIP, MEK, MEK$_F$, MEK$_S$, MEK.p2, MEK-RKIP, ERK, ERK.p2, ERK.p2-TF.p2, ERK.p2-RSK.A, ERK.p2-RSK.A-TF.p2, Grb2, Shc, Shc$_m$, SOS, Grb2-SOS, Grb2-SOS$_m$, c-Fos, c-Fos.RNA, MKP, MKP.RNA, EGFR, TF, PAK, PP5, RKIP, RKIP.p, PKC, and RSK) among 51 species that represent the whole pathway. When the inference begins, 7 proteins (ERK.p2-RSK.A-TF.p2, TF, PAK, PP5, RKIP.p, PKC, and RSK) are discarded within this initial observed set due to their structural dependencies (see Section 4 for details) on other proteins. Therefore the estimation is, indeed, based on 28 observed and 6 (Raf.I$_m$, Raf.A-Ras.GTP$_m$, MEK$_F$-RKIP, MEK$_S$-RKIP, ERK.p1, c-Fos.p) unobserved substrates.

Finally in our analysis we assume that the observed values do not have any measurement error, such that the noise of the data merely originates from the stochasticity of the protein interactions. Although this assumption can be regarded as strong, we use it for simplicity. The additional error term for each observation would increase the computational time due to the increase of the number of estimated parameters in such a complex system where the current estimation is already computationally very demanding.

6 Application to the simulated data

The estimated parameters via two inference algorithms are presented in Table 1 and Table 2. The results from two schemes mostly indicate the success of Scheme 2 with respect to Scheme 1 in terms of the precision. This shows that often there is dependence within the states and within the reaction rates in the sense that the dependency in the algorithm significantly affects the outputs. However typically Scheme 1 has higher acceptance ratios than Scheme 2 owns. This implies the difficulty of proposing the candidate values under the dependent structure. Moreover in terms of the computational cost, Scheme 1 is more efficient (Table 3). On the other hand when comparing the estimates with true values, we see that some of the estimates are precise in both schemes, whereas some of them have a very large bias. Figure 3 and Figure 4 display several examples from probability distributions of rate constants of two schemes with different acceptance ratios. The average errors of each estimate from both schemes, on the other side, are represented in Figure 5. In these plots the errors are calculated as Average error = |Estimated value − True value|/True value where |.| indicates the absolute value of the given number. In the assessment we consider that the average errors less than 60% can be seen as precise estimates for this number of iterations. The study of Golightly and Wilkinson (2005) shows that the performance of algorithms is significantly dependent on the number of observations, iterations, and augmented states between each pair of observed time points. They find that as the underlying values increase, the estimates improve considerably. Therefore with respect to their analysis, we already work with a small dataset and a smaller number of augmented states as well as a smaller number of iterations. Moreover our system of interest has more complex structure than the network which they use in their evaluation. Consequently from the beginning of the application, it is expected that the accuracies of estimates may not be satisfied for all parameters and is considered that the imprecise estimates are mostly caused by our preferences for the underlying variables, rather than the limitation of the method. We see that even when the number of observed time points is slightly increased, the accuracy of estimates is higher than that in Table 1 and Table 2. The estimated values via Scheme 1 for 50 time points under the same conditions (i.e. $m = 3$ thereby $T = 197$, 35 observed substrates with the common initial rate constants and number of molecules) are displayed in Table 6 and Figure 6 in the Appendix. From the comparison of results of Figure 5(a) and Figure 6, it is also observed that the gain from the accuracy is particularly seen for the estimates whose average errors lie between 0.32 and 0.61 although the estimates whose errors are greater than 0.61 do not change very much. The reason can be explained as the direct influence of such an improvement (i.e. the increase of the observed measurements) on the model parameters whose convergence rates are faster than those of other components. From our analysis we find that even though all the estimates indicate good mixing properties at the end of burn-in, i.e. the acceptance ratios are between 5% and 60%, some of the estimates reach convergence very fast, whereas some of them have very slow convergence. Hence it is concluded that in order to get higher convergence rates from the imprecise results, we should not use only more observations but also utilize higher number of iterations and higher number of augmented states between each observed pair despite of the high computational cost in estimation of such a complex system. Therefore we believe that our finding can be evaluated as one of the worst scenario in inference of the realistic complexity and the performance of algorithms can be better displayed at least, for instance, by raising the number of iterations like 1,000,000 or 10,000,000 MCMC runs as Golightly and Wilkinson (2005) applied in their study. However if we prefer such a high number of iterations in our inference for improving accuracies, we suggest an efficient programme language like `C`, in place of `R`, to reduce the computational cost efficiently. In this study where we execute our `R` codes on a high power computer, the calculation takes at least seven days (Table 3).

In our assessment, apart from the highlighted reasons of the inaccuracy of estimates, we investigate other sources of imprecisions, analysing each reaction and its corresponding substrates one by one. We find that, particularly, when more than one substrate is missing in a reaction due to the exclusion of structurally dependent substrates (Section 4), the relevant estimated rate constant typically has large error. For instance, in Reaction 9, 12, 18, 26, 31, 51, 54, 59, and 65 from the list of reactions in the Appendix, we have more than one missing substrate according to the classification in Table 5. Accordingly we get a low accuracy of estimates of related rates as stated in Table 1 and Table 2. Moreover since we allow unobserved substrates in our model, the reactions whose components consist of both linearly dependent and unobserved substrates such as Reaction 16, 42, and 43 from the same list (Appendix) indicate inaccurate estimates.

According to the individual evaluation of each parameter, seeing that the precision of reaction rates is highly dependent on the number of missing species in relevant reactions, we consider comparing the ratios of rate constants. The reason is that the ratios of rates could be more invariant to missing substrates if the reactions share the same missing terms. In order to decide on the pairs of rate constants for comparison, we consider the following scenarios whose associated reactions have common species:

(a) $\mathrm{A} \xrightarrow{c_1} \mathrm{B}$

(b) $\mathrm{B} \xrightarrow{c_2} \mathrm{C} + \mathrm{D}$

(c) $\mathrm{E} \xrightarrow{c_3} \mathrm{F}$

(d) $\mathrm{G} \xrightarrow{c_4} \mathrm{F}$

(e) $\mathrm{H} + \mathrm{J} \xrightarrow{c_5} \mathrm{K}$

(f) $\mathrm{H} + \mathrm{M} \xrightarrow{c_6} \mathrm{K}.$

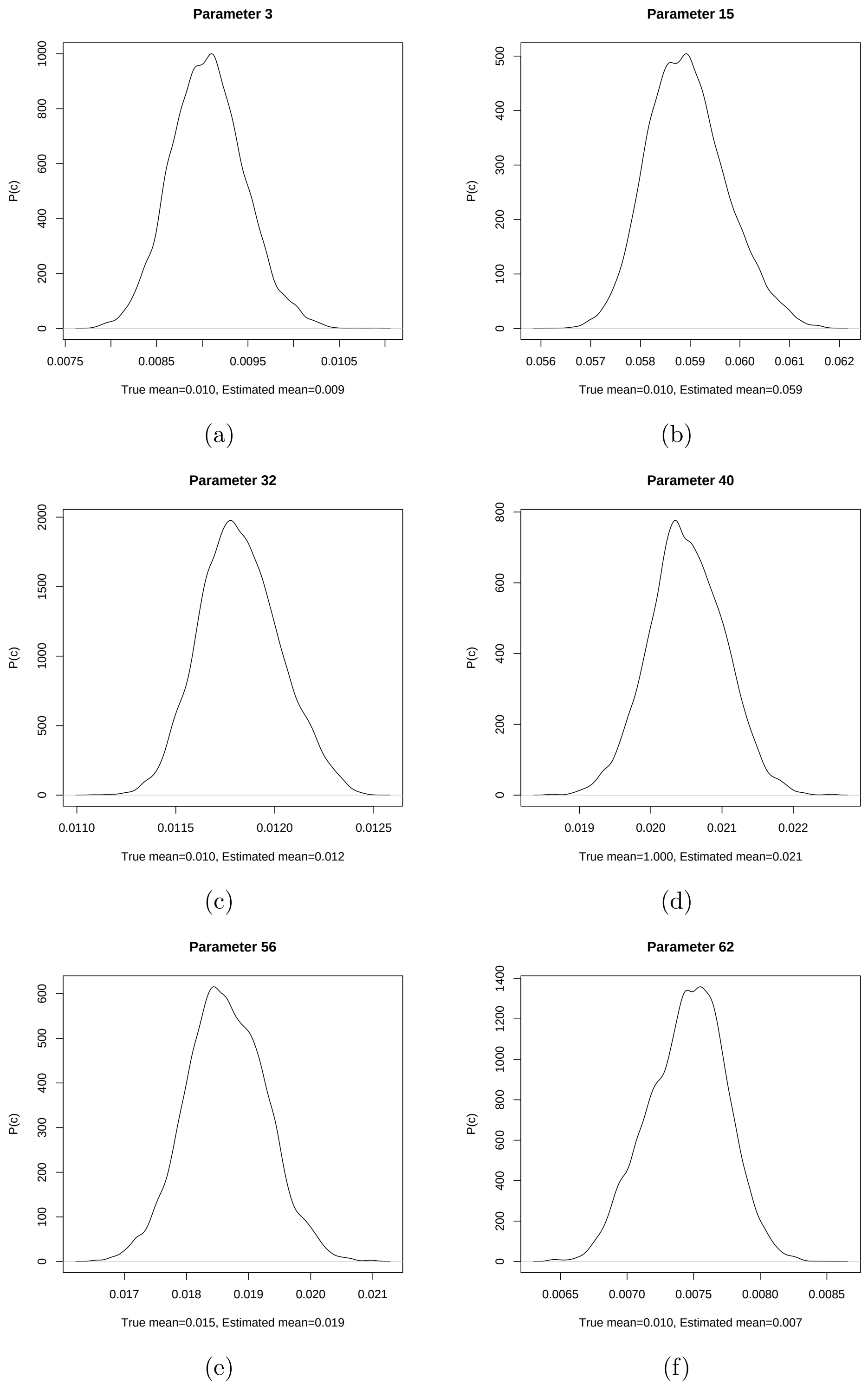
Among those reactions we expect a constant ratio between c_1 and c_2 since there is a positive correlation between Reaction (a) and (b) as both own the species B. On the other hand since a negative correlation exists between Reaction (c) and (d) because

Table 1: Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants of the MAPK/ERK pathway from the simulated data. The estimates are based on algorithms in Scheme 1 and Scheme 2 with $100,000$ MCMC runs in which the first $85,000$ runs are taken as burn-in.

		Scheme 1			Scheme 2		
Reaction	True rate	μ	σ	p	μ	σ	p
c_1	0.010	0.007	0.000	0.545	0.009	0.000	0.526
c_2	0.010	0.055	0.001	0.541	0.060	0.001	0.523
c_3	0.010	0.009	0.000	0.545	0.023	0.001	0.521
c_4	0.010	0.022	0.001	0.541	0.020	0.001	0.525
c_5	1.000	2.744	0.022	0.496	4.239	0.077	0.517
c_6	1.000	2.252	0.093	0.583	1.221	0.011	0.342
c_7	1.000	1.539	0.007	0.500	1.058	0.010	0.354
c_8	1.000	0.680	0.019	0.590	0.979	0.010	0.390
c_9	0.010	0.198	0.004	0.577	0.202	0.007	0.388
c_{10}	0.010	0.000	0.000	0.596	0.000	0.000	0.395
c_{11}	1.000	1.469	0.010	0.562	1.035	0.006	0.425
c_{12}	0.015	0.653	0.022	0.632	0.116	0.010	0.527
c_{13}	0.010	3.933	0.079	0.614	0.949	0.032	0.532
c_{14}	0.010	0.058	0.001	0.629	0.054	0.001	0.528
c_{15}	0.010	0.059	0.001	0.627	0.058	0.001	0.529
c_{16}	0.010	3.080	0.126	0.814	1.149	0.027	0.377
c_{17}	1.000	0.272	0.006	0.803	0.912	0.006	0.388
c_{18}	0.010	2.871	0.093	0.807	1.161	0.022	0.371
c_{19}	1.000	0.264	0.007	0.812	0.864	0.016	0.394
c_{20}	1.000	1.450	0.025	0.737	0.966	0.004	0.389
c_{21}	0.010	0.001	0.001	0.568	0.014	0.010	0.141
c_{22}	0.010	0.702	0.021	0.564	5.155	0.066	0.140
c_{23}	0.015	0.827	0.084	0.567	0.228	0.016	0.140
c_{24}	0.010	0.008	0.000	0.565	0.008	0.001	0.141
c_{25}	0.010	0.190	0.005	0.544	0.265	0.009	0.140
c_{26}	0.010	4.132	0.082	0.601	5.487	0.070	0.267
c_{27}	0.010	0.434	0.004	0.598	0.438	0.015	0.268
c_{28}	0.010	0.058	0.003	0.615	0.050	0.008	0.269
c_{29}	0.010	0.046	0.002	0.614	0.041	0.009	0.269
c_{30}	0.010	0.013	0.000	0.592	0.013	0.001	0.268
c_{31}	0.010	0.027	0.008	0.609	0.114	0.039	0.278
c_{32}	0.010	0.012	0.000	0.575	0.014	0.002	0.278
c_{33}	1.000	4.706	0.053	0.583	5.214	0.083	0.276

Table 2: Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants of the MAPK/ERK pathway from the simulated data. The estimates are based on algorithms in Scheme 1 and Scheme 2 with $100,000$ MCMC runs in which the first $85,000$ runs are taken as burn-in.

		Scheme 1			Scheme 2		
Reaction	True rate	μ	σ	p	μ	σ	p
c_{34}	0.010	1.101	0.007	0.503	0.684	0.051	0.278
c_{35}	0.010	1.695	0.041	0.577	2.567	0.023	0.271
c_{36}	0.010	0.207	0.006	0.516	0.291	0.040	0.323
c_{37}	0.010	0.372	0.014	0.518	0.351	0.035	0.323
c_{38}	1.000	0.981	0.012	0.520	0.774	0.055	0.324
c_{39}	1.000	0.946	0.016	0.519	1.287	0.028	0.316
c_{40}	1.000	0.021	0.001	0.515	0.025	0.006	0.324
c_{41}	1.000	0.002	0.000	0.174	0.003	0.001	0.211
c_{42}	0.010	0.028	0.002	0.174	0.056	0.021	0.211
c_{43}	0.010	0.019	0.003	0.174	0.037	0.011	0.211
c_{44}	1.000	0.528	0.002	0.172	0.492	0.018	0.210
c_{45}	0.015	1.588	0.021	0.168	1.726	0.022	0.205
c_{46}	0.010	1.621	0.016	0.766	1.845	0.052	0.305
c_{47}	0.010	0.176	0.002	0.787	0.142	0.008	0.307
c_{48}	0.010	5.426	0.116	0.829	3.396	0.052	0.307
c_{49}	0.010	3.321	0.080	0.827	1.702	0.141	0.308
c_{50}	1.000	1.357	0.028	0.798	1.248	0.038	0.305
c_{51}	0.010	3.972	0.156	0.685	2.732	0.055	0.282
c_{52}	1.000	0.187	0.002	0.651	0.233	0.012	0.285
c_{53}	1.000	0.165	0.003	0.653	0.174	0.004	0.282
c_{54}	0.010	0.171	0.004	0.668	0.216	0.004	0.282
c_{55}	0.010	1.932	0.060	0.665	1.456	0.024	0.275
c_{56}	0.015	0.019	0.001	0.388	0.016	0.008	0.249
c_{57}	0.010	1.453	0.028	0.370	3.475	0.189	0.249
c_{58}	0.010	1.723	0.014	0.357	1.777	0.114	0.248
c_{59}	0.010	0.956	0.010	0.387	4.372	0.102	0.248
c_{60}	0.010	0.000	0.000	0.392	0.016	0.013	0.249
c_{61}	0.010	0.265	0.010	0.544	0.948	0.146	0.274
c_{62}	0.010	0.007	0.000	0.551	0.021	0.005	0.275
c_{63}	0.010	1.338	0.007	0.467	1.271	0.084	0.273
c_{64}	0.010	0.758	0.017	0.547	0.344	0.215	0.275
c_{65}	0.010	1.157	0.012	0.485	1.839	0.104	0.273
c_{66}	1.000	5.338	0.185	0.714	17.706	0.153	0.280

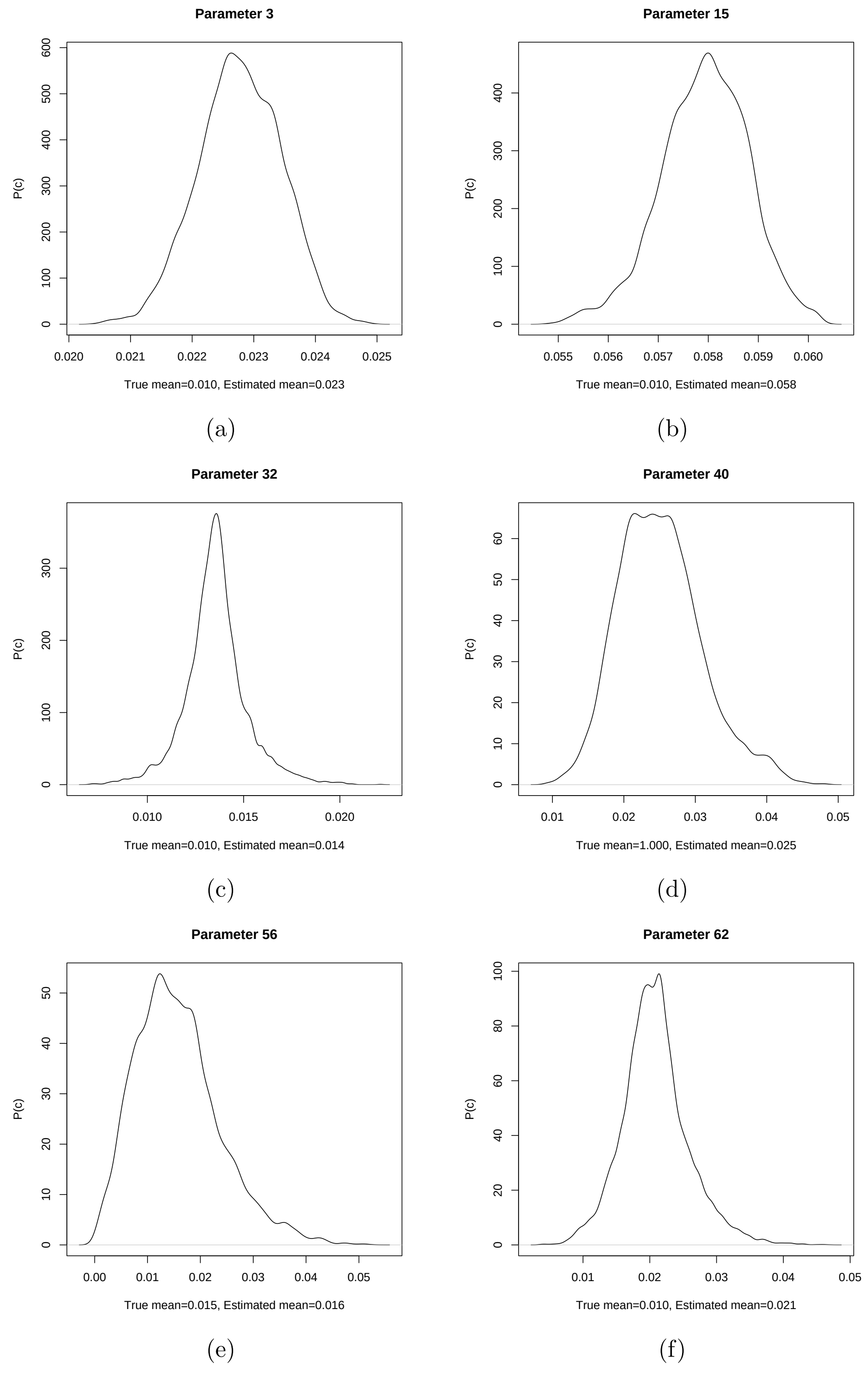


(a) (b)

(c) (d)

(e) (f)

Figure 3: Probability distribution of the reaction rate (a) 3, (b) 15, (c) 32, (d) 40, (e) 56, and (f) 62, respectively, by using Scheme 1 after 100,000 MCMC runs in which the first 85,000 runs are burn-in.



(a) (b)

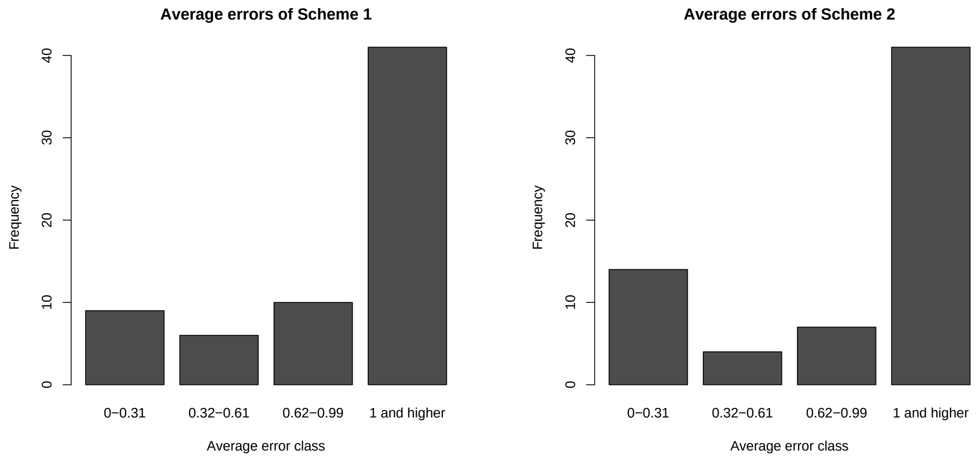
(c) (d)

(e) (f)

Figure 4: Probability distribution of the reaction rate (a) 3, (b) 15, (c) 32, (d) 40, (e) 56, and (f) 62, respectively, by using Scheme 2 after 100,000 MCMC runs in which the first 85,000 runs are burn-in.

Table 3: Total CPU (Central processing unit) time and total real computational time of Scheme 1 (with 20 and 50 observed time points) and Scheme 2, respectively, in R and with a high power computer (3.00 GHz Dual Core Xeon Process - Single Trade Application) for estimating rate constants via 100,000 MCMC runs.

	Total CPU	Total computational time
Scheme 1 with 20 observed time points	135.06	175 hr 8 min 1 sec
Scheme 1 with 50 observed time points	360.63	230 hr 47 min 29 sec
Scheme 2 with 20 observed time points	183.67	231 hr 1 min 36 sec



(a) (b)

Figure 5: Frequencies of average errors for the estimates presented in Table 1 and 2. Figure (a) shows the results obtained by Scheme 1 and Figure (b) indicates the results found by Scheme 2. The estimates are based on 20 observed time points and 100,000 MCMC runs in which the first 85,000 runs are burn-in.

Table 4: Posterior means (μ) and standard deviations (σ), of ratios of estimated reaction rate constants of the MAPK/ERK pathway from the simulated data. The estimates are based on algorithms in Scheme 1 and Scheme 2 with $100,000$ MCMC runs in which the first $85,000$ runs are taken as burn-in.

		Scheme 1		Scheme 2	
Ratio of reaction rate	True ratio	μ	σ	μ	σ
c_4/c_5	0.010	0.008	0.001	0.005	0.000
c_3/c_7	0.010	0.006	0.000	0.022	0.001
c_1/c_8	0.010	0.010	0.000	0.009	0.000
c_{14}/c_{15}	1.000	0.977	0.026	0.927	0.018
c_{30}/c_{38}	0.010	0.013	0.000	0.017	0.002
c_{52}/c_{53}	1.000	1.132	0.015	1.342	0.067
c_{57}/c_{58}	1.000	0.844	0.011	1.960	0.108

of the species F, similarly between Reaction (e) and (f) due to the species H and K, we suggest that the ratios of associated rates (i.e. the ratios of c_3/c_4 and c_5/c_6) may not be constant. Therefore in our analysis we couple the parameters whose common terms have the form as given in Reaction (a) and (b) without additionally investigating whether our selected ratios have any biological meaning. Table 4 shows examples of those estimated ratios and their corresponding true values. The results show that the selected ratios are precise.

7 Conclusion and discussion

We have illustrated the implementation of MCMC algorithms and data augmentation schemes in the estimation of reaction rates of the MAPK/ERK pathway, which was described as a set of quasi reactions by integrating several sources from the biological literature. Due to the size and complexity of the MAPK/ERK network, any diffusion approximation of its dynamic progression is riddled with singularities of the diffusion matrix, i.e. effective linear dependencies in the states. Therefore, we have introduced two MCMC updating regimes that deal with these singularities at different stages of the update. The first plan simply rejects states or rates that lead to singular diffusions. The second one, tolerates reaction rates and states with singular diffusions by temporarily reducing the dimension of the state space. In simulation studies, it has been observed that the second plan is more accurate than the first updating scheme.

The Euler method with data augmentation for approximating diffusions is a promising inference tool. But the large number of missing values and high correlations between substrates can cause biased estimates, particularly obvious when the amount of observations of the system are realistically small. Performance of the estimation improves, when the number of observations increases. Apart from an increase of the sample size, estimates can also be improved by updating the missing data in blocks of random size, rather than single block at a time (Golightly and Wilkinson 2006a,b), which enables better mixing in the end.

The existence of linear dependencies in the system is always a serious problem in estimation. Therefore we suggest that when estimating model parameters of a complex stochastic dynamic system, like our pathway, the substrates that are eliminated because of their structural dependencies, can be implicitly updated within the MCMC algorithm, rather than completely excluded from the dataset. Every dependent substrate can be simulated as a linear combination of other substrates. In that way, linearly dependent substrates can be included in the computation of acceptance probabilities of reaction rates and states when calculating the ratio of both likelihoods and transition kernels. The implementation of this type of the inference, i.e. the inference under structural dependency, is the topic of ongoing research.

8 Appendix

8.1 Description of the MAPK/ERK pathway and related estimates with a larger dataset

We use the following list of reactions for estimating reaction rates of the MAPK/ERK pathway. The estimated rates and associated statistics are summarized in Table 1 and Table 2. The list of substrates, on the other hand, is given in Table 5.

In Table 6 we list the estimated model parameters and associated statistics calculated by Scheme 1 from a dataset which has 50 time points.

1. Grb2 + SOS $\longrightarrow$ Grb2-SOS
2. EGFR + Shc $\longrightarrow$ EGFR + Shc_m
3. EGFR + Grb2-SOS $\longrightarrow$ EGFR + Grb2-SOS_m
4. Shc_m + Grb2-SOS_m $\longrightarrow$ Shc-Grb2-SOS_m
5. Shc-Grb2-SOS_m $\longrightarrow$ Shc_m + Grb2-SOS_m
6. Shc_m $\longrightarrow$ Shc
7. Grb2-SOS_m $\longrightarrow$ Grb2-SOS
8. Grb2-SOS $\longrightarrow$ Grb2 + SOS
9. Shc-Grb2-SOS_m + Ras.GDP $\longrightarrow$ Shc-Grb2-SOS_m + Ras.GTP
10. Grb2-SOS_m + Ras.GDP $\longrightarrow$ Grb2-SOS_m + Ras.GTP

11. Ras.GTP $\longrightarrow$ Ras.GDP

12. GAP + Ras.GTP $\longrightarrow$ GAP + Ras.GDP

13. Raf + PP2A $\longrightarrow$ Raf.I + PP2A

14. Raf.I + Ras.GTP $\longrightarrow$ Raf.I$_m$ + Ras.GTP

15. Raf.I$_m$ + Ras.GTP $\longrightarrow$ Raf.I-Ras.GTP$_m$

16. Raf.I-Ras.GTP$_m$ + PAK $\longrightarrow$ Raf.A-Ras.GTP$_m$ + PAK

17. Raf.A-Ras.GTP$_m$ $\longrightarrow$ Raf.A$_m$ + Ras.GTP

18. PP5 + Raf.A$_m$ $\longrightarrow$ PP5 + Raf.I$_m$

19. Raf.I$_m$ $\longrightarrow$ Raf

20. Raf.I-Ras.GTP$_m$ $\longrightarrow$ Raf.I$_m$ + Ras.GTP

21. Raf.A$_m$ + MEK $\longrightarrow$ Raf.A$_m$ + MEK.p2

22. PAK + MEK $\longrightarrow$ PAK + MEK$_F$

23. MEK$_F$ + Raf.A$_m$ $\longrightarrow$ MEK.p2 + Raf.A$_m$

24. Raf.I + RKIP $\longrightarrow$ Raf.I-RKIP

25. Raf.I-RKIP + Ras.GTP $\longrightarrow$ Raf.I-RKIP$_m$ + Ras.GTP

26. Raf.I-RKIP$_m$ + Ras.GTP $\longrightarrow$ Raf.I-RKIP-Ras.GTP$_m$

27. MEK + RKIP $\longrightarrow$ MEK-RKIP

28. MEK$_F$ + RKIP $\longrightarrow$ MEK$_F$-RKIP

29. MEK$_S$ + RKIP $\longrightarrow$ MEK$_S$-RKIP

30. MEK.p2 + RKIP $\longrightarrow$ MEK.p2-RKIP

31. PKC + Raf.I-RKIP $\longrightarrow$ PKC + Raf.I + RKIP.p

32. ERK.p2 + Raf.I-RKIP $\longrightarrow$ ERK.p2 + Raf.I + RKIP.p

33. Raf.I-RKIP-Ras.GTP$_m$ $\longrightarrow$ Raf.I-RKIP$_m$ + Ras.GTP

34. Raf.I-RKIP$_m$ $\longrightarrow$ Raf.I-RKIP

35. MEK-RKIP $\longrightarrow$ MEK + RKIP

36. MEK$_F$-RKIP $\longrightarrow$ MEK$_F$ + RKIP

37. MEK$_S$-RKIP $\longrightarrow$ MEK$_S$ + RKIP

38. MEK.p2-RKIP $\longrightarrow$ MEK.p2 + RKIP

39. RKIP.p ⟶ RKIP

40. MEK.p2 + ERK ⟶ MEK.p2 + ERK.p1

41. MEK.p2 + ERK.p1 ⟶ MEK.p2 + ERK.p2

42. MEK.p2-RKIP + ERK ⟶ MEK.p2-RKIP + ERK.p1

43. MEK.p2-RKIP + ERK.p1 ⟶ MEK.p2-RKIP + ERK.p2

44. ERK.p2 + MEK ⟶ ERK.p2 + MEK$_S$

45. MEK$_S$ + Raf.A$_m$ ⟶ MEK.p2 + Raf.A$_m$

46. ERK.p2 + Shc-Grb2-SOS$_m$ ⟶ ERK.p2 + Shc-Grb2$_m$ + SOS

47. ERK.p2 + Grb2-SOS$_m$ ⟶ ERK.p2 + Grb2$_m$ + SOS

48. Shc-Grb2$_m$ ⟶ Shc + Grb2

49. Grb2$_m$ ⟶ Grb2

50. ERK.p2 + TF ⟶ ERK.p2-TF.p2

51. ERK.p2-TF.p2 + c-Fos.DNA ⟶ ERK.p2-TF.p2 + c-Fos.DNA + c-Fos.RNA

52. c-Fos.RNA ⟶ c-Fos

53. ERK.p2 + c-Fos ⟶ ERK.p2 + c-Fos.p

54. ERK.p2-TF.p2 + MKP.DNA ⟶ ERK.p2-TF.p2 + MKP.DNA + MKP.RNA

55. MKP.DNA ⟶ MKP

56. MKP + ERK.p2 ⟶ MKP + ERK

57. ERK.p2 + RSK ⟶ ERK.p2-RSK.A

58. ERK.p2-RSK.A + TF ⟶ ERK.p2-RSK.A-TF.p2

59. ERK.p2-RSK.A-TF.p2 + c-Fos.DNA ⟶ ERK.p2-RSK.A-TF.p2 + c-Fos.DNA
+ c-Fos.RNA

60. ERK.p2-RSK.A + c-Fos ⟶ ERK.p2-RSK.A + c-Fos.p

61. ERK.p2-RSK.A-TF.p2 + MKP.DNA ⟶ ERK.p2-RSK.A-TF.p2 + MKP.DNA
+ MKP.RNA

62. MKP + ERK.p2-RSK.A ⟶ MKP + ERK + RSK

63. ERK.p2-TF.p2 $\longrightarrow$ ERK.p2 + TF

64. ERK.p2-RSK.A $\longrightarrow$ ERK.p2 + RSK

65. ERK.p2-RSK.A-TF.p2 $\longrightarrow$ ERK.p2-RSK.A + TF

66. EGFR $\longrightarrow \emptyset$

8.2 Reaction of degradations of the MAPK/ERK pathway

Apart from the degradation of EGFR which is denoted as the 66th reaction in Section 8.1, we define the following list of degradations which may execute after dissociations of proteins for the MAPK/ERK pathway.

1. Grb2 $\longrightarrow \emptyset$

2. SOS $\longrightarrow \emptyset$

3. Shc $\longrightarrow \emptyset$

4. Ras.GDP $\longrightarrow \emptyset$

5. GAP $\longrightarrow \emptyset$

6. Raf $\longrightarrow \emptyset$

7. Raf.I $\longrightarrow \emptyset$

8. Raf.A_m $\longrightarrow \emptyset$

9. PP2A $\longrightarrow \emptyset$

10. PAK $\longrightarrow \emptyset$

11. PP5 $\longrightarrow \emptyset$

12. MEK $\longrightarrow \emptyset$

13. MEK_F $\longrightarrow \emptyset$

14. MEK_S $\longrightarrow \emptyset$

15. MEK.p2 $\longrightarrow \emptyset$

16. RKIP $\longrightarrow \emptyset$

17. Raf.I-RKIP $\longrightarrow \emptyset$

18. PKC $\longrightarrow \emptyset$

Table 5: List of proteins used in inference of the MAPK/ERK pathway. The proteins written in bold type are taken as measured proteins in the computation.

Independent	**Ras.GDP**, **Ras.GTP**, **Raf**, **Raf.I**, Raf.I$_m$, **Raf.I-Ras.GTP**$_m$, **Raf.A**$_m$, Raf.A-Ras.GTP$_m$, **Raf.I-RKIP**, **RKIP**, **MEK**, **MEK**$_F$, **MEK**$_S$, **MEK.p2**, **MEK-RKIP**, MEK$_F$-RKIP, MEK$_S$-RKIP, **EGFR**, **ERK**, ERK.p1, **ERK.p2**, **ERK.p2-TF.p2**, **ERK.p2-RSK.A**, **Grb2**, **Shc**, **Shc**$_m$, **SOS**, **Grb2-SOS**, **Grb2-SOS**$_m$, **c-Fos**, **c-Fos.RNA**, c-Fos.p, **MKP**, **MKP.RNA**.
Dependent	Raf.I-RKIP$_m$, Raf.I-RKIP-Ras.GTP$_m$, MEK.p2-RKIP, MKP.DNA, **TF**, **PAK**, **ERK.p2-RSK.A-TF.p2**, Shc-Grb2-SOS$_m$, Grb2$_m$, Shc-Grb2$_m$, GAP, **PKC**, c-Fos.DNA, PP2A, **PP5**, **RKIP.p**, **RSK**.

19. ERK $\longrightarrow \emptyset$

20. ERK.p1 $\longrightarrow \emptyset$

21. ERK.p2 $\longrightarrow \emptyset$

22. TF $\longrightarrow \emptyset$

23. c-Fos.RNA $\longrightarrow \emptyset$

24. c-Fos $\longrightarrow \emptyset$

25. MKP.RNA $\longrightarrow \emptyset$

26. MKP $\longrightarrow \emptyset$

27. c-Fos.p $\longrightarrow \emptyset$

28. RSK $\longrightarrow \emptyset$

8.3 Description of protein states used in the MAPK/ERK pathway

We use the following substrates in the description of the MAPK/ERK pathway with the degradation of the EGF receptor.

1. Ras protein states
   (a) Ras.GDP: The inactive Ras protein near the cell membrane
   (b) Ras.GTP: The active Ras near the cell membrane
2. Raf protein states
   (a) Raf: The inactive and non-phosphorylated Raf protein in the cytosol

Table 6: Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants of the MAPK/ERK pathway from the simulated data which have 50 time points. The data are generated from the Gillespie simulation of the system and are gathered by moving 0.05 unit of time between $t = 17.55$ and $t = 20$ under the same condition given in Section 5 and 5.1. The estimates are based on algorithms in Scheme 1 with $100,000$ MCMC runs in which the first $85,000$ runs are taken as burn-in.

Reaction	True rate	μ	σ	p	Reaction	True rate	μ	σ	p
c_1	0.010	0.009	0.000	0.536	c_{34}	0.010	1.155	0.015	0.458
c_2	0.010	0.051	0.000	0.531	c_{35}	0.010	1.846	0.038	0.497
c_3	0.010	0.038	0.001	0.530	c_{36}	0.010	0.285	0.005	0.497
c_4	0.010	0.031	0.001	0.535	c_{37}	0.010	0.448	0.009	0.501
c_5	1.000	3.876	0.085	0.529	c_{38}	1.000	0.880	0.008	0.499
c_6	1.000	1.178	0.005	0.133	c_{39}	1.000	0.998	0.010	0.488
c_7	1.000	1.144	0.003	0.125	c_{40}	1.000	0.020	0.001	0.498
c_8	1.000	0.968	0.002	0.144	c_{41}	1.000	0.002	0.000	0.159
c_9	0.010	0.268	0.001	0.142	c_{42}	0.010	0.001	0.001	0.159
c_{10}	0.010	0.000	0.000	0.146	c_{43}	0.010	0.006	0.002	0.159
c_{11}	1.000	1.815	0.038	0.563	c_{44}	1.000	0.471	0.002	0.156
c_{12}	0.015	0.727	0.022	0.589	c_{45}	0.015	0.979	0.012	0.158
c_{13}	0.010	3.554	0.071	0.574	c_{46}	0.010	1.033	0.018	0.705
c_{14}	0.010	0.037	0.001	0.585	c_{47}	0.010	0.074	0.001	0.773
c_{15}	0.010	0.054	0.001	0.583	c_{48}	0.010	7.443	0.171	0.773
c_{16}	0.010	2.603	0.092	0.774	c_{49}	0.010	3.641	0.148	0.774
c_{17}	1.000	0.265	0.003	0.760	c_{50}	1.000	1.118	0.013	0.659
c_{18}	0.010	2.467	0.086	0.773	c_{51}	0.010	4.574	0.182	0.645
c_{19}	1.000	0.276	0.003	0.747	c_{52}	1.000	0.187	0.004	0.624
c_{20}	1.000	1.240	0.022	0.698	c_{53}	1.000	0.194	0.006	0.627
c_{21}	0.010	0.002	0.001	0.440	c_{54}	0.010	0.209	0.004	0.624
c_{22}	0.010	1.026	0.018	0.402	c_{55}	0.010	2.713	0.103	0.637
c_{23}	0.015	0.632	0.019	0.434	c_{56}	0.015	0.018	0.001	0.205
c_{24}	0.010	0.004	0.000	0.432	c_{57}	0.010	1.563	0.026	0.199
c_{25}	0.010	0.193	0.008	0.428	c_{58}	0.010	1.958	0.008	0.188
c_{26}	0.010	3.143	0.083	0.465	c_{59}	0.010	0.315	0.005	0.203
c_{27}	0.010	0.452	0.002	0.447	c_{60}	0.010	0.000	0.000	0.205
c_{28}	0.010	0.065	0.001	0.472	c_{61}	0.010	0.223	0.017	0.539
c_{29}	0.010	0.106	0.003	0.457	c_{62}	0.010	0.009	0.000	0.540
c_{30}	0.010	0.009	0.000	0.470	c_{63}	0.010	1.118	0.005	0.412
c_{31}	0.010	0.004	0.003	0.522	c_{64}	0.010	0.826	0.015	0.538
c_{32}	0.010	0.009	0.000	0.518	c_{65}	0.010	1.196	0.005	0.425
c_{33}	1.000	3.491	0.058	0.504	c_{66}	1.000	5.860	0.233	0.673

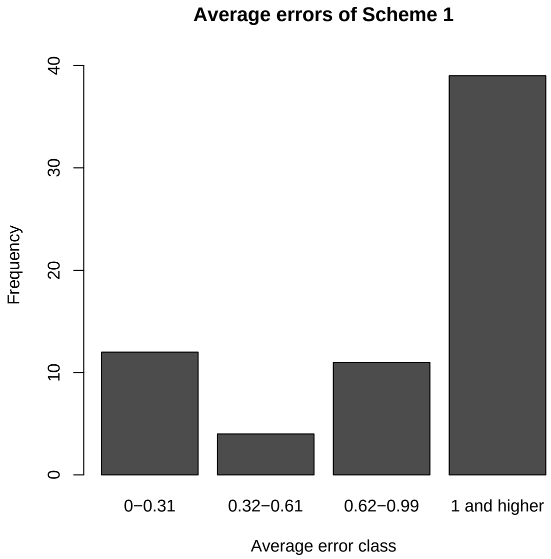


Figure 6: Frequencies of average errors for the estimates presented in Table 6. Figure shows the results calculated by Scheme 1. The estimates are based on 50 observed time points and 100,000 MCMC runs in which the first 85,000 runs are burn-in.

(b) Raf.A$_m$: The active Raf phosphorylated on the S338 and the S471 binding sites near the cell membrane

(c) Raf.A-Ras.GTP$_m$: The complex of the active Raf and Ras.GTP near the cell membrane

(d) Raf.I: The inactive Raf phosphorylated on the S259 binding site in the cytosol

(e) Raf.I$_m$: The inactive Raf phosphorylated on the S259 binding site and recruited from the cytosol to the cell membrane by Ras.GTP

(f) Raf.I-Ras.GTP$_m$: The complex of the inactive Raf and Ras.GTP near the cell membrane

(g) Raf.I-RKIP: The complex of the inactive Raf and RKIP, whose binding site is S338, in the cytosol

(h) Raf.I-RKIP$_m$: The complex of the inactive Raf and RKIP which is recruited to the membrane by Ras.GTP

(i) Raf.I-RKIP-Ras.GTP$_m$: The complex of the inactive Raf, RKIP, and Ras.GTP near the cell membrane

3. MEK protein states

(a) MEK: The inactive and non-phosphorylated MEK protein in the cytosol

(b) MEK$_F$: The inactive MEK in the cytosol which is mono phosphorylated by the activator PAK on the S298 binding site

(c) MEK$_S$: The inactive MEK in the cytosol which is mono phosphorylated by the active ERK on the T292 binding site

(d) MEK.p2: The double-phosphorylated MEK (active MEK) on the S218 and S222 binding sites in the cytosol

(e) MEK-RKIP: The complex of MEK and RKIP in the cytosol

(f) MEK$_F$-RKIP: The complex of MEK$_F$ and RKIP in the cytosol

(g) MEK$_S$-RKIP: The complex of MEK$_S$ and RKIP in the cytosol

(h) MEK.p2-RKIP: The complex of the active MEK.p2 and RKIP in the cytosol

4. ERK protein states

(a) ERK: The inactive and non-phosphorylated ERK protein in the cytosol

(b) ERK.p1: The inactive, mono phosphorylated ERK in the cytosol

(c) ERK.p2: The double phosphorylated ERK (active ERK) in the cytosol

(d) ERK.p2-RSK.A: The complex of the active ERK and active RSK, which is activated by ERK, in the nucleus

(e) ERK.p2-RSK.A-TF.p2: The complex of the active ERK, active RSK, and double phosphorylated transcription factor in the nucleus

(f) ERK.p2-TF.p2: The complex of the active ERK and a transcription factor (like Elk or SAP proteins), which is double phosphorylated by the active ERK, in the nucleus

5. Grb2, Shc, and SOS protein states
   - (a) Grb2: A protein in the cytosol
   - (b) Grb2_m: Grb2 near the cell membrane after dissociation of SOS by active ERK
   - (c) Grb2-SOS: The complex of Grb2 and SOS in the cytosol
   - (d) Grb2-SOS$_m$: The complex of Grb2 and SOS near the cell membrane, where it is able to activate Ras
   - (e) Shc: A protein in the cytosol
   - (f) Shc$_m$: Shc near the cell membrane after the activation of EGFR
   - (g) Shc-Grb2_m: The complex of Shc and Grb2 near the cell membrane after the dissociation of SOS by the active ERK
   - (h) Shc-Grb2-SOS$_m$: The complex of Shc, Grb2, and SOS near the cell membrane, where it is able to activate Ras
   - (i) SOS: A protein, which is an exchange factor, in the cytosol

6. c-Fos and MKP protein states
   - (a) c-Fos: A protein in the nucleus
   - (b) c-Fos.DNA: The gene sequence of c-Fos
   - (c) c-Fos.p: c-Fos phosphorylated by ERK
   - (d) c-Fos.RNA: The transcription of c-Fos gene into mRNA
   - (e) MKP: A protein in the cytosol
   - (f) MKP.DNA: The gene sequence of MKP
   - (g) MKP.RNA: The transcription of MKP gene into mRNA

7. Other proteins
   - (a) EGF: A protein which triggers the activation of the pathway by attaching its receptor (EGFR) in the cell membrane
   - (b) EGFR: A receptor that is equated with activated tyrosine kinase receptors
   - (c) GAP: A protein near the cell membrane
   - (d) PAK: A protein near the cell membrane
   - (e) PKC: A protein in the cytosol
   - (f) PP2A: A protein near the cell membrane or in the cytosol
   - (g) PP5: A protein near the cell membrane
   - (h) RKIP: A protein in the cytosol
   - (i) RKIP.p: RKIP mono phosphorylated either by PKC or ERK on the binding sites S153 and S99, respectively
   - (j) RSK: An inactive protein in the cytosol
   - (k) RSK.A: The active RSK, which is activated by ERK.p2
   - (l) TF: A transcription factor (like Elk or SAP proteins), which will be double phosphorylated by the active ERK, in the nucleus

References

Bower, J. M. and Bolouri, H. (2001). *Computational Modelling of Genetic and Biochemical Networks*. Massachusetts Institute of Technology, second edition. 851, 855, 858

Brent, R. (2004). "A partnership between biology and engineering." *Nature Biotechnology*, 22: 469–482. 854

— (2005). "A fishing buddy for hypothesis generators." *Science*, 308: 504–506. 852

Cao, Y., Gillespie, D. T., and Petzold, L. R. (2005). "Avoiding negative populations in explicit Poisson tau-leaping." *Journal of Chemical Physics*, 123: 054104.1–054104.8. 852

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC, second edition. 858, 859

Chang, L. and Karin, M. (2001). "Mammalian MAP kinase signalling cascades." *Nature*, 410(6824): 37–40. 853

Endy, D. and Brent, R. (2001). "Modelling cellular behaviour." *Nature*, 409: 391–395. 852, 854

Eraker, B. (2001). "MCMC analysis of diffusion models with application to finance." *Journal of Business and Economic Statistics*, 19(2): 177–191. 858, 863

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC. 861

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC. 859

Gibson, M. A. and Bruck, J. (2000). "Efficient exact stochastic simulation of chemical systems with many species and many channels." *Journal of Physical Chemistry*, A(104): 1876–1889. 851

Gillespie, D. T. (1977). "Exact stochastic simulation of coupled chemical reactions." *Journal of Physical Chemistry*, 81(25): 2340–2361. 851, 852, 855, 865

— (1992). "A rigorous derivation of the chemical master equation." *Physica A*, 188: 404–425. 851

— (1996). "The multivariate Langevin and Fokker-Planck equations." *American Journal of Physics*, 64(10): 1246–1257. 858

— (2000). "The chemical Langevin equation." *Journal of Chemical Physics*, 113: 297–306. 858

— (2001). "Approximate accelerated stochastic simulation of chemically reacting systems." *Journal of Chemical Physics*, 115: 1716–1733. 852

Golightly, A. and Wilkinson, D. J. (2005). "Bayesian inference for stochastic kinetic models using a diffusion approximation." *Biometrics*, 61(3): 781–788. 852, 858, 860, 862, 863, 867, 868

— (2006a). "Bayesian sequential inference for nonlinear multivariate diffusions." *Statistics and Computing*, 16: 323–338. 875

— (2006b). "Bayesian sequential inference for stochastic kinetic biochemical network models." *Journal of Computational Biology*, 13(3): 838–851. 875

— (2008). "Bayesian inference for nonlinear multivariate diffusion models observed with error." *Computational Statistics and Data Analysis*, 52(3): 1674–1693. 858

Hornberg, J. J. (2005). "Towards integrative tumor cell biology control of MAP kinase signalling." Ph.D. thesis, Vrije Universiteit, Amsterdam. 852

Kampen, N. G. V. (1981). *Stochastic Processes in Physics and Chemistry*. Amsterdam: North-Holland. 857, 858

Kolch, W. (2000). "Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions." *Biochemical Journal*, 351: 289–305. 852, 853, 854

Kolch, W., Calder, M., and Gilbert, D. (2005). "When kinases meet mathematics: the systems biology of MAPK signalling." *FEBS Letters*, 579: 1891–1895. 853, 854

Lawrence, E. (2005). *Henderson's Dictionary of Biology*. Pearson Prentice Hall. 852

Morton-Firth, C. and Bray, D. (1998). "Predicting temporal fluctuations in an signalling pathway." *Journal of Theoretical Biology*, 192: 117–128. 851

Orton, R., Sturm, O. E., Vyshemirsky, V., Calder, M., Gilbert, D. R., and Kolch, W. (2005). "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway." *Biochemical Journal*, 392: 249–261. 851

Purutçuoğlu, V. and Wit, E. (2006). "Exact and approximate stochastic simulations of the MAPK pathway and comparisons of simulations' results." *Journal of Integrative Bioinformatics*, 3(2). 853

— (2008). "Inclusion of convoluted measurements in Bayesian inference of the MAPK/ERK pathway via multivariate diffusion model." In Sezerman, U. (ed.), *Proceeding of the Third International Symposium on Health, Informatics and Bioinformatics*. Sabancı University, İstanbul, Turkey, 18-20 May, 2008. CD-Rom. 866

Risken, H. (1984). *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer-Verlag. 858

Roberts, G. O. and Stramer, O. (2001). "On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm." *Biometrika*, 88(3): 603–621. 858

Schoeberl, B., Eichler-Jonsson, C., Gilles, E. D., and Müller, G. (2002). "Computational modelling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors." *Nature Technology*, 20: 370–375. 852, 853

Tian, T. and Burrage, K. (2004). "Binomial leap methods for simulating stochastic chemical kinetics." *Journal of Chemical Physics*, 121(21): 10356–10364. 852

Turner, T. E., Schnell, S., and Burrage, K. (2004). "Stochastic approaches for modelling in vivo reactions." *Computational Biology and Chemistry*, 28: 165–178. 851, 857

Vyshemirsky, V., M.Girolami, Gormand, A., and Kolch, W. (2006). "A Bayesian analysis of the ERK signalling pathway." DCS Technical Report Series TR-2006-227, Department of Computing Science, University of Glasgow. 852, 854, 866

Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC. 852, 855, 857, 858, 865

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley and Sons. 852

Yeung, K., Janosch, P., McFerran, B., Rose, D. W., Mischak, H., Sedivy, J. M., and Kolch, W. (2000). "Mechanism of suppression of the Raf/MEK/Extracellular signal-regulated kinase pathway by the Raf Kinase inhibitor protein." *Molecular and Cellular Biology*, 20(9): 3079–3085. 853

Yeung, K., Seitz, T., Li, S., Janosch, P., McFerran, B., Kaiser, C., Fee, F., Katsanakis, K. D., Rose, D. W., Mischak, H., Sedivy, J. M., and Kolch, W. (1999). "Suppression of Raf-1 kinase activity and MAP kinase signalling by RKIP." *Nature*, 401: 173–177. 853

Acknowledgments

We would like to thank to anonymous referees for their comments which contributed to the improvement of the paper.