Simulation Methods for Optimal Experimental Design in Systems Biology

D. Faller
Freiburg Center for Data Analysis and Modeling
Eckerstr. 1, 79104
Freiburg, Germany

U. Klingmüller
Max-Planck-Institute for Immunology
Stübeweg 51, 79108
Freiburg, Germany

J. Timmer
Freiburg Center for Data Analysis and Modeling
Eckerstr. 1, 79104 Freiburg, Germany
jeti@fdm.uni-freiburg.de

To obtain a systems-level understanding of a biological system, the authors conducted quantitative dynamic experiments from which the system structure and the parameters have to be deduced. Since biological systems have to cope with different environmental conditions, certain properties are often robust with respect to variations in some of the parameters. Hence, it is important to use optimal experimental design considerations in advance of the experiments to improve the information content of the measurements. Using the MAP–Kinase pathway as an example, the authors present a simulation study investigating the application of different optimality criteria. It is demonstrated that experimental design significantly improves the parameter estimation accuracy and also reveals difficulties in parameter estimation due to robustness.

Keywords: Experimental design, systems biology, simulation methods

1. Introduction

Systems biology examines how individual components of biological systems dynamically interact with each other to obtain a systems-level understanding of the biological process under consideration [1]. For this approach, quantitative measurements are made. These measurements are then used to statistically assess the performance of a proposed mathematical model describing the biological system under investigation [2]. Since in general, the parameters are unknown, the deviances of the model predictions can either be caused by wrong parameters or by a wrong model structure. Hence, the model parameters are estimated from the experimental data. Based on these parameter estimates, the model can be statistically validated if it is able to reproduce the observed dynamic behavior [3].

SIMULATION, Vol. 79, Issue 12, December 2003 717-725

DOI: 10.1177/0037549703040937

To successfully identify the model structure and reliably estimate the system parameters, these have to be identifiable given the current experimental design. If this is not the case, the experimental design has to improved. Moreover, even if identifiability is given, an enhanced experimental design may drastically improve the estimation accuracy. Since quantitative time-resolved measurements are time and cost intensive, these improvements will also reduce the experimental costs needed to achieve a prespecified accuracy.

In the following, we discuss how optimality of an experimental design can be defined and calculated. Then a simulation study using the MAP–Kinase signaling pathway, which is known to be a robust pathway involved in a variety of different regulatory systems, is presented. The main aim of this simulation study is to investigate how computer simulations in advance of the experiments can be used to improve the experimental design. Experimental design can be used to optimize the selection of the time points at which the measurements are recorded and to calculate optimal

stimulations of the system. Here we are mainly concerned with optimization of the input to the MAP–Kinase pathway. All measurements are recorded on a prespecified uniform grid of time points. Using this pathway as an example, the advantages and disadvantages of different optimality criteria are discussed.

This article is structured as follows. Section 2 introduces the basic system identification theory and briefly reviews some concepts of optimal experimental design. After a short introduction of the MAP–Kinase pathway in section 3, the results of the simulation study are presented in section 4.

2. System Identification

Dynamical biological systems can be described by a variety of different mathematical models, ranging from partial or ordinary differential equations to stochastic differential equations. In the following, we will restrict ourselves to models defined in terms of ordinary differential equations. In this case, the time evolution of the system state $\boldsymbol{x}(t) \in \mathbb{R}^K$ is given by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\theta}, u(t)), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^P$ denotes the parameters of the system, and $u(t)$ is the input to the system. Note that $\boldsymbol{\theta}$ may not only include dynamic parameters such as reaction rates but also unknown initial conditions of the system state vector $\boldsymbol{x}(0)$. For the sake of simplicity, the input function $u(t)$ is assumed to be scalar, but the methods presented in the following do not depend on this assumption.

Often, not all components of the system can be measured directly. The observation function $\boldsymbol{g}$ describes which properties $\boldsymbol{y}^M(t, \boldsymbol{\theta}) \in \mathbb{R}^L$ of the system can be measured,

$$\boldsymbol{y}^M(t_i, \boldsymbol{\theta}, u) = \boldsymbol{g}(\boldsymbol{x}(t_i, \boldsymbol{\theta}, u)) \quad i = 1, \ldots, N\ . \tag{2}$$

The observation function $\boldsymbol{g}$ and the input function $u(t)$, together with the specification of the time points t_i at which the measurements are recorded, completely specify the experimental design used. The observations $\boldsymbol{y}^D(t_i) \in \mathbb{R}^L$ are given by

$$\boldsymbol{y}^D(t_i) = \boldsymbol{y}^M(t_i, \boldsymbol{\theta}_0, u) + \boldsymbol{\epsilon}_i \quad i = 1, \ldots, N\ . \tag{3}$$

The true parameter vector is denoted by $\boldsymbol{\theta}_0$, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^L$ describes the observation error at time t_i. Most often, the observation error is assumed to be distributed according to

$$\epsilon_{ij} = N(0, \sigma_{ij}^2), \quad i = 1, \ldots, N \quad j = 1, \ldots, L\ . \tag{4}$$

The variances σ_{ij} can be estimated from repetitions of the experiments.

Knowledge of the system structure (equation (1)) and the experimental design (equation (2)) allows for evaluating the identifiability of the system structure and the parameters. This is a prerequisite for the inference of system properties, which is the final goal of systems biology.

2.1 Identifiability

One distinguishes between structural, local, and practical identifiability. Structural identifiability of parameters is a theoretical property of the model structure depending only on the observation function $\boldsymbol{g}$ and the input function $u(t)$. It does not depend on the observational noise or the number of data points measured but rather is an asymptotic property in the limit of an infinite number of observations.

The parameters $\boldsymbol{\theta}$ of a model are structural identifiable if

$$\begin{aligned} &\forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^P\ , \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Rightarrow \exists\, t \text{ with} \\ &\boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_1, u)) \neq \boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_2, u))\ . \end{aligned} \tag{5}$$

This definition for structural identifiability is rather strict. One can easily imagine realistic situations in which the parameters are not identifiable according to this definition but nevertheless would be identifiable for a reasonably restricted set of all possible parameters. Hence, for all practical purposes, it is crucial to restrict the set of possible parameters to decide if parameters can be identified with the current experimental protocol. This leads to the definition of local identifiability.

The parameters $\boldsymbol{\theta}$ of a model are locally identifiable in a ϵ neighborhood of a parameter $\boldsymbol{\theta}_0$ if

$$\begin{aligned} &\forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \{\boldsymbol{\theta} \in \mathbb{R}^P \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon\}\ , \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \\ &\quad \Rightarrow \exists\, t \text{ with } \boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_1, u)) \neq \boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_2, u))\ . \end{aligned} \tag{6}$$

In contrast to the theoretical properties of structural and local identifiability, the practical identifiability of parameters is limited by the finite amount of data and observational noise. Hence, if there are large observation errors or few data, and thus no reliable estimate of the parameters is possible, these parameters are called *practical nonidentifiable*.

There are different analytical approaches to prove the structural identifiability of parameters (see [4–6] or [7] for an overview). However, for large nonlinear systems, these analytical approaches are quite complicated. For the purposes needed in this study, local identifiability near the true parameter value is sufficient. To prove local identifiability, we used an approach based on the parameter estimation accuracy [8], which will be discussed in section 2.3.

Identifiability of parameters has to be distinguished from model selection, whereby one tries to discriminate between different possible models. If the set of all models under investigation is given by $\mathcal{M} = \{\boldsymbol{f} \mid \text{possibly true model}\}$, the true model is structural identifiable if

$$\begin{aligned} &\forall\, M_1(\boldsymbol{\theta}_1), M_2(\boldsymbol{\theta}_2) \in \mathcal{M}\ , M_1 \neq M_2 \\ &\quad \Rightarrow \exists\, t \text{ with } \boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_1, u)) \neq \boldsymbol{g}(\boldsymbol{x}(t, \boldsymbol{\theta}_2, u))\ . \end{aligned} \tag{7}$$

2.2 Parameter Estimation

In all practical applications, parameters of the system have to be estimated since they cannot be measured directly. In

addition, the measured values (see equation (3)) contain observational errors. Due to its superior statistical properties, the most popular approach to parameter estimation is the maximum likelihood method [9–11].

In the case of Gaussian observational noise, the maximum likelihood estimation corresponds to a minimization of the weighted residual sum of squares,

$$\chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{L}\sum_{i=1}^{N}\left(\frac{y_j^D(t_i) - y_j^M(t_i,\boldsymbol{\theta},u)}{\sigma_{ij}}\right)^2 . \quad (8)$$

The asymptotic distribution of the least squares estimate $\hat{\boldsymbol{\theta}}$ can be computed analytically. To this end, one assumes that in the limiting case of an infinite number of observations, the deviation $\Delta\boldsymbol{\theta}$ between the real and estimated parameters is also small. Hence, the observation function can be expanded in a Taylor series, yielding

$$\begin{aligned} y_j^M(t_i,\boldsymbol{\theta},u) &= y_j^M(t_i,\boldsymbol{\theta}_0,u) + \nabla_{\boldsymbol{\theta}}\, y_j \mid_{t_i,\boldsymbol{\theta}_0} (\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\ &= y_j^M(t_i,\boldsymbol{\theta}_0,u) + \left\{\nabla_{\boldsymbol{\theta}}\, g_j \right. \\ &\left. \nabla_{\boldsymbol{x}}\, g_j\left([\nabla_{\boldsymbol{\theta}}\, x_1]^T,\ldots,[\nabla_{\boldsymbol{\theta}}\, x_K]^T\right)^T\right\} \mid_{t_i,\boldsymbol{\theta}_0} (\boldsymbol{\theta}-\boldsymbol{\theta}_0). \end{aligned} \quad (9)$$

After inserting this result in the minimization functional, one obtains

$$\begin{aligned} \chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{L}\sum_{i=1}^{N}\Bigg[&\frac{\epsilon_{ij}^2}{\sigma_{ij}^2} - \frac{2\,\epsilon_{ij}}{\sigma_{ij}^2}\nabla_{\boldsymbol{\theta}}\, y_j(t_i,\boldsymbol{\theta}_0)\;\Delta\boldsymbol{\theta} \\ &+ [\Delta\boldsymbol{\theta}]^T\left(\frac{1}{\sigma_{ij}^2}\left[\nabla_{\boldsymbol{\theta}} y_j\right]^T\left[\nabla_{\boldsymbol{\theta}} y_j\right]\right)\Big|_{t_i,\boldsymbol{\theta}_0}\Delta\boldsymbol{\theta}\Bigg]. \end{aligned} \quad (10)$$

The minimization of $\chi^2(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ yields the following equation for the estimated deviation of the parameter vector $\Delta\boldsymbol{\theta}$,

$$\begin{aligned} &\left\{\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{1}{\sigma_{ij}^2}\left[\nabla_{\boldsymbol{\theta}}\, y_j\right]^T\left[\nabla_{\boldsymbol{\theta}}\, y_j\right]\right\}\Delta\boldsymbol{\theta} =: F\,\Delta\boldsymbol{\theta} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{\epsilon_{ij}}{\sigma_{ij}^2}\left[\nabla_{\boldsymbol{\theta}}\, y_j\right]^T , \end{aligned} \quad (11)$$

where the so-called Fisher information matrix F was introduced [9]. The equation above can be solved, and one obtains

$$\Delta\boldsymbol{\theta} = F^{-1}\sum_{j=1}^{L}\sum_{i=1}^{N}\frac{\epsilon_{ij}}{\sigma_{ij}^2}\left[\nabla_{\boldsymbol{\theta}}\, y_j\right]^T . \quad (12)$$

Finally, the covariance matrix Σ of the estimated parameter vector is computed by

$$\Sigma = \langle\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T\rangle = F^{-1} . \quad (13)$$

To evaluate this covariance matrix, we need the derivations of the observation function with respect to the parameters, $\nabla_{\boldsymbol{\theta}}\; y_j(t_i)$. Taking into account equation (9), one hence needs the derivation of $\boldsymbol{g}$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{x}$. In addition, the derivations $\nabla_{\boldsymbol{\theta}}\, x_k(t_i)$ have to be computed from the system of ordinary differential equations,

$$\frac{\partial}{\partial t}\left(\nabla_{\boldsymbol{\theta}}\; x_k\right) = \sum_{r=1}^{K}\frac{\partial f_k(\boldsymbol{x},\boldsymbol{\theta})}{\partial x_r}\;\nabla_{\boldsymbol{\theta}} x_r + \nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{x},\boldsymbol{\theta}) , \quad (14)$$

with the initial conditions $(\nabla_{\boldsymbol{\theta}}\; x_k)(0) = \nabla_{\boldsymbol{\theta}}\; x_k(0)$. If the Fisher information matrix and thus the covariance matrix of the estimated parameters are known, the asymptotic confidence intervals for the estimates can be computed from the multivariate normal distribution,

$$p(\boldsymbol{\theta}) = \frac{\sqrt{\mathrm{Det}(F)}}{[2\pi]^{P/2}}\exp(-\frac{1}{2}\boldsymbol{\theta}^T F\boldsymbol{\theta}) . \quad (15)$$

2.3 Local Identifiability

Local identifiability of the parameters in a small neighborhood of the true parameter values can be assessed using the parameter estimation accuracy [8]. If at least one of the parameters is not identifiable, there must exist a functional relationship between some of the parameters. This would result in a joint probability distribution $p(\boldsymbol{\theta})$, which is not a multivariate normal distribution. Technically, such a nonidentifiability would result in a covariance matrix that does not have full rank. Hence, the condition number, the ratio of the largest eigenvalue to the smallest eigenvalue, would asymptotically tend to infinity. Therefore, the asymptotic behavior of the condition number can be used to assess local parameter identifiability.

Note that if there are large differences in the size of the parameters to be estimated, it will be advantageous to use relative estimation errors. In such a setting, the correlation matrix is used instead of the covariance matrix.

2.4 Optimal Experimental Design

Basically, the information content of a measurement can be quantified by the covariance matrix Σ of the estimated parameters. Simply speaking, the smaller the joint confidence intervals for the estimated parameters are, the more information is contained in the experiment. Mostly, four measures of the information content are distinguished [12-15]:

- A-optimal design: $\max(\mathrm{Tr}(F))$
- D-optimal design: $\min(\det(\Sigma))$
- E-optimal design: $\min(\lambda_{max}(\Sigma))$
- Modified E-optimal design: $\min(\lambda_{max}(\Sigma)/\lambda_{min}(\Sigma))$

The A-optimal design tries to maximize the trace of the Fisher information matrix F. However, this criterion is rarely used since it can lead to noninformative experiments

[16], with a covariance matrix that is not positive definite. A D-optimal design minimizes the determinant of the covariance matrix Σ and can thus be interpreted as minimizing the geometric mean of the errors in the parameters. The largest error is minimized by the E-optimal design, which corresponds to a minimization of the largest eigenvalue of the covariance matrix. The modified E-optimal design minimizes the ratio of the largest to the smallest eigenvector and thus optimizes the functional shape of the confidence intervals. Other measures for the information content are possible (see, e.g., [17] for a criterion based on the largest actual estimation error).

The optimization of the different optimality criteria can be quite intensive in terms of computing time. For the computation of the optimal stimulation of the MAP–Kinase pathway in section 3, a polynomial parameterization of the input function is used. Then, a minimization algorithm based on a sequential quadratic programming method, implemented in the NAG numerical libraries [18], is used to obtain the optimal stimulation.

3. Application to the MAP–Kinase Cascade

Upon external stimulation, cell surface receptors initiate a network of internal signaling pathways that are essential to cell function by transmitting the signal through the cytoplasm to the nucleus. One of these signaling pathways, the mitogen-activated protein kinase (MAPK) signal cascade, is a highly conserved pathway found in a variety of eukaryotic organisms. Therefore, cumulative efforts over the past decade have been carried out to explore the functionality and properties of this module [19–25; for an overview, see 26–28].

MAP–Kinase is assumed to be composed of three kinases: MAPK kinase kinase (Raf), MAPK kinase (Mek), and MAPK (Erk). These kinases can be activated by phosphorylation. The activated form then catalyzes the activation of the downstream kinase. Double-phosphorylated Erk, the final step of the signaling cascade, translocates into the nucleus and triggers gene expression. The MAP–Kinase forms the final step of numerous signaling pathways, as shown in Figure 1.

In the following, we will concentrate on the last step of the MAPK cascade, which is shown in Figure 2.

The reactions occurring in this last step of the signaling cascade are

$$[\mathrm{Erk}]+[\mathrm{Mek}^{**}] \underset{b_1}{\overset{a_1}{\rightleftarrows}} [\mathrm{Erk}-\mathrm{Mek}^{**}] \xrightarrow{c_1} [\mathrm{Erk}^{*}]+[\mathrm{Mek}^{**}],$$

$$[\mathrm{Erk}^{*}]+[\mathrm{Mek}^{**}] \underset{b_3}{\overset{a_3}{\rightleftarrows}} [\mathrm{Erk}^{*}-\mathrm{Mek}^{**}] \xrightarrow{c_3} [\mathrm{Erk}^{**}]+[\mathrm{Mek}^{**}],$$

$$[\mathrm{Erk}^{**}]+[\mathrm{Pase}] \underset{b_4}{\overset{a_4}{\rightleftarrows}} [\mathrm{Erk}^{**}-\mathrm{Pase}] \xrightarrow{c_4} [\mathrm{Erk}^{*}]+[\mathrm{Pase}],$$

$$[\mathrm{Erk}^{*}]+[\mathrm{Pase}] \underset{b_2}{\overset{a_2}{\rightleftarrows}} [\mathrm{Erk}^{*}-\mathrm{Pase}] \xrightarrow{c_2} [\mathrm{Erk}]+[\mathrm{Pase}].$$

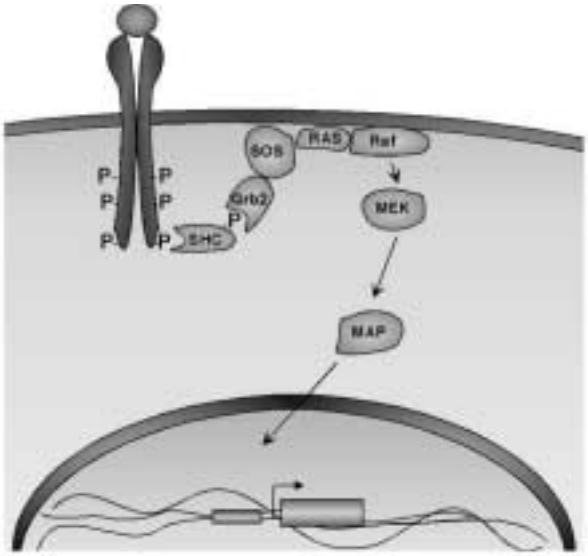


Figure 1. The typical graphical representation of signaling pathways. The receptor spans the cell membrane. Its activation upon extracellular stimulation is mediated to the DNA in the cell nucleus by a cascade of phosphorylations. The last steps of the cascade form the MAP–Kinase cascade.

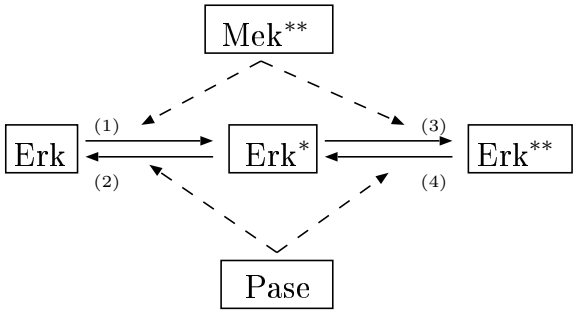


Figure 2. The last step of the mitogen-activated protein kinase (MAPK) cascade. Activated MAPK kinase (Mek**) catalyzes the activation of MAPK (Erk) by phosphorylation, resulting in the activated form Erk**. The deactivation of the active form is catalyzed by the phosphatase (Pase).

Assuming the mass action law for each of these reactions, a system of coupled ordinary differential equations for the involved proteins can be derived:

$$\begin{aligned}[\mathrm{Erk}^{*}] &= c_1[\mathrm{Erk}-\mathrm{Mek}^{**}] - a_3[\mathrm{Erk}^{*}][\mathrm{Mek}^{**}] \\ &\quad + b_3[\mathrm{Erk}^{*}-\mathrm{Mek}^{**}] + c_4[\mathrm{Erk}^{**}-\mathrm{Pase}] \\ &\quad - a_2[\mathrm{Erk}^{*}][\mathrm{Pase}] + b_2[\mathrm{Erk}^{*}-\mathrm{Pase}], \\ [\mathrm{Erk}^{**}] &= c_3[\mathrm{Erk}^{*}-\mathrm{Mek}^{**}] - a_4[\mathrm{Erk}^{**}][\mathrm{Pase}]\end{aligned}$$

$$
\begin{aligned}
&+ b_2[\mathrm{Erk}^* - \mathrm{Pase}],\\
[\mathrm{Erk} - \mathrm{Mek}^{**}] &= a_1[\mathrm{Erk}][\mathrm{Mek}^{**}]\\
&- (b_1 + c_1)[\mathrm{Erk} - \mathrm{Mek}^{**}],\\
[\mathrm{Erk}^* - \mathrm{Mek}^{**}] &= a_3[\mathrm{Erk}^*][\mathrm{Mek}^{**}]\\
&- (b_3 + c_3)[\mathrm{Erk}^* - \mathrm{Mek}^{**}],\\
[\mathrm{Erk}^* - \mathrm{Pase}] &= a_2[\mathrm{Erk}^*][\mathrm{Pase}]\\
&- (b_4 + c_4)[\mathrm{Erk}^* - \mathrm{Pase}],\\
[\mathrm{Erk}^{**} - \mathrm{Pase}] &= a_3[\mathrm{Erk}^{**}][\mathrm{Pase}]\\
&- (b_4 + c_4)[\mathrm{Erk}^{**} - \mathrm{Pase}].
\end{aligned}
$$

Since the model assumes that the total Erk concentration is constant, the nonactivated Erk is given in terms of the total Erk concentration,

$$[\mathrm{Erk}](t) = [\mathrm{Erk}]_{\mathrm{total}} - [\mathrm{Erk}^*](t) - [\mathrm{Erk}^{**}](t)\ . \qquad (16)$$

Altogether, there are 14 dynamical parameters involved: the 12 reaction rates $a_i, b_i, c_i, i \in 1, \ldots, 4$ and the total phosphatase (Pase) and kinase (Erk) concentrations. The parameters a_i denote the rates at which the substrate binds to the enzyme, b_i denotes the corresponding breaking rates, and c_i denotes the rate at which the actual activation reaction occurs. For this system, Mek^{**} serves as input, while Erk^{**} can be regarded as the output of the system. The initial concentrations of all phosphorylated Erks and complexes of phosphorylated Erks with Meks or phosphatases are zero.

4. Results

To understand the functionality of the MAPK signaling pathway, a quantitative description of the signal dynamic is necessary [24]. The crucial question for such a systems-level approach is how the information content of measurements can be quantified and how experiments can be optimized to yield a maximal amount of information about the observed system.

To answer these questions, we performed a simulation study, using the last step of the MAPK cascade as an example. We assumed that the dynamical behavior of both the inactive (Erk) and the active (Erk^{**}) form of MAPK can be measured subject to a known stimulation of the activated MAPK (Mek^{**}) (see Fig. 2).

The dynamic behavior of the system and thus the parameter estimation accuracy depend on the actual parameter values. To demonstrate the prospects of experimental design considerations, the following parameters have been chosen for the purpose of the presented simulation study:

$$
\begin{aligned}
&a_i = 0.5,\ b_i = 0.6,\ c_i = 0.9,\quad i = 1, \ldots, 4\\
&\mathrm{Pase}_{tot} = 20,\quad \mathrm{Erk}_{tot} = 50.
\end{aligned}
\qquad (17)
$$

In general, for optimal experimental design investigations, a rough knowledge of the actual parameters is necessary.

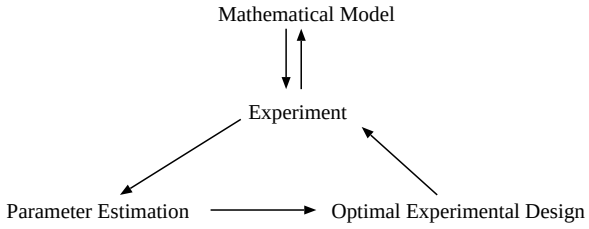


Figure 3. Illustration of the model-building process. Starting from a hypothetical mathematical model, experiments are performed to validate and improve the model. After each parameter estimation, optimal experimental design can be used to perform further informative experiments, which increase the information about the system.

Usually, a first simple experiment is performed to get a first estimate of the parameters. Then, these parameters are used to design new experiments. In an iterative process, then, the information from these new experiments is included in the design process (see Fig. 3).

A typical time evolution of the system response—the two measured components Erk and Erk^{**}, subject to a stimulus with a quartic input function (see equation (18)), including Gaussian observational noise ($\sigma = 0.3$)—is shown in Figure 4a.

To simplify the discussion of the results, we restrict ourselves to the estimation of two parameters from the first reaction in Figure 2. We will discuss the estimation of a_1, b_1 and a_1, c_1; since these two cases showed quite different behavior, all other parameters were fixed. However, the concepts discussed can be generalized to a larger parameter space.

4.1 Parameter Identifiability

Before one tries to estimate these parameters, one must ensure that the parameters are identifiable at all. To investigate identifiability, we simulated data according to the model, and observational noise was added. Then, minimization of $\chi^2(\boldsymbol{\theta})$ (see equation (8)) yielded an estimate of the parameters and the covariance matrix. As discussed in section 2.3, the asymptotic behavior of the condition number of this covariance matrix can be used to assess the identifiability of the parameters [8]. Figure 4b shows the behavior of the condition number if the parameters a_1 and b_1 are estimated and c_1 is fixed. It can be seen that the condition number does not tend to infinity with an increasing number N of data points; hence, in this setting, the parameters are locally identifiable. However, the large values of the condition number already indicate that there will be large estimation errors, which probably limit the practical identifiability. Estimation of a_1 and c_1 showed the same qualitative behavior, but the values of the condition number were much smaller.

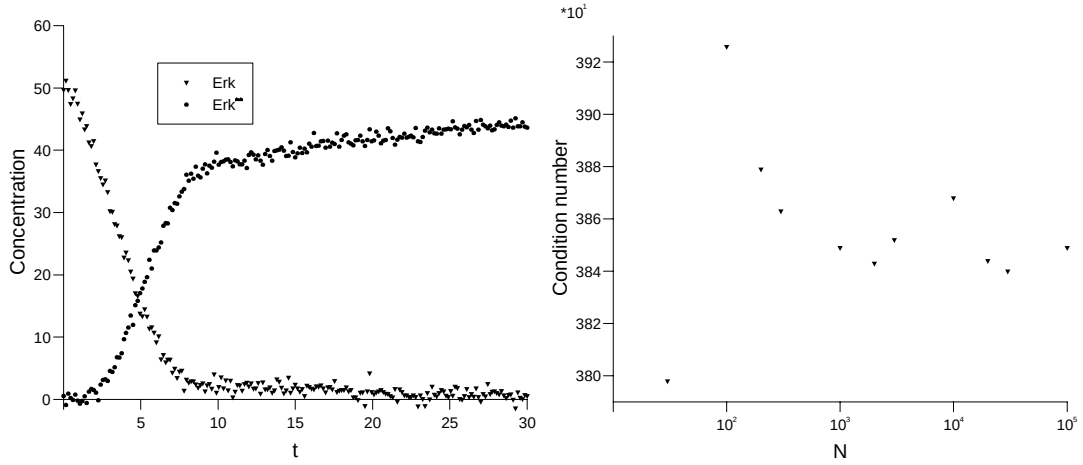


Figure 4. (a) Typical time evolution of the measured activated and inactivated mitogen-activated protein kinase (MAPK) concentrations, Erk and Erk**, for a linear input signal with $p_1 = 0.267$ (see equation (18)) and parameters for the MAP–Kinase cascade $a_i = 0.5$, $b_i = 0.6$, $c_i = 0.9$ and $\text{Pase}_{tot} = 20$, $\text{Erk}_{tot} = 50$. (b) Asymptotic behavior of the condition number of the covariance matrix when the rate constants a_1 and b_1 are estimated.

4.2 Optimal Experimental Design

Designing optimal experiments typically involves decisions about what components of the system are measured at which points in time. In addition, the stimulation of the system has to be chosen. For our simulation study, we assume that the time evolution of the inactive (Erk) and activated (Erk**) MAPK is measured at a given time resolution, and we try to find the most informative input profile for Mek**. To this end, we used a polynomial parameterization of the input function

$$\begin{aligned} \text{Mek}^{**}(t) &= 2 + \sum_{k=1}^{d} p_k\, t^k, \\ 2 \le \text{Mek}^{**}(t) &\le 10 \quad \forall t \in [0, 30], \end{aligned} \tag{18}$$

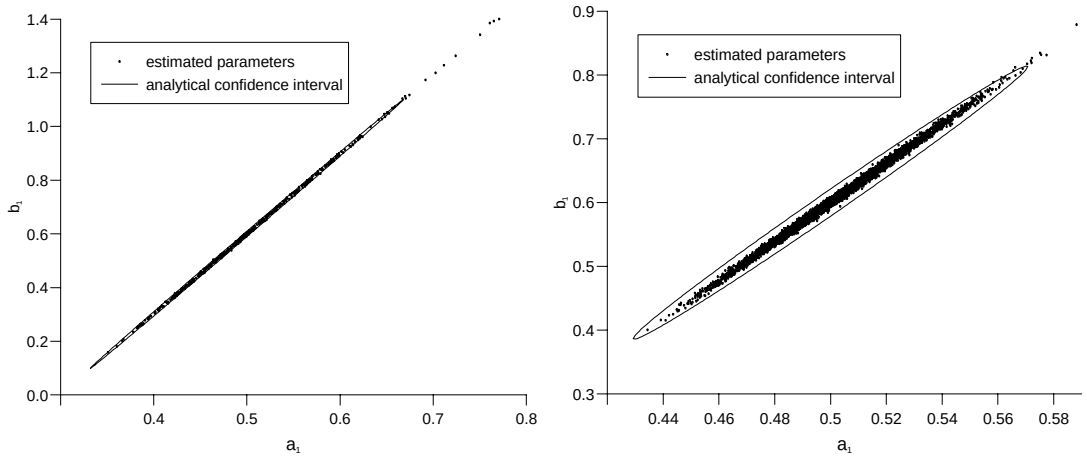
where the maximal and minimal Mek** concentrations are bounded. Using this parameterization of the input, we optimized the different design criteria discussed in section 2.4 to obtain the optimal input for different degrees d of the input function.

The confidence intervals calculated according to equation (15) are valid in the asymptotic case of an infinite number of observations. To assess the validity of this assumption in the present case, observational noise is added to simulated data, and the parameters are estimated. Figure 5 shows the 95% confidence ellipsoids calculated according to equation (15) for a modified E-optimal design, together with 200 samples of estimated parameters for the optimal linear and 500 samples for the quartic input function. It can be seen that the shape and size of the computed confidence ellipsoids resemble the distribution of the estimated parameters, while the estimated parameters are slightly shifted toward higher parameter values. This indicates that in the current setting, the asymptotic covariance matrix can be used to optimize the experimental design. Note that also if this assumption would not be valid, one can still use the estimated parameters to reconstruct the covariance matrix and thus optimize the experimental design. However, the numerical efforts in this case are much larger.

Figure 6 shows the shape of the 95% confidence ellipsoid calculated according to equation (15) for an estimation of the parameters a_1 and b_1, using two optimal input functions of different degree d. The modified E-optimal design was used to find the optimal linear and quartic input function. It can be seen that there is a substantial improvement in parameter estimation accuracy in moving from an optimal linear input to the optimal quartic input function. In this case, the estimation error in the parameter a_1 was reduced by approximately 60%.

Figure 7a shows the 95% confidence intervals for estimation of a_1, b_1 for the D-optimal, E-optimal, and modified E-optimal design using the largest degree of the input function. There is a strong correlation between the estimates of these two parameters. These correlations can lead to practical nonidentifiability for less optimal input functions, despite the structural identifiability that is present in this case.

Especially in the case of correlated estimates, it is important to be aware of the multidimensionality of the confidence interval. For example, the estimation error of a_1, if



(a) (b)

Figure 5. The 95% confidence intervals for the parameters a_1 and b_1 in the case of a modified E-optimal design for the best linear (a) and quartic (b) input function, together with 200 (a) and 500 (b) samples of estimated parameters

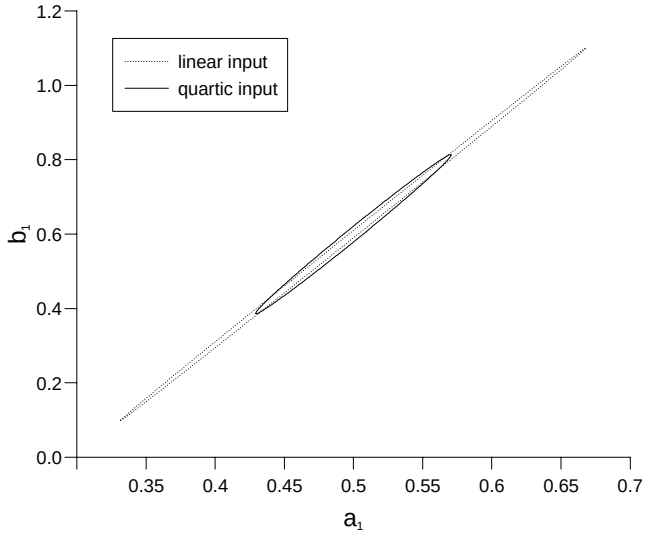
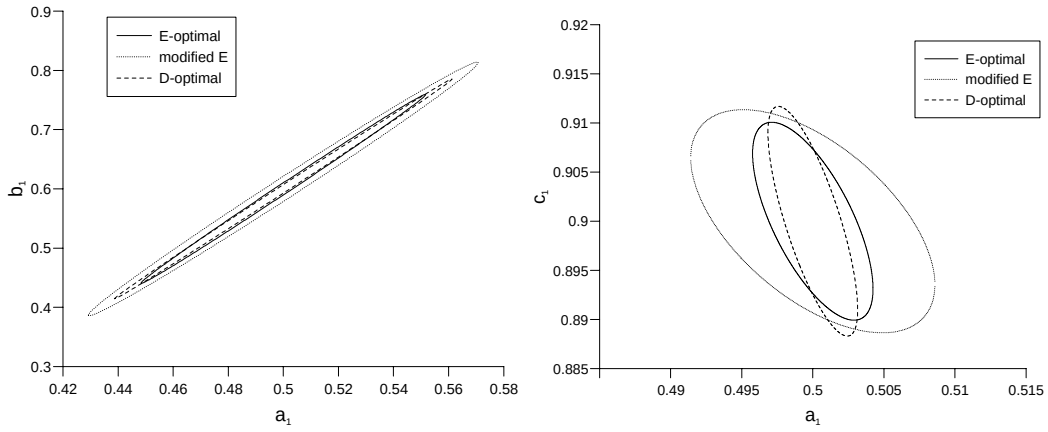


Figure 6. The 95% confidence intervals for the parameters a_1 and b_1 in the case of a modified E-optimal design for the best linear and quartic input function

the parameter b_1 would be fixed, is about 10 times smaller than in the case when both parameters are estimated. The estimation error of a_1, if b_1 is fixed, can be seen in Figure 7a as the projection of the intersection of the $b_1 = 0.6$ line and the confidence ellipsoid to the a_1 axis.

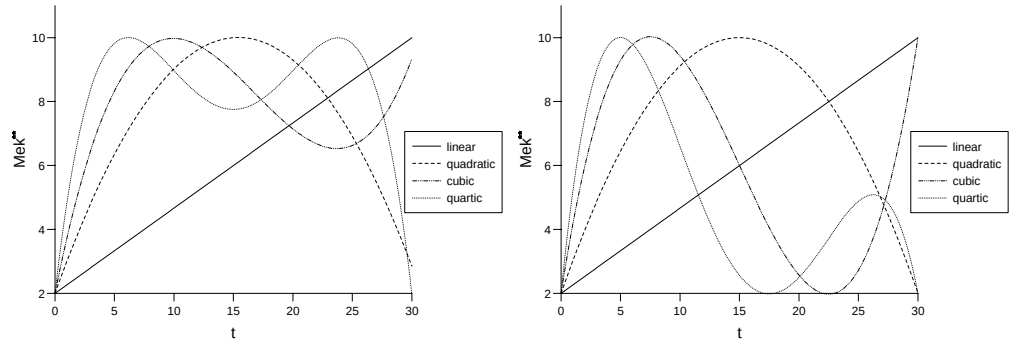
Figure 7b shows the confidence regions for an estimation of a_1 and c_1. There is only a small correlation between these two parameters present, and hence these parameters are better suited to a discussion of the different design criteria. As expected, the D-optimal design results in the smallest volume in the parameter space of the confidence region, while the E-optimal design minimizes the largest principal axis of the confidence ellipsoid. In contrast to these criteria, the modified E-optimal design, which optimizes the shape of the confidence region, results in larger estimation errors, especially for the parameter a_1.

Figure 8 shows the optimal input functions for different degrees of the polynom from equation (18) for the

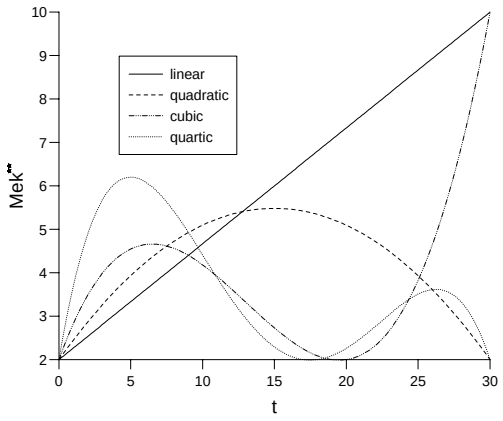


(a) (b)

Figure 7. (a) The 95% confidence intervals for the estimation of a_1, b_1 and a_1, c_1 (b) for different design criteria using a optimal quartic input function



(a) (b)



(c)

Figure 8. The optimal input functions for the E-optimal design (a), the modified E-optimal design (b), and the D-optimal design (c) for different degrees of the input function polynom

E-optimal design, the modified E-optimal design, and the D-optimal design.

5. Conclusion

Systems biology is based on quantitative dynamic measurements of biological systems, from which the system structure and the parameters have to be deduced. We have used the MAP–Kinase signaling cascade, a highly conserved regulatory module, to demonstrate the impact of improved experimental designs on parameter estimation accuracy. We have shown that by moving from a linear input function to a specific quartic input function, the parameter estimation error could be reduced by 60%.

In addition to improving the estimation accuracy, computer simulations can also help in detecting possible practical nonidentifiabilities in the parameter space. Practical nonidentifiability will result in a large condition number of the estimated covariance matrix (see Fig. 6 as an example). This example illustrates that, for example, for an analysis of robustness, where the sensitivity of the output to variations in the parameters is studied, it does not suffice to use variations in single parameters. In general, combinations of parameters are the most prospective candidates for nonidentifiabilities.

Since the modified E-optimal design resulted in larger confidence intervals, this simulation study would prefer the E- and D-optimal experimental designs. These designs resulted in roughly the similar shape and size of the confidence regions.

The software used for the presented simulations is available on request from the authors (jeti@fdm.uni-freiburg.de).

6. References

[1] Kitano, H. 2002. Systems biology: A brief overview. *Science* 295:1662-4.

[2] Swameye, I., T. G. Müller, J. Timmer, O. Sandra, and U. Klingmüller. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *Proceedings of the National Academy of Science* 100:1028-33.

[3] Cox, D. R., and D. V. Hinkley. 1994. *Theoretical statistics*. London: Chapman & Hall.

[4] Pohjanpalo, H. 1978. System identifiability based on power series expansion of the solution. *Mathematical Biosciences* 41:21-33.

[5] Ritt, J. F. 1932. *Differential equations from the algebraic standpoint*. Ann Arbor: American Mathematical Society.

[6] Wu, W. T. 1986. On zeros of algebraic equations—an application of Ritt principle. *Kexue Tongbao* 31:1-5.

[7] Müller, T. G. 2002. Modeling complex systems with differential equations. Ph.D. diss., Albert-Ludwigs Universität Freiburg [Online]. Available: www.freidok.uni-freiburg.de/volltexte/556/

[8] Müller, T. G., N. Noykova, M. Gyllenberg, and J. Timmer. 2002. Parameter identification in dynamical models of anaerobic wastewater treatment. *Mathematical Biosciences* 177-178:147-60.

[9] Fisher, R. A. 1912. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* 41:155-60.

[10] Schittkowski, K. 1994. Parameter estimation in systems of nonlinear equations. *Numerische Mathematik* 68:129-42.

[11] Ljung, L. 1999. *System identification*. Englewood Cliffs, NJ: Prentice Hall.

[12] Box, G. E. P., and W. J. Hill. 1967. Discrimination among mechanistic models. *Technometrics* 9:57-71.

[13] Munack, A. 1991. Optimization of sampling. In *Biotechnology, a multi-volume comprehensive treatise*, edited by K. Schügerl, 251-64. Weinheim: VCH, Weinheim.

[14] Munack, A. 1992. Some improvements in the identification of bioprocesses. In *Modeling and control of biotechnical processes*, 2nd IFAC Symposium, edited by M. N. Karim and G. Stephanopoulos, 89-94. San Diego: Elsevier.

[15] Baltes, M., R. Schneider, C. Sturm, and M. Reuss. 1994. Optimal experimental design for parameter estimation in unstructured growth models. *Biotechnology Progress* 10:480-8.

[16] Goodwin, G. C. 1987. Identification: Experiment design. In *Systems and control encyclopedia*, vol. 4, edited by M. Singh, 2257-64. Oxford: Pergamon Press.

[17] Lohmann, T., H. G. Bock, and J. P. Schlöder. 1992. Numerical methods for parameter estimation and optimal experimental design in chemical reaction systems. *Industrial and Engineering Chemistry Research* 31:54-7.

[18] Numerical Algorithms Group. 2002. *The NAG Fortran library manual, Mark 20*. Cambridge: Cambridge University Press.

[19] Koshland, D. E. 1987. Switches, thresholds and ultrasensitivity. *Trends in Biochemical Sciences* 12:225-9.

[20] Huang, C.-Y. F., and J. E. Ferrell. 1996. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Science* 93:10078-83.

[21] Asthagiri, A. R., and D. A. Lauffenburger. 2001. A computational study of feedback effects on signal dynamics in a mitogen–activated protein kinase (MAPK) pathway model. *Biotechnology Progress* 17:227-39.

[22] Heinrich, R., B. G. Neel, and T. A. Rapoport. 2002. Mathematical models of protein kinase signal transduction. *Molecular Cell* 9:957-70.

[23] Bhalla, U. S., P. T. Ram, and R. Iyengar. 2002. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signalling network. *Science* 297:1018-23.

[24] Ingolia, N. T., and A. W. Murray. 2002. History matters. *Science* 297:948-9.

[25] Kholodenko, B. N. 2002. MAP kinase cascade signaling and endocytic trafficking: A marriage of convenience. *TRENDS in Cell Biology* 12(4):173-7.

[26] Blüthgen, N., and H. Herzel. 2001. MAP–Kinase-cascade: Switch, amplifier or feedback controller. In *2nd Workshop on Computation of Biochemical Pathways and Genetic Networks*, edited by U. Kummer R. Gauges, and C. van Gend, 55-62. Berlin: Logos Verlag.

[27] Pearson, G., Robinson, T. B. Gibson, B. Xu, M. Karandikar, K. Berman, and M. H. Cobb. 2001. Mitogen-activated protein (MAP) kinase pathways: Regulation and physiological functions. *Endocrine Reviews* 22:153-8.

[28] Pouysségur, J., V. Volmat, and P. Lenormand. 2002. Fidelity and spatio-temporal control in MAP kinase (ERKs) signalling. *Biochemical Pharmacology* 64:755-63.

D. Faller *is a PhD student at the Freiburg Center for Data Analysis and Modeling, Freiburg, Germany.*

U. Klingmüller *is a group leader at the Max-Planck-Institute for Immunology, Freiburg, Germany.*

J. Timmer *is a group leader at the Freiburg Center for Data Analysis and Modeling, Freiburg, Germany.*