10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.2, June 4-6, 2007, Cancún, Mexico

OPTIMAL DYNAMIC EXPERIMENTAL DESIGN IN SYSTEMS BIOLOGY: APPLICATIONS IN CELL SIGNALING.

Eva Balsa-Canto * **Antonio A. Alonso** *
Julio R. Banga *

* *Process Eng. Group, IIM-CSIC. Vigo-Spain*

Abstract:
Mathematical models of complex biological systems often consist of sets of ordinary differential equations which depend on several non measurable parameters that must be estimated by fitting the model to experimental data. However this fitting can be only accomplished for the cases that practical identifiability may be guaranteed.
This work proposes an iterative optimal experimental design procedure, consisting of three main steps (identifiability analysis, ranking of parameters and the design of optimal dynamic experiments), so as to maximize identifiability, that is the ratio quantity/quality of information for model calibration. The applicability and advantages of using such procedure are illustrated by considering an example related to the modelling of a cell signaling cascade.

Keywords: model identification, optimal experimental design, global optimization, systems biology, cell signaling

1. INTRODUCTION

Since the recognition of the role of the malfunction of cell signaling pathways, particularly those involving phosphorylation cascades, in the development of diseases such as cancer, many efforts have been devoted to their mathematical modeling. The aim is to provide a systematic framework to generate hypothesis and make predictions "in silico", to get a better insight into the disease process and ultimately to identify potential drug targets (Ideker et al., 2001; Kitano, 2002).

Most models are based on viewing cellular signaling pathways as networks of biochemical reactions (Kholodenko, 2006). Such models consist of sets of non-linear ordinary differential equations that depend on several parameters (kinetic constants, initial concentration of some proteins, etc.) which are not measurable and must therefore be estimated by fitting the model to experimental data. The model calibration is performed by minimizing a cost function which quantifies the differences between model predictions and measurements. However the results will be satisfactory only if the sources of information are of a sufficiently high quality. Unfortunately, experiments in molecular biology rarely produce large and accurate data sets (Kutalik et al., 2004), thus often resulting in the impossibility of calculating unique values for the parameters.

Optimal experimental design (OED) of dynamic experiments consists of the determination of the scheme of measurements and stimuli profiles that maximize the amount and quality of information extracted from the experiemt(s), as measured by the Fisher Information Matrix, with the aim of estimating the parameters with the greatest precision and/or decorrelation (Banga et al., 2002;; Asprey and Macchietto, 2002).

Although the potential of the design of optimal dynamic experiments has been exploited in other scientific areas, this seems not to be the case in the context of systems biology, where only a few studies have recently appeared. Faller et al.(2003) made use of simulations to calculate polynomial input profiles in order to enhance parameter estimation accuracy for a MAP kinase cascade; Kutalik et al. (2004) proposed the calculation of optimal sampling times so as to reduce the variation of the parameter estimates.

This work proposes an iterative experimental design procedure which involves several steps: identifiability analysis, ranking of parameters and the rigorous solution of the optimal experimental design problem. Particular attention will be paid to the OED problem which is formulated as a general dynamic optimization problem (Banga et al., 2002) and solved using the so called control vector parameterization approach (CVP). As a result, a usually multimodal non-linear programming problem (NLP) is obtained therefore the use of global optimization methods is required.

2. OPTIMAL DYNAMIC EXPERIMENTAL DESIGN ITERATIVE PROCEDURE

Model development can be regarded as a cycle comprising a number of phases. Once the model structure has been established based on *a priori* phenomenological knowledge and hypothesis, experimental data is used to obtain a first estimate of the model unknown parameters. This task is often rather complicated, mainly due to the following reasons (Rodriguez-Fernandez et al., 2006):

- large number of parameters
- multimodality (several sub-optimal solutions)
- presence of practical identifiability problems, that is, the impossibility of calculating unique values for all parameters.

Global optimization methods, particularly stochastic global methods, have shown excellent properties in dealing with the multimodality problem even for the cases when the number of parameters is large and/or the order of magnitude is unknown, as recently illustrated by Rodriguez et al., (2006) and Egea et al., (2006).

However practical identifiability problems pose new difficulties which are hardly solvable unless an appropriate experimental scheme is used. This work proposes an iterative experimental design procedure (Figure 1) which involves (i) performing a ranking of the parameters; (ii) the computation of the correlation matrix and robust confidence intervals for the parameters so as to evaluate identifiability problems and finally, (iii) the solution of an optimal experimental design problem via dynamic optimization.

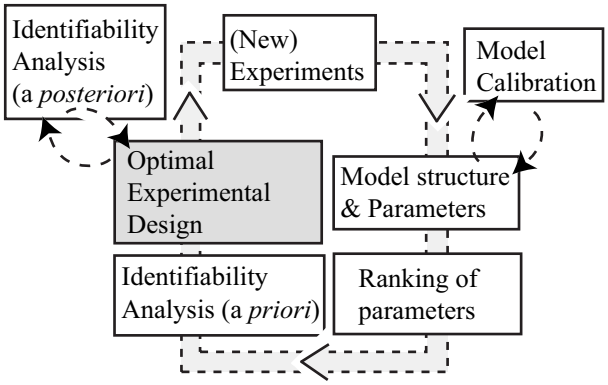

Fig. 1. Optimal dynamic experimental design iterative procedure

3. RANKING OF PARAMETERS

Let us assume a general dynamic model in ordinary differential equations:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, t) \quad (1)$$

$$\mathbf{y}(t_s^k, \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, t_s^k), \boldsymbol{\theta}, t_s^k) \quad (2)$$

where $\mathbf{x} \in X \subset \mathbb{R}^{n_x}$ are the state variables, $\mathbf{y} \in Y \subset \mathbb{R}^{n_{obs}}$ is the vector of n_s discrete time measurements t_k^s, $\mathbf{u} \in U \subset \mathbb{R}^{n_u}$ corresponds to the external factors (inputs or stimuli), $\mathbf{v} \in V \subset \mathbb{R}^{n_v}$ includes the sampling times, experiment durations and the initial conditions, and $\boldsymbol{\theta} \in \Theta \subset \Re^{n_\theta}$ is the vector of model parameters.

Local parametric sensitivities for a given experiment ix at a given sampling time t_s^k are then defined as follows:

$$S_{ij}^{ix}(t_s^k) = \frac{\partial y_i^{ix}}{\partial \theta_j}(t_s^k); i = 1 \dots n_{obs}; j = 1 \dots n_\theta \quad (3)$$

The relative sensitivities, $s_{ij}^{ix} = \frac{\triangle \theta_j}{\triangle y_i^{ix}} \frac{\partial y_i^{ix}}{\partial \theta_j}$, can then be used to asses the individual local parameter importance, that is to establish a ranking of parameters. Brun and Reichert (2001) suggested several importance factors, that may be generalized for the case of having several observables and experiments as follows:

$$\delta_j^{msqr} = \frac{1}{n_{exp} n_{obs} n_s} \sqrt{\sum_{ix=1}^{n_{exp}} \sum_{i=1}^{n_{obs}} \sum_{k=1}^{n_s} \left(s_{ij}^{ix}(t_s^k)\right)^2} \quad (4)$$

$$\delta_j^{mabs} = \frac{1}{n_{exp} n_{obs} n_s} \sum_{ix=1}^{n_{exp}} \sum_{i=1}^{n_{obs}} \sum_{k=1}^{n_s} \left| s_{ij}^{ix}(t_s^k) \right| \quad (5)$$

$$\delta_j^{mean} = \frac{1}{n_{exp} n_{obs} n_s} \sum_{ix=1}^{n_{exp}} \sum_{i=1}^{n_{obs}} \sum_{k=1}^{n_s} s_{ij}^{ix}(t_s^k) \quad (6)$$

$$\delta_j^{max} = \max_{k,i,ix} s_{ij}^{ix}(t_s^k) \quad (7)$$

$$\delta_j^{min} = \min_{k,i,ix} s_{ij}^{ix}(t_s^k) \quad (8)$$

Ranking the parameters according to these criteria, preferably in decreasing order, results in a parameter importance ranking. δ^{msqr} and δ^{mabs} quantify how sensitive is a model to a given parameter considering in δ^{mabs} interactions between parameters. δ^{max} and δ^{min} indicate the presence of outliers and provide information about the sign. δ^{mean} provides information about the sign of the averaged effect a change in a parameter has on the model output.

4. (PRACTICAL) IDENTIFIABILITY ANALYSIS

Once an experimental scheme has been established it is necessary to numerically investigate its properties. In this regard, the use of the confidence regions for the parameter estimates provides information on practical identifiability (the smaller and rounder the confidence region, the better the experimental scheme) and allows for the comparison among alternative experiments.

The most widely used approach to calculate confidence regions is based on the so called Cramér-Rao inequality (Ljung, 1999) which relates the Fisher Information Matrix with the inverse of the parameters correlation matrix. However, and due to the nonlinear nature of the models under consideration, the use of the Cramér-Rao inequality may give a wrong estimation of the real confidence regions.

This work proposes, instead, the use of a more robust Monte-Carlo based approach (Walter and Pronzato, 1997). The underlying idea is to generate different sets of noisy simulated measurements and to solve the corresponding parameter estimation problem; different experiments will lead to different identified parameters allowing to obtain a "cloud" of solutions which represents the confidence region. The results achieved are plotted by pairs of parameters, revealing identifiability problems, correlation type between parameters and degree of precision in the estimation.

5. OPTIMAL EXPERIMENTAL DESIGN

5.1 Fisher Information Matrix

Here we assume the model calibration is performed by minimizing the so called least squares function:

$$J(\boldsymbol{\theta}) = \sum_{ix=1}^{n_{exp}} \sum_{i=1}^{n_{obs}} \left[\Delta Y_i^{ix}\right]^T \mathbf{Q}_i^{ix} \left[\Delta Y_i^{ix}\right] \quad (9)$$

where $\Delta Y_i^{ix} = \tilde{\mathbf{y}}_i^{ix} - \mathbf{y}_i^{ix}(\boldsymbol{\theta})$, $\tilde{\mathbf{y}}_i^{ix} \in \tilde{Y} \subset \Re^{n_s}$ is the vector of sampling data for the observable i in the experiment ix, $\mathbf{y}_i^{ix} \in Y \subset \Re^{n_s}$ the corresponding model predictions and $\mathbf{Q}_i^{ix} \in \Omega \subset \Re^{n_s \times n_s}$ is a nonnegative definite symmetric weighting matrix related to the experimental error.

Under certain assumptions it may be concluded that the practical identifiability can be improved through the maximization of the so called the Fisher Information Matrix (FIM) (Ljung, 1999):

$$FIM = \sum_{ix=1}^{n_{exp}} \sum_{i=1}^{n_{obs}} \left[\nabla_\theta y_i^{ix}\right]^T (\boldsymbol{\theta}^*) \mathbf{Q}_i^{ix} \left[\nabla_\theta y_i^{ix}\right] (\boldsymbol{\theta}^*) \quad (10)$$

where $\left[\nabla_\theta y_i^{ix}\right]$ is the matrix ($\in \Re^{n_s \times n_\theta}$) of parametric sensitivities for the observable i in the experiment ix as calculated for a given vector of parameters $\boldsymbol{\theta}^*$ assumed to be close the "real" one.

5.2 Mathematical formulation of the OED problem

The optimal experimental design (OED) problem may be formulated as a general dynamic optimization problem as follows: *Calculate the time-varying manipulable variables (stimuli), sampling times, experiment durations and (possibly) initial conditions so as to maximize (or minimize) a scalar measure of the FIM*:

$$J_{OED} = \phi(FIM) \quad (11)$$

subject to the system dynamics as summarized in Eqns. 2 and other algebraic constraints related to experimental limitations: $\mathbf{u}^L(t) \leq \mathbf{u}(t) \leq \mathbf{u}^U(t)$ and $\mathbf{v}^L \leq \mathbf{v} \leq \mathbf{v}^U$.

The maximization of the FIM may be achieved through the definition of suitable cost functionals J_{OED} (Vanrolleghem and Dochain, 1998). The most popular are:

- *D-optimality*: $max\ \phi_D = det(FIM)$
- *E-optimality*: $max\ \phi_E = \lambda_{min}(FIM)$
- *Modified E-optimality*: $min\ \phi_\varepsilon = \frac{\lambda_{max}(FIM)}{\lambda_{min}(FIM)}$

The following interpretation can be given to each of these criteria: *D-optimality* designs result in the smallest volume of the confidence region in the parameter space and indicates the quantity of information provided by the experiments. *E-optimality* intends to minimize the maximum error on the parameter estimates. And *Modified E-optimality* regards the relationship between the maximum and minimum error, the closer its value to one, the more homogeneous the distribution of the information among the parameters so the maximum

decorrelation among them. Note however that whereas *D-* and *E-optimality* tend to minimize the size of the confidence hyper-ellipsoid, *Modified E-optimality* tends to make it rounder regardless the size.

Every FIM based criteria may lead to different experimental designs and without no extra information it will be impossible to decide which will be the most convenient. Here we propose to use a practical identifiability analysis to compare the properties of the different designs.

5.3 Numerical techniques

The most widely used approaches to solve dynamic optimization problems, as recently reviewed by Banga et al. (2005), are based on the transformation of the original infinite dimension optimization problem into a nonlinear programming problem (NLP). From the different alternatives, the control vector parameterization (CVP) approach is selected here as it allows for the design of a number of simultaneous experiments with several inputs and for the general case of large scale models without solving excessively large nonlinear optimization problems.

The CVP method proceeds dividing the duration of the experiments into a number of elements and approximating the stimuli profiles using low order polynomials. The linear or constant approximations are the most convenient since they can be implemented in practice, meeting experimental constraints.

As a result, a nonlinear optimization problem must be solved with an initial value problem embedded. The decision variables the polynomial coefficients, the experimental durations, the sampling times and the initial conditions. The evaluation of the FIM dependent cost function requires the simulation of the system dynamics and the calculation of the parametric sensitivities, computed here by means of ODESSA (Leis and Kramer, 1988).

Local and global optimization methods

The NLPs arising from the application of the CVP method are frequently multimodal (presenting multiple local optima) (Banga et al., 2003). Therefore, deterministic (gradient based) local optimization techniques may converge to local optima, especially if they are started far away from the global solution. In order to surmount these difficulties, global optimization methods must be used.

In this regard, stochastic GO methods have been successfully applied to solve nonlinear dynamic optimization problems (Banga and Seider, 1996; Banga et al., 2005), being therefore good candidates for solving optimal experimental design problems. Note that this type of approaches can not guarantee global optimality, but they may approach its vicinity (sometimes the best known solution) with relative efficiency.

This work makes use of a population based method, Differential Evolution (DE, Storn and Price, 1997) due to its demonstrated robustness in the solution of a collection of nonlinear optimization problems.

6. ILLUSTRATIVE EXAMPLE: A MAP KINASE SIGNALING PATHWAY

MAP kinase family members have been found to regulate diverse biological functions by phosphorylation of specific target molecules (such as transcription factors, other kinases, etc.) found in cell membrane, cytoplasm and nucleus. We consider here the so called ERK module, that involves the activation of ERK via MEK.

6.1 Mathematical model

The last step of the signaling cascade is represented as a biochemical reaction network (as in Faller et al., 2003). The application of the mass action law to each of those reactions result in the following set of non-linear ordinary differential equations:

$$\begin{aligned}
\dot{x}_1 &= c_1 * x_3 - a_3 * x_1 * u + b_3 * x_4 + \\
&\quad c_4 * x_6 - a_2 * x_1 * P + b_2 * x_5 \\
\dot{x}_2 &= c_3 * x_3 - a_4 * x_2 * P + b_4 * x_6 \\
\dot{x}_3 &= a_1 * E * u - (b_1 + c_1) * x_3 \\
\dot{x}_4 &= a_3 * x_1 * u - (b_3 + c_3) * x_4 \\
\dot{x}_5 &= a_2 * x_1 * P - (b_2 + c_2) * x_5 \\
\dot{x}_6 &= a_3 * x_2 * P - (b_4 + c_4) * x_6
\end{aligned} \tag{12}$$

where x_i for $i = 1 \dots 6$, stands for the concentrations of ERK^*, ERK^{**}, $ERK{-}MEK^*$, $ERK^*{-}MEK^{**}$, $ERK^* - P$ and $Erk^{**} - P$, respectively. P regards the phosphatase ($P = P_{tot} - x_5 - x_6$) and E the kinase ERK ($E = E_{tot} - x_1 - x_2$) concentrations. The parameters a_i denote the rates at which the substrate binds to the enzyme, b_i denote the corresponding breaking rates, and c_i denote the rate at which the actual activation reaction occurs. The initial concentrations $x_i(t_0)$ of all phosphorylated Erks and complexes of phosphorilated Erks with Meks or phosphatases are zero. The stimulus u corresponds to $MEK^{**}(t)$ verifying $2 \leq u(t) \leq 12$ (arbitrary units) and the observables are the activated ERK^{**} and the

total ERK. The nominal values for the parameters and total quantities are: $a_i = 0.5$, $b_i = 0.6$, $c_i = 0.9\ \forall i = 1, \ldots, 4$; $Erk_{tot} = 50$ and $P_{tot} = 20$.

6.2 Ranking of parameters

In order to check which are the most relevant parameters in the model a ranking was performed using a number of different constant stimulus experiments over the accepted range for u.

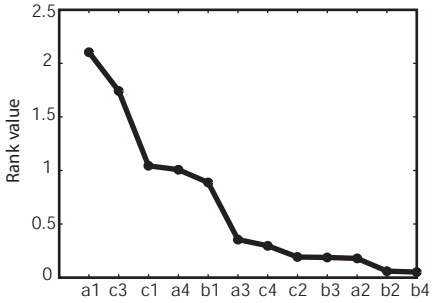

Fig. 2. Ranking of parameters. Ordered by decreasing δ_{msqr}.

The model is specially sensitive to a_1, c_2, c_1, a_4 and b_1. If a structural identifiability analysis is performed it is concluded that from the given observables it is impossible to simultaneously estimate a_1 and all the other parameters. Therefore we will assume a_1 known and perform the optimal experimental design for the remaining most important parameters c_2, c_1, a_4 and b_1.

6.3 Practical identifiability analysis for a typical experiment

In practice the experiments are usually performed under constant stimulus. What would happen if we try to estimate c_3, c_1, a_4 and b_1 from such an experiment? The practical identifiability analysis may help to answer this question. A constant stimulus of $u = 4$, 20 equidistant sampling times and 10% Gaussian noise in the experiments were used. The corresponding correlation matrix is shown in Figure 3.

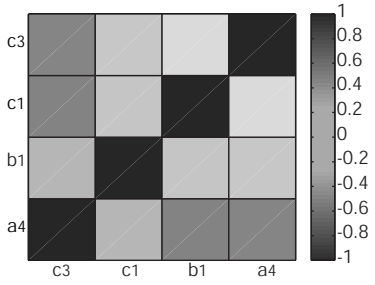

Fig. 3. Correlation matrix for a typical constant stimulus experiment.

The parameters are not highly correlated but the robust confidence ellipsoids reveal errors up to 85% when trying to estimate their values.

6.4 Design of optimal dynamic experiments

Based on conversations with experimental biologists the following constraints were imposed for the optimal experimental design problem:

- Two experiments with two steps each.
- The duration of the experiment is free: $30 \leq t_f^{ix} \leq 90\ min.$
- 20 equidistant or 15 optimal sampling times.
- 10% Gaussian noise.
- As the parameters are not highly correlated, *D-optimality* criterion is chosen so as to minimize the size of the confidence regions.

In order to check for the multimodality of the optimization problem, a multistart of a local method was used. Results are illustrated in Figure 4.

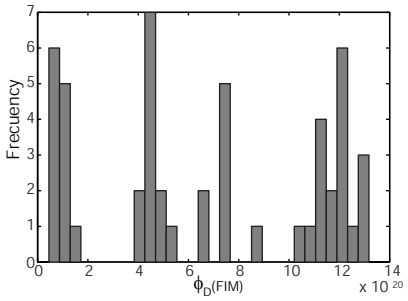

Fig. 4. Multistart of a local NLP solver.

From this Figure it becomes apparent the presence of multiple suboptimal solutions, therefore the use of a global optimization method is required.

The use of the population based method DE (Storn and Price, 1997) lead to the optimal experimental design in Figures 5 and 6.

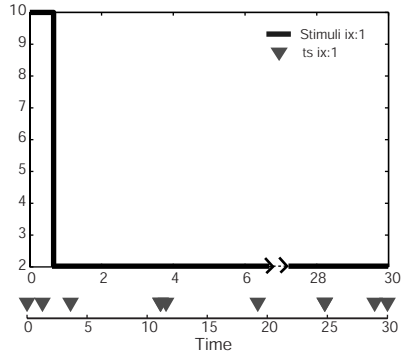

Fig. 5. Optimal (dynamic) experiment 1.

The comparison of the confidence regions for the constant stimulus case (Exp. Scheme (a)) with the two optimal dynamic experiments with equidistant (Exp. Scheme (b)) and optimally located sampling times (Exp. Scheme (c)), reveals that,

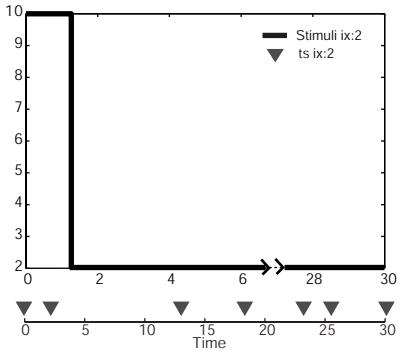

Fig. 6. Optimal (dynamic) experiment 2.

even for the worst case, OED may largely improve results, as shown in Figure 7. Remark that the maximum predicted errors for the Schemes (b) and (c) correspond to values of 30% and 16% respectively.

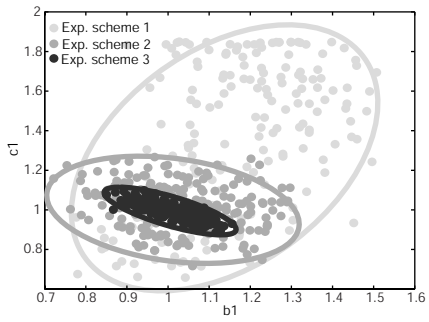

Fig. 7. Comparison of confidence intervals for 3 experimental schemes (*worst* case).

7. CONCLUSIONS

Reliable model calibration in systems biology largely depends on the quantity and quality of the experimental data. This work proposes the use of an iterative procedure based on the computation of a ranking of parameters, identifiability analysis and optimal dynamic experimental designs with the aim of maximizing practical identifiability.

The results obtained for a simple signaling pathway clearly indicate that dynamic experiments combined with optimal sampling times yield better results than the classical experiments using constant stimulus and equidistant measurements.

Acknowledgments

This work was supported by the European Community as part of the FP6 COSBICS Project (STREP FP6-512060) and by Xunta de Galicia (PGIDIT05PXIC40201PM).

REFERENCES

S.P. Asprey and S. Macchietto. Designing robust optimal dynamic experiments. *J. Process Control*, 12:545–556, 2002.

J. R. Banga, C.G. Moles, and A. A. Alonso. *Global optimization of bioprocesses using stochastic and hybrid methods.*, pages 45–70. Frontiers In Global Optimization. C.A. Floudas and P. M. Pardalos, (Eds.), Kluwer Acad. Pub. 2003.

J. R. Banga, E. Balsa-Canto, C.G. Moles, and A. A. Alonso. Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. *J. of Biotechnology*, 117:407–419, 2005.

J.R. Banga, K. J. Versyck, and J.F. Van Impe. Computation of optimal identification experiments for nonlinear dynamic process models: an stochastic global optimization approach. *Ind. & Eng. Chem. Res.*, 41:2425–2430, 2002.

R. Brun and P. Reichert. Practical identifiability analysis of large environmental simulation models. *Water Resources Res.*, 37:1015–1030, 2001.

J. A. Egea, M. Rodriguez-Fernandez, J. R. Banga, and R. Marti. Scatter search for chemical and bio-process optimization. *J. Glob. Opt.*, DOI:10.1007/s10898-006-9075-3, 2006.

D. Faller, U. Klingmüller, and J. Timmer. Simulation methods for optimal experimental design in systems biology. *Simulation*, 79:717–725, 2003.

T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2:343–372, 2001.

B.N. Kholodenko. Cell-singalling dynamics in time and space. *Nature Reviews, Molecular Cell Biology*, 7:165–176, 2006.

H. Kitano. Systems biology: A brief overview. *Science*, 295:1662–1664, 2002.

Z. Kutalik, K-H. Cho, and O. Wolkenhauer. Optimal sampling time selection for parameter estimation in dynamic pathway modelling. *BioSystems*, (75):43–55, 2004.

J.R. Leis and M.A. Kramer. Odessa- an ordinary differential-equation solver with explicit simultaneous sensitivity analysis. *ACM Trans. Math. Soft.*, 14:61–67, 1988.

L. Ljung. *System identification: Theory for the user*. Prentice Hall, New Jersey, 1999.

M. Rodriguez-Fernandez, P. Mendes, and J.R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2-3):24, 2006.

R. Storn and K. Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J. Global. Optim.*, 11:341–359, 1997.

P.A. Vanrolleghem and D. Dochain. *Bioprocess model identification. Advanced instrumentation, data interpretation, and control of biotechnological process*, pages 251–318. Kluwer Acad. Pub., 1998.

E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Springer, Masson, 1997.