# Data- and model-based identification of biochemical processes

Dissertation
zur
Erlangung des Grades
Doktor-Ingenieur

der
Fakultät für Maschinenbau
der Ruhr-Universität Bochum

von
Tom Quaiser
aus Redhill, UK

Bochum 2011

Dissertation eingereicht am: 21. November 2011
Tag der mündlichen Prüfung: 10. Februar 2012

Erster Referent: Universitätsprofessor Dr.-Ing. M. Mönnigmann
Zweiter Referent: Universitätsprofessor Dr.-Ing. W. Marquardt
Dritter Referent: Universitätsprofessor Dr. rer. nat. F. Schaper

Schriftenreihe des Lehrstuhls für Regelungstechnik und Systemtheorie

**Tom Quaiser**

# Data- and model-based identification of biochemical processes

Shaker Verlag
Aachen 2012

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Bochum, Univ., Diss., 2012

Copyright Shaker Verlag 2012
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-1018-3
ISSN 2195-0113

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

# Vorwort

Die vorliegende Arbeit entstand während meiner Zeit als wissenschaftlicher Mitarbeiter bei der Aachener Verfahrenstechnik - Prozesstechnik der RWTH Aachen und am Lehrstuhl für Regelungstechnik und Systemtheorie an der Ruhr-Universität Bochum. Sie wurde zu wesentlichen Teilen von der DFG finanziert.

Ganz herzlich möchte ich mich bei Prof. Dr.-Ing. Mönnigmann für die Betreuung und Begutachtung dieser Arbeit bedanken. Neben der hervorragenden wissenschaftlichen Betreuung möchte ich mich besonders dafür bedanken, dass er es mir durch Fernbetreuung ermöglicht hat bei der Aachener Verfahrenstechnik zu bleiben. Hiermit hat er für mich einen sehr großen Anteil zur Vereinbarkeit von Forschung und Familie geleistet.

Obwohl ich "virtuell" schon an diversen Universitäten beschäftigt war, ist meine Heimat immer die Aachener Verfahrenstechnik - Prozesstechnik geblieben. Daher gilt mein herzlicher Dank Prof. Dr.-Ing. Marquardt, der mir an seinem Lehrstuhl ein Zuhause gegeben hat. Die wissenschaftliche Zusammenarbeit und die menschliche Stimmung am Lehrstuhl waren immer von freundschaftlichem Miteinander geprägt. Ganz besonders möchte ich meinen Bürokollegen Claas Michalik und Fady Assassa danken. Neben Diskussionen über fachlichen Themen, blieb mit beiden Kollegen Zeit für viele andere Themen, humoristische Abschweifungen, Fitnesstraining und Obstsalat. Auch möchte ich mich bei Ralf Hannemann bedanken, der meinen Start am Lehrstuhl begleitet hat, und der eine Seele von Mensch ist.

Prof. Dr. Schaper möchte für das Begutachten meiner Arbeit danken. Mein besonderer Dank gilt auch meiner Kooperationspartnerin Anna Dittrich für die hervorragende fachliche und menschliche Zusammenarbeit. Ebenfalls möchte ich meinen Bochumer Kollegen danken. Hier vor allem Martin Kastsian für das "Bändigen" des Rechenclusters.

Nicht genug kann ich meiner Familie danken. Meiner Frau für die Unterstützung und Liebe. Meinen Kindern, die mir jeden Tag zeigen, was wirklich wichtig ist. Meinen Eltern, die immer an mich glauben. Meinem Vater für das Korrekturlesen. Und meinen Schwiegereltern, die immer für uns da waren.

# Contents

# Notation

## Abbreviations

| | |
|---|---|
| AB | antibody |
| A20 | zink-finger protein |
| AD | automatic differentiation |
| AIC | Akaike information criterion |
| AICc | Akaike information criterion adapted for few samples |
| BDF | backward differentiation formula |
| DNA | deoxyribonucleic acid |
| EGF | epidermal growth factor |
| ELISA | enzyme-linked immunosorbent assay |
| FIM | Fisher information matrix |
| gp80 | glycoprotein 80 |
| gp130 | glycoprotein 130 |
| IFN | interferon-$\gamma$ |
| I$\kappa$B$\alpha$ | inhibitor of NF-$\kappa$B |
| IKK | I$\kappa$B kinase |
| IL-6 | interleukin 6 |
| IL-10 | interleukin 10 |
| IL-13 | interleukin 13 |
| IP | immunoprecipitation |
| JAK | Janus kinase |
| pJAK | phosphorylated JAK |
| LHS | Latin hypercube sampling |
| LMA | law of mass action |
| MAP | mitogen-activated protein |
| MLE | maximum likelihood estimate |
| MPI | message passing interface |

| | |
|---|---|
| N-SH2 | N-terminal SH2 domain of SHP2 |
| C-SH2 | C-terminal SH2 domain of SHP2 |
| NF-$\kappa$B | nuclear factor-$\kappa$B |
| ODE | ordinary differential equation |
| OED | optimal experimental design |
| PCA | principal component analysis |
| PDE | partial differential equation |
| POI | protein of interest |
| PPX | cytoplasmic phosphatase |
| pSHP2 | phosphorylated SHP2 |
| pSTAT | phosphorylated STAT |
| PTP | protein tyrosine phosphatase domain |
| PTPN11 | SHP2 encoding gene |
| R | receptor |
| RNA | ribonucleic acid |
| SDS | sodium dodecyl sulfate |
| SH2 | Src homology 2 |
| SHC | SH2 domain-containing protein |
| SHP2 | SH2 domain-containing tyrosine phosphatase 2 |
| siRNA | small interfering RNA |
| SOCS | suppressor of cytokine signaling |
| SQP | sequential quadratic programming |
| STAT | signal transducer and activator of transcription |
| STAT1 | IFN-$\gamma$-activated STAT1 |
| STAT1c | cytoplasmic STAT1 |
| TGF-$\alpha$ | transforming growth factor $\alpha$ |
| TNF | tumor necrosis factor |
| WB | Western blot |
| WSSR | weighted sum of squared residuals |
| Y | one letter code for the amino acid tyrosine |

## Greek letters

| | |
|---|---|
| $\delta_{\epsilon_c}$ | amount by which $\epsilon_c$ is incremented in the correlation algorithm |
| $\Delta_k$ | AICc difference |
| $\epsilon_{\chi^2}$ | threshold for the $\epsilon$ uncertainty region around an MLE parameter |
| $\epsilon$ | cut-off value for the eigenvalue method |

| | |
|---|---|
| $\epsilon_c$ | cut-off value for the correlation method |
| $\epsilon_o$ | cut-off value for the orthogonal method |
| $\lambda^i$ | $i$th eigenvalue of $H$ |
| $\lambda_{\min}$ | smallest eigenvalue of $H$ |
| $\tilde{\lambda}^{i,j}$ | $j$th eigenvalue of $\tilde{S}^{iT}\tilde{S}^i$ |
| $\sigma_{ij}$ | standard deviation of $\tilde{y}_i(t_j)$ |
| $\sigma_{h,i,j}$ | standard deviation of $\tilde{y}_{h,i}(t_j)$ |
| $\sigma(a)$ | standard deviation of values in vector $a$ |
| $\sigma^2(p_i)$ | variance of parameter $p_i$ |
| $\nu$ | degrees of freedom of maximum likelihood estimation |
| $\phi(p)$ | sum of squares function |
| $\chi^2$ | $\chi^2$ function |

## Latin letters

| | |
|---|---|
| $\bar{a}$ | arithmetic mean of $a$ |
| $C$ | matrix of absolute correlation values above the threshold $1 - \epsilon_c$ |
| $corr(a, b)$ | correlation between two vectors $a$ and $b$ |
| $corr^\star(a, b)$ | tailor made correlation function used in the correlation method |
| $c_i^{tot}(K)$ | sum of correlations of parameter $p_i$ to all $p_j$ with $j \neq i$ and $j \in K$ |
| $cov(a, b)$ | covariance between two vectors $a$ and $b$ |
| $H$ | Hessian matrix |
| $H_0$ | null hypothesis |
| $H_A$ | alternative hypothesis |
| $I$ | index set of identifiable parameters |
| $J$ | final parameter ranking of the PCA-based method |
| $K$ | index set used in the correlation method |
| $L$ | likelihood function |
| $L_i$ | index set of parameters used in the $i$th iteration of the PCA-based and orthogonal method |
| $M$ | model |
| $M^k$ | model created in the $k$th iteration of the simplification work flow |
| $M^k_{\text{SHP2}}$ | SHP2 model resulting from iteration $k$ of the model identification cylce |

| | |
|---|---|
| $M^{5,\text{pSHP2+}}_{\text{SHP2}}$ | $M^5_{\text{SHP2}}$ extended for positive feedback by phosphorylated SHP2 |
| $M^{5,\text{actSTAT+}}_{\text{SHP2}}$ | $M^5_{\text{SHP2}}$ extended for positive feedback by active STAT |
| $M^{5,\text{actR+}}_{\text{SHP2}}$ | $M^5_{\text{SHP2}}$ extended for positive feedback by active IL-6-bound receptor |

| | |
|---|---|
| $M_{\text{SHP2}}^{6,\text{actSTAT+}}$ | $M_{\text{SHP2}}^{6}$ after exchange of positive pSHP2 against positive actSTAT feedback |
| $M_{\text{SHP2}}^{6,\text{actR+}}$ | $M_{\text{SHP2}}^{6}$ after exchange of positive pSHP2 against positive active IL-6-bound receptor feedback |
| $N$ | number of starting points for multi-start parameter estimation |
| $n$ | number of measurement reproductions |
| $N_q$ | index set of unidentifiable parameters in iteration $q$ of the PCA-based method |
| $n_{\text{accept}}$ | number of accepted estimates in a multi-start estimation run |
| $n_{\text{exp}}$ | number of experiments |
| $n_p$ | number of parameters |
| $n_t$ | number of measurement times |
| $n_y$ | number of model outputs |
| $n_x$ | number of state variables |
| $n_I$ | number of elements in the index set $I$ of identifiable parameters |
| $n_J$ | number of elements of parameter ranking $J$ created by the principal component analysis based method |
| $P(\chi^2\|\nu)$ | probability for the value of the $\chi^2$-distribution with $\nu = n_y n_t - n_p$ degrees of freedom to be less than an observed $\chi^2$-value |
| $p$ | vector of model parameters |
| $\hat{p}$ | vector of MLEs of $p$ |
| $p^\star$ | vector of true parameters |
| $p', p''$ | vector of nominal parameters |
| $p^{\text{lit}}$ | vector of parameter values taken from the literature |
| $\hat{p}^k$ | MLE of iteration $k$ of the simplification work flow |
| $P_q$ | orthogonal projection used in iteration $q$ of the orthogonal method |
| $Q$ | matrix containing the acceptable estimates of a multi-start estimation |
| $R(\Delta p)$ | sum of squared errors between linearized and regular model output |
| $R_j^i$ | parameter ranking created by the PCA-based method with criterion $j$ for model output $y_i$ |
| $R_{j,k}^i$ | parameter index at position $k$ of the PCA-based ranking $R_j^i$ |
| $S$ | sensitivity matrix |
| $\tilde{S}^i$ | truncated sensitivity matrix used in the PCA-based method |
| $S_{\cdot i}$ | $i$th column of $S$ |
| $S_{\cdot k}^{proj}$ | orthogonal projection of $S_{\cdot k}$ onto $V$ |
| $S_{\cdot k}^{\perp}$ | perpendicular connection between $V$ and $S_{\cdot k}$ |
| $s(t_i)$ | submatrix of matrix $S$ for measurement time $t_i$ |
| $s_{ij}(t_k)$ | sensitivity coefficient for parameter $p_j$ with respect to $y_i$ at $t_k$ |

| | |
|---|---|
| $t$ | time |
| $t_i$ | $i$th measurement time |
| $T$ | test statistic |
| $\delta t$ | time interval between measurements |
| $u$ | vector of time variant model input |
| $u^i$ | $i$th eigenvector of $H$ |
| $\tilde{u}^{i,j}$ | $j$th eigenvector of $\tilde{S}^{iT}\tilde{S}^i$ |
| $U$ | set of indices of unidentifiable parameters |
| $v$ | reaction rate |
| $v(p_i)$ | coefficients of variation of parameter $p_i$ |
| $\bar{v}$ | cut-off value for variance-based identifiability testing |
| $V$ | vector space spanned by the columns of $X_q$ in the orthogonal method |
| $w_k$ | AICc weights |
| $W$ | inverse of the measurement variance matrix |
| $W_j^i(q)$ | set of the first $q$ elements of $R_j^i$ |
| $W^{ortho}$ | ranking of parameter indices created by the orthogonal method |
| $x$ | vector of state variables |
| $x_0$ | vector of initial values of state variables |
| $X^q$ | collection of columns $S_{\cdot i}$ corresponding to identifiable parameters in iteration $q$ of the orthogonal method |
| $y$ | vector of model outputs |
| $y_i(t_j)$ | $i$th model output at time $t_j$ |
| $y_{h,i}(t_j)$ | $i$th model output of experiment $h$ at time $t_j$ |
| $\tilde{y}_i(t_j)$ | measurement corresponding to $y_i(t_j)$ |
| $\tilde{y}_{h,i}(t_j)$ | measurement corresponding to $y_{h,i}(t_j)$ |
| $\tilde{y}$ | vector of measurements |

## Calligraphic letters

| | |
|---|---|
| $\mathbb{D}$ | space of feasible experimental designs |
| $\mathcal{D}$ | a particular experimental design |
| $\mathcal{D}^\star$ | an optimal experimental design |
| $\mathcal{M}()$ | metric to quantify the information content of an experimental design |
| $\mathcal{U}$ | space of feasible parameters |
| $\mathcal{V}(p)$ | neighborhood of parameter $p$ |

# Abstract

In the last decade a paradigm shift has taken place in biochemical research: while traditionally biochemical processes have often been studied on a qualitative level, more and more research now focuses on quantitative time-resolved aspects of biochemical processes. However, on a quantitative dynamic level the complexity of these processes increases significantly and the need of mathematical models arises. Once a model is fitted to experimental data it can be used to simulate and study the dynamic behavior of a given process. Furthermore, a fitted model allows it to test new experiments and hypothesis in silico before time and cost intensive real experiments need to be conducted. The interplay between biochemical experimentation and mathematical modeling - known as systems biology - is an integral part of this thesis.

Identifying a predictive model starts with the formulation of an initial model, which combines a priori knowledge with new to be tested hypotheses. The initial model is refined in an iterative process of performing quantitative experiments, estimating unknown model parameters, model validation and hypothesis testing. When constructing a model, it is tempting to incorporate all known interactions between biochemical species, which results in models with a large number of unknown parameters, which subsequently have to be estimated from experimental data. However, parameter estimation can only provide valid results, if the complexity of the model and the amount and quality of data are in balance with one another. If this is the case the model is said to be identifiable for the given data. In Chapter 2 of this thesis we describe a new automatic approach to test the identifiability of model parameters. We compare our new method - the eigenvalue method - to three well established methods for identifiability testing. For three published models of signaling cascades our eigenvalue methods outperforms the other methods in terms of efficiency and effectiveness. Furthermore, we find that even when assuming abundant and noise-free measurement data, the three models are not identifiable.

If a model turns out to be unidentifiable, two steps can be taken. Either additional experiments need to be conducted to increase the information content of the

data, or the model has to be simplified. In Chapter 3 we follow the latter path and describe an iterative approach that combines multi-start parameter estimation, identifiability testing, sampling-based variance analysis and goodness-of-fit testing into a work flow for model simplification. We demonstrate the effectiveness of this work flow by simplifying a published model of a signaling cascade under the assumption of realistic measurements until a good fitting model with identifiable and barely varying parameters results.

Finally, in Chapter 3 we demonstrate the power of a data-driven model-based approach for process identification by discriminating between different hypotheses on the function of SHP2 in the early phase of JAK-STAT signaling. Furthermore, we identify key processes that are essential for the dynamics of early pathway activation. In addition to the techniques presented in Chapters 1 and 2 we apply a brute-force method for optimal experimental design to propose new informative experiments. Using an initial and the optimal designed data, we iteratively refine our model until an identifiable and predictive model of early JAK-STAT signaling results that adequately describes the data.