International Journal of Bifurcation and Chaos, Vol. 14, No. 6 (2004) 2069–2079

MODELING THE NONLINEAR DYNAMICS OF CELLULAR SIGNAL TRANSDUCTION

J. TIMMER* and T. G. MÜLLER
Freiburger Zentrum für Datenanalyse und Modellbildung, Eckerstr. 1, 79104 Freiburg, Germany
Fakultät für Physik, Hermann – Herder Str. 3, 79104 Freiburg, Germany
**jeti@fdm.uni-freiburg.de*

I. SWAMEYE, O. SANDRA and U. KLINGMÜLLER
Max-Planck-Institut für Immunbiologie, Stübeweg 51, 79108 Freiburg, Germany

Received March 13, 2002; Revised April 5, 2002

During the past decades the components involved in cellular signal transduction from membrane receptors to gene activation in the nucleus have been studied in detail. Based on the qualitative biochemical knowledge, signalling pathways are drawn as static graphical schemes. However, the dynamics and control of information processing through signalling cascades is not understood. Here we show that based on time resolved measurements it is possible to quantitatively model the nonlinear dynamics of signal transduction. To select an appropriate model we performed parameter estimation by maximum likelihood and statistical testing. We apply this approach to the JAK-STAT signalling pathway that was believed to represent a feed-forward cascade. We show by comparison of different models that this hypothesis is insufficient to explain the experimental data and suggest a new model including a delayed feedback.

Keywords: Modeling; signal transduction; nonlinear dynamics; model selection.

1. Introduction

Deciphering the genome of an organism is only the very first step towards understanding the mechanisms in living cells. A more detailed picture is required to understand the regulatory properties of metabolic, genetic and cellular signalling networks. These networks are characterized by their dynamic behavior. Nonetheless, the extensive biochemical knowledge about these systems is predominantly represented in a static and qualitative manner by drawing arrows connecting interacting components of the network, see e.g. [KEGG] and Fig. 1. However, as stated in a recent *Editorial* in *Nature*: "But, to really understand the biochemical networks thus represented, one needs to have numbers attached to the arrows" [Campbell, 1999], for similar claims, see also [Koshland, Jr., 1998; Zheng & Flavel, 2000; Endy & Brent, 2001; Editorial, 2000; Downward, 2001].

A first step in this direction of analyzing the dynamics is the simulation of these networks. Therefore, the qualitative scheme is translated into a set of parameterized differential equations. Choosing the parameters is a difficult task and they are

*Author for correspondence.

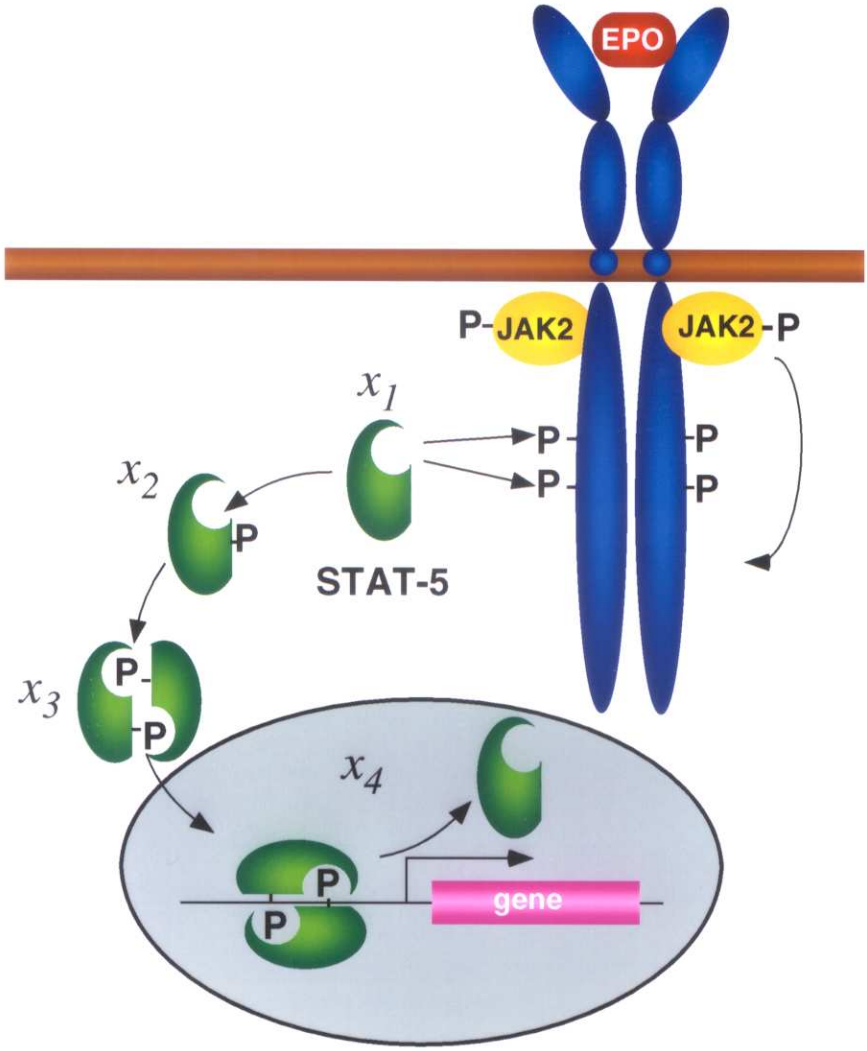


Fig. 1. Graphical representation of the JAK-STAT pathway of the Epo-receptor.

usually taken from the literature, see e.g. [Kholodenko *et al.*, 1999; Kremling & Gilles, 2000]. Unfortunately, biochemical parameters can differ by orders of magnitude depending on the experimental conditions. Therefore, this approach faces the *simulation dilemma* [Timmer *et al.*, 2000], i.e. it is difficult to decide whether discrepancies between simulated and measured data result from inadequate parameters or from an insufficient model. Yet, this approach has been applied to investigate genetic networks [Loomis & Sternberg, 1995; McAdams & Shapiro, 1995], robustness of chemotaxis [Barkai & Leibler, 1997; Alon *et al.*, 1999] and the segment polarity network [von Dassow *et al.*, 2000], short-term signaling of the epidermal growth factor receptor [Kholodenko *et al.*, 1999], optimality of metabolic networks [Edwards *et al.*, 2001] stem cell overproducing in colon cancer [Boman *et al.*, 2001], apoptosis [Fussenegger *et al.*, 2000], emergent properties of signal pathways [Bhalla & Iyengar, 1999] and whole cell behavior [Tomita *et al.*, 1999]. Usually, dependence of the result with respect to the chosen parameters is investigated by sensitivity analysis.

To solve the simulation dilemma we follow a different approach by estimating the parameters from experimental data. Having optimized the parameters of the model, discrepancies between the measured and the simulated data allow for the conclusion that the model is not sufficient, i.e. the biochemical concept the mathematical model was based upon has to be reconsidered.

We apply this data-based approach by modeling the dynamics of the JAK-STAT signalling path-

way. Translating the believed graphical representation of this pathway, a feed-forward cascade, in a set of coupled differential equations, and fitting the parameters revealed that this model is insufficient to explain the measured data. Therefore we propose a generalized model. We validate this model with independent measurements.

The paper is organized as follows: In the next section we briefly describe the investigated pathway and the acquisition of data. In Sec. 3, the methods of parameter estimation, model selection and identifiability are discussed. The application of the methods to infer a dynamical model of the JAK-STAT signalling pathway is given in Sec. 4.

2. The JAK-STAT Pathway and Data Acquisition

Functioning cells originate from undifferentiated progenitor cells. Differentiation of progenitor cells is triggered by hormones. Erythropoietin (Epo) is the hormone that promotes the development of progenitor cells to red blood cells.[1]

Upon binding of Epo to cell surface receptors, multiple signalling pathways transduce the signal to the nucleus where the respective genes are activated. Here, we concentrate on the JAK-STAT pathway, see Fig. 1; for a more detailed description, see [Darnell, Jr., 1997; Pellegrini & Dusanter-Fourt, 1997].

Binding of Epo to the extracellular part of the receptor leads to activation by phosphorylation of the so-called Janus kinase (JAK) at intracellular, cytoplasmic domain of the receptor. In turn, this leads to phosphorylation of monomeric STAT-5, a member of the STAT (signal transduction and activator of transcription) family of transcription factors. The phosphorylated monomeric STAT-5 forms dimers and these dimers migrate into the nucleus where they bind to promotor region of the DNA and initiate gene transcription. It was believed that the active role of STAT-5 ends in the nucleus by dedimerization, dephosphorylation and export to the cytoplasm where it is eventually degradated [Haspel *et al.*, 1996; Haspel & Darnell, Jr., 1999]. Thus, the JAK-STAT signalling pathway represents a feed-forward cascade. Its graphical representation is given in Fig. 1.

[1]Therefore Epo serves as a doping substance for athletes.

Biochemically, the time courses of the activation of the Epo-receptor, the phosphorylated (momomeric and dimeric) STAT-5 in the cytoplasm and the total amount of STAT-5 in the cytoplasm were determined. The measured values represent

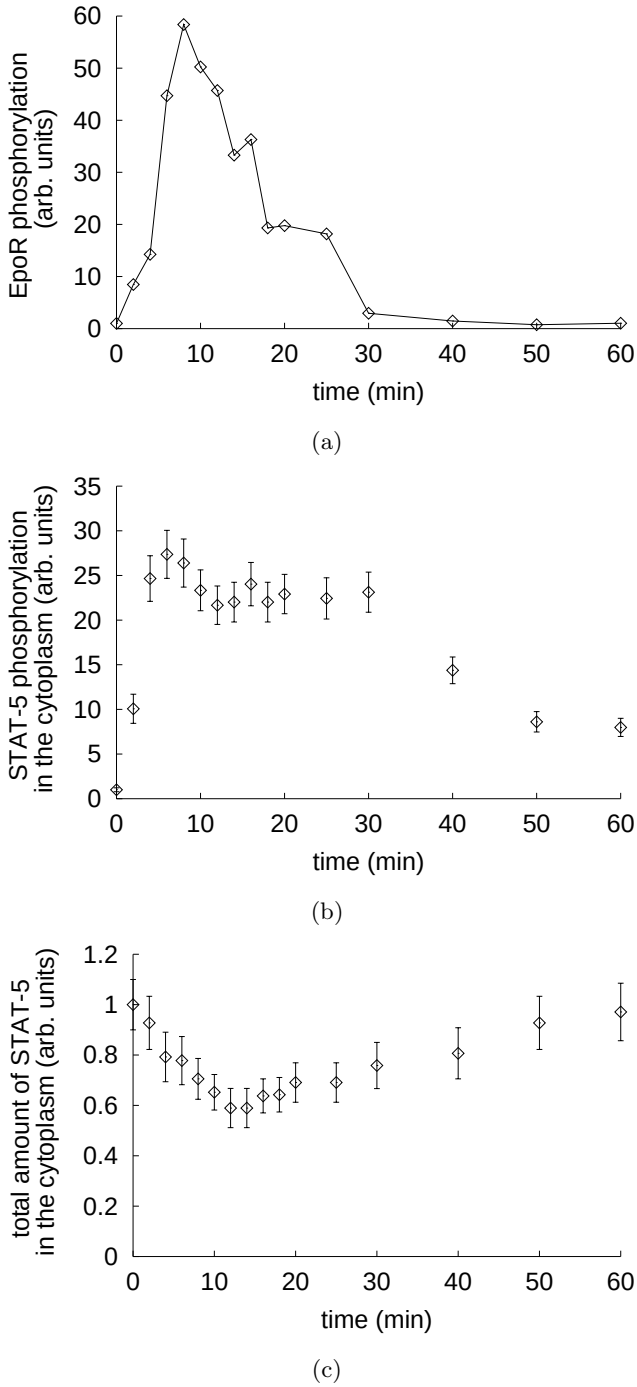


Fig. 2. Examples of the measured time series. (a) Activation of the Epo-receptor. (b) Phosphorylated STAT-5 in the cytoplasm. (c) Total amount of STAT-5 in the cytoplasm.

relative units. Unfortunately, it is currently not possible to simultaneously quantify STAT-5 in the nucleus. For a detailed description of the biochemical techniques to measure the different components, see [Swameye *et al.*, 2003].

Figure 2 displays (a) the time courses of Epo-receptor activation, (b) phosphorylated STAT-5 in the cytoplasm and (c) the total amount of STAT-5 in the cytoplasm for one representative experiment. In the time series of phosphorylated STAT-5 a plateau is reproducibly detected between 10 and 30 min.

3. Methods

In this section we discuss the topics of parameter estimation, identifiability and model selection that will be applied in Sec. 4.

To derive a dynamical model for the pathway under consideration, one has to first decide between a discrete or continuous state space. Then one has to chose between a dynamically deterministic or stochastic model. From the measured data and the underlying nature of a chemical reaction we conclude that a continuous state space is the adequate description. Chemical reactions are intrinsically stochastic. But, since each cell comprises in the order of 10^4 STAT-5 molecules, the deterministic limit should be reached. Therefore, we aim to model this system by a deterministic differential equation.

To render the following discussion concerning parameter estimation and identifiability not too abstract, we exemplify it for the model that corresponds to Fig. 1: Assuming mass-action kinetics and denoting the amount of activated Epo-receptors by EpoR_A, unphosphorylated monomeric STAT-5 by x_1, phosphorylated monomeric STAT-5 by x_2, phosphorylated dimeric STAT-5 in the cytoplasm by x_3 and phosphorylated dimeric STAT-5 in the nucleus by x_4, we arrive at the following model (model 1):

$$\dot{x}_1 = -k_1 x_1 \,\mathrm{Epo}\, R_A \tag{1}$$

$$\dot{x}_2 = +k_1 x_1 \,\mathrm{Epo}\, R_A - k_2 x_2^2 \tag{2}$$

$$\dot{x}_3 = -k_3 x_3 + 0.5 k_2 x_2^2 \tag{3}$$

$$\dot{x}_4 = +k_3 x_3 \tag{4}$$

The initial values for x_2, x_3, x_4 are zero, the initial value for x_1 is a free parameter that also has to be estimated from the data. The observed quantities are:

$$y_1 = k_5(x_2 + 2x_3) \tag{5}$$

$$y_2 = k_6(x_1 + x_2 + 2x_3) \tag{6}$$

$$y_3 = k_7 \,\mathrm{Epo}\, R_A\,, \tag{7}$$

where $k_5 - k_7$ have to be included as scaling parameters since only relative units can be measured. The factors of 2 in Eqs. (5) and (6) reflect the fact that a dimer produces a signal twice as high as a monomer. Note, that EpoR_A, measured by y_3, is not a dynamical variable but an external input. y_1 and y_2 will be used to estimate the parameters.

3.1. *Parameter estimation*

Formally, the problem of parameter estimation in the present context reads:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\, u) \tag{8}$$

$$\mathbf{y}(t_i) = \mathbf{g}(\mathbf{x}(t_i)) + \varepsilon(t_i)\,, \tag{9}$$

where u presents the time course of the activated Epo-receptor and $\varepsilon(t_i)$ the observational noise.

The simplest approach to parameter estimation in differential equations is based on estimating time derivatives from the observed data and transferring the task to a regression problem, see [Hegger *et al.*, 1998] for a more detailed description and successful application of this approach. This approach requires that the observational noise is small and that the observations offer enough information about the dynamical variables. Both needs are not fulfilled in the present setting since the noise is substantial and only linear combinations of the dynamical variables are observed.

A more promising approach is the *initial value approach* [Schittkowski, 1994] which takes into account the dynamical nature of the task. Choosing initial guesses for the initial value $x_1(t=0)$ and the parameters $\mathbf{k}$, here, one aims to minimize:

$$\chi^2(x_1(t=0),\,\mathbf{k}) = \sum_{i=1}^{N}\sum_{j=1}^{2} \frac{(y_j^D(t_i) - y_j^M(t_i;\, x_1(t=0),\,\mathbf{k})^2}{\sigma_{ij}^2}\,,$$

with $y_j^D(t_i)$ the measured data and $y_j^M(t_i;\, x_1(t=0),\,\mathbf{k})$ the output of the model. Assuming a Gaussian distribution for the measurement errors, this approach yields the maximum likelihood estimates for the parameters and the initial value which allows for statistical inference as discussed in Sec. 3.3.

Depending on the underlying dynamics, this approach might run into the problem of numerous

local minima [Timmer *et al.*, 1998]. In this case one should use the so-called *multiple shooting approach* developed in [Bock, 1981, 1983]; for applications see [Timmer *et al.*, 2000; Horbelt *et al.*, 2001]. Fortunately, the present task turned out to be well behaved enough to be handled by the *initial value approach*.

3.2. *Identifiability*

Given a dynamical system as in Eqs. (1)–(4) and observation functions as in Eqs. (5) and (6), it might be possible that not all parameters can be estimated [Vajda *et al.*, 1989; Ljung & Glad, 1994]. For example, solving Eq. (7) for Epo R_A and plugging it in Eqs. (1) and (2) leads to:

$$\dot{x}_1 = -k_1/k_7 x_1 y_3 \tag{10}$$

$$\dot{x}_2 = +k_1/k_7 x_1 y_3 - k_2 x_2^2 \tag{11}$$

showing that k_1 and k_7 cannot be estimated separately, but only their ratio.

Analogous calculations show that the identifiable parameter combinations are: $k_2 x_1(0)$, k_5/k_2, k_6/k_2. Furthermore, all dynamical variables x_i are identifiable only in combination with k_2 of the form $k_2 x_i$. Only k_3 is uniquely determined.

For the purpose of parameter estimation and model selection the model should be transferred into an identifiable one by introducing the identifiable parameter combinations as new parameters.

3.3. *Model selection*

Selecting an adequate model structure is the most difficult part of the modeling process for which no perfect solution exists. Here, we follow the forward selection strategy, i.e. we start with the simplest reasonable model and refine it in a way suggested by biochemical knowledge until further refinement does not improve the fit. We expect that more resolved future measurements will be called for a further refinement of our final model proposed here.

As a measure of a significant improvement we choose the likelihood ratio test (LRT) [Cox & Hinkley, 1994; Timmer & Klein, 1997]. If a more general model $\mathcal{M}_1$ with r_1 parameters does not offer a sufficient improvement of the fit compared to a simpler submodel $\mathcal{M}_2$ with r_2 parameters, the ratio of the likelihoods or, for convenience, twice the difference of the log likelihoods $\mathcal{L}$ are distributed as:

$$2(\mathcal{L}(\mathcal{M}_1) - \mathcal{L}(\mathcal{M}_2)) \sim \chi^2_{r_1 - r_2}$$

In the case of one additional parameter the critical value at 1% level of confidence for the LRT is 6.635. In this way, the LRT penalizes overparameterization. The above distributional result holds if the so-called standard conditions are fulfilled [Cox & Hinkley, 1994; Vuong, 1989]. The most important of these are:

(1) The models are nested.
(2) The true parameters are not part of the boundary of the parameter space.
(3) The parameters are identifiable under the null hypothesis.

The last point explains the above-mentioned advice to formulate models in such a way that the parameters are identifiable. Point 2 is of special importance here, since the parameters in the present setting mainly represent rate constants which are constrained to be non-negative. In the case that this type of nonstandard condition is given, it has been shown that the LRT statistic becomes a mixture of χ^2 distributions, in the simplest case of one additional parameter [Self & Liang, 1987]:

$$2(\mathcal{L}(\mathcal{M}_1) - \mathcal{L}(\mathcal{M}_2)) \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1\,,$$

where χ^2_0 represents a Dirac measure at zero. Disregarding distributional results of this kind renders the tests conservative, i.e. disabling the detection of a violation of the null hypothesis. Combined with an adjustment of the α niveau with increasing number of data, LRTs provide a consistent model selection strategy, i.e. if the more complex model is the true one, it will be recovered with probability one if the number of data points increases [Neyman & Pearson, 1933; Bauer *et al.*, 1988].

If the first standard condition is not fulfilled, i.e. the models are not nested, statistical model selection becomes more cumbersome. Here, we follow a bootstrap strategy suggested by [Hall & Wilson, 1991]. The basic idea is to investigate whether the empirical difference of χ^2 values of the two models is consistent with the distribution of the difference of χ^2 values based on one of the fitted models. In order to avoid confusion with superscripts we denote χ^2 values by C in the following.

The detailed procedure is as follows:

(1) Calculate the χ^2 values of both models

$$C_k = \sum_{i=1}^{N}\sum_{j=1}^{2} \frac{(y^D(i,j) - y^{M_k}(i,j))^2}{\sigma_{ij}^2}\,, \quad k = 1,\, 2$$

and their difference

$$C_{12} = C_1 - C_2$$

(2) Assume model 1 to be correct and simulate the time series $y^{M_1}(i, j)$. Generate bootstrap time series $y^{M_1^*}(i, j)$ by adding noise with variance σ_{ij}^2 to the simulated time series.
(3) For each bootstrap time series $y^{M_1^*}(i, j)$ fit both models and calculate

$$C_{12}^{*1} = (C_1^* - C_2^*) - C_{12}$$

(4) Reject the null hypothesis "Model 1 is correct" if C_{12} is not consistent with the hypothesis of being drawn from the distribution $C_{12}^{*1} - C_{12}$ at a given significance level.
(5) Repeat steps 2–4 assuming model 2.

Note, that possible outcomes of this procedure include the cases of rejecting as well as not rejecting both models.

In the context of model selection a remark on why not apply the popular *Akaike Information Criterion* (AIC) [Akaike, 1973, 1974] is in order. Akaike suggested to compare two models by:

$$\mathrm{AIC}(\mathcal{M}_i) = 2(\mathcal{L}(\mathcal{M}_i)) + 2r_i\,,$$

and choosing the one with smaller AIC. For the above setting with one additional parameter for the more general model, this exactly equals the LRT with $\alpha = 0.156$, consequently leading to false positive results in 15.6% of the cases independent from the number of data points [Atkinson, 1981; Teräsvirta & Mellin, 1986]. Thus, AIC is not a consistent model selection criterion. Furthermore, its derivation assumes the above-mentioned standard conditions and it is not known how it behaves if they are not fulfilled. The same holds for the classical F-test which is asymptotically equivalent to the LRT in the present setting.

4. A Dynamical Model of the JAK-STAT Pathway

In this section, we first describe the model selection process. The selected model is finally valitated by application to time series obtained from an independent experiment.

4.1. *Model selection*

In this section we describe the iterative modeling process of the dynamics of the JAK-STAT signalling pathway. The parameter estimation is based on three repetitions of the experiment. The dynamical parameters are fitted simultaneously for all experiments while the scaling parameters $k_7 - k_9$ are fitted separately. Only the resulting fit for one of the experiments will be displayed.

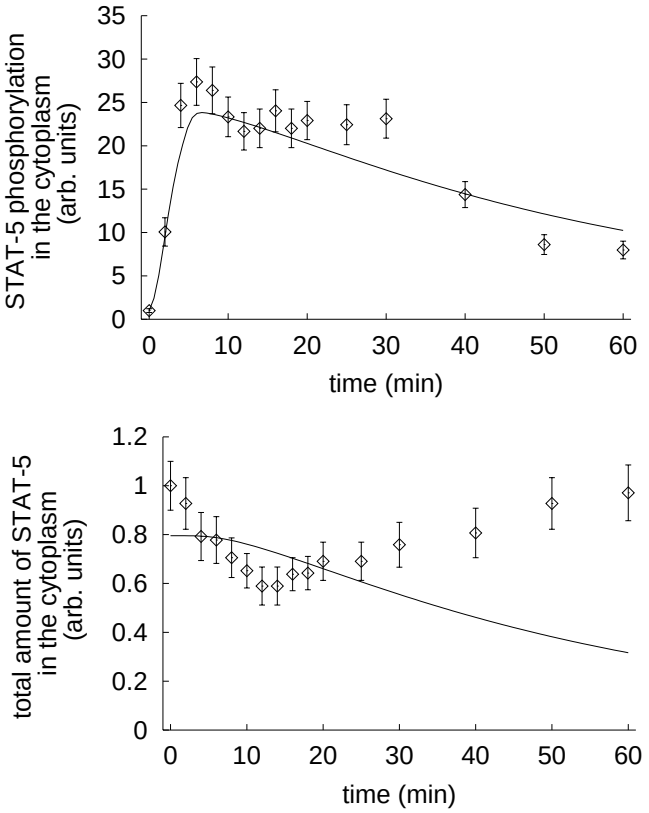


Fig. 3. Fit of model 1 (feed-forward) to the measured time series of phosphorylated and total STAT-5 in the cytoplasm.

We start the modeling procedure with model 1 already briefly discussed in Sec. 3 given by

$$\dot{x}_1 = -k_1 x_1 \,\mathrm{Epo}\, R_A \tag{12}$$

$$\dot{x}_2 = +k_1 x_1 \,\mathrm{Epo}\, R_A - k_2 x_2^2 \tag{13}$$

$$\dot{x}_3 = -k_3 x_3 + 0.5 k_2 x_2^2 \tag{14}$$

$$\dot{x}_4 = +k_3 x_3 \tag{15}$$

This model summarizes the hitherto point of view of a feed-forward cascade underlying the signal transduction, see Fig. 1. Figure 3 displays the resulting fit.

Even without any statistics, it can be concluded that this model is not able to describe the measured data. Note, that apart from the qualitative wrong behavior of the total amount of STAT-5, the model cannot reproduce the plateau in the time series

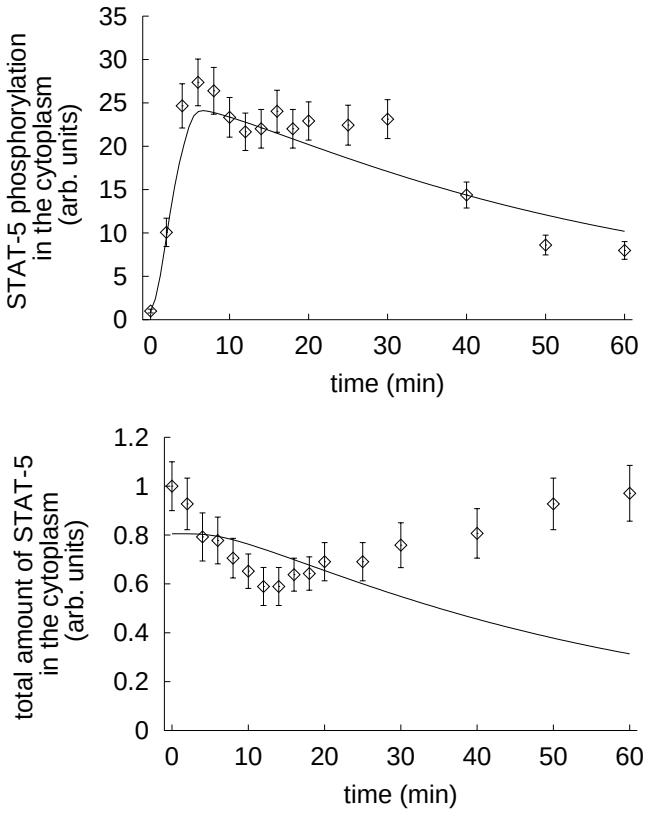


Fig. 4. Fit of model 2 (feed-forward model with a back-reaction) to the measured time series of phosphorylated and total STAT-5 in the cytoplasm.

for the phosphorylated STAT-5 in the period from 10–30 min.

Typically, chemical reactions are to some degree reversible. To test whether the inclusion of a back-reaction from the dimer to the monomer (model 2) improves the fit, Eqs. (13) and (14) of model 1 are replaced by:

$$\dot{x}_2 = +k_1 x_1 \,\mathrm{Epo}\, R_A - k_2 x_2^2 + 2k_3' x_3 \qquad (16)$$

$$\dot{x}_3 = -k_3 x_3 + 0.5 k_2 x_2^2 - k_3' x_3 \qquad (17)$$

Although this generalization improves the fit significantly ($LR = 7.8$, $p < 0.01$) the resulting fit in Fig. 4 again shows that this model cannot be sufficient.

The marginal differences between the fits of the two models is reflected by the fact that the back-reaction rate constant is only 3% of the forward-reaction rate constant.

The unsuccessful models considered so far assumed that the active role of STAT-5 ends in the nucleus. This triggers the idea that STAT-5, after dedimerization and dephosphorylation in the nucleus, might reenter the cytoplasm and is involved into another round of activation. A detailed description of STAT-5 in the nucleus would require at least four components: the free dimer, the dimer bound to promotor regions of the DNA, the dedimerized activated monomer and the deactivated monomer. Without measurements of at least some of these components the models would not be identifiable. Therefore, it is necessary to search for effective models to describe the behavior in the nucleus. The two simplest approaches are:

(1) One effective compartment.
This changes Eqs. (12) and (15) to

$$\dot{x}_1 = -k_1 x_1 \,\mathrm{Epo}\, R_A + 2k_4 x_4 \qquad (18)$$

$$\dot{x}_4 = +k_3 x_3 - k_4 x_4 \qquad (19)$$

(2) An effective delay.
This assumes that the sojourn time in the nucleus can be captured by a fixed delay. Equations (12) and (15) are replaced by

$$\dot{x}_1 = -k_1 x_1 \,\mathrm{Epo}\, R_A + 2k_4 x_3(t-\tau) \qquad (20)$$

$$\dot{x}_4 = +k_3 x_3 - k_4 x_3(t-\tau) \qquad (21)$$

To ensure mass conservation, the condition $k_3 \geq k_4$ has to hold.

Identifiability analysis as discussed in Sec. 3.2 shows that the parameters k_4 and τ are identifiable.

We first treat the second alternative of an effective delay (model 3). Figure 5 displays the resulting fits.

Apart from the likelihood ratio test that yields a test statistic of 838.0 compared to model 1, corresponding to $p < 10^{-5}$, the figures show an extremely accurate fit that even reproduces the plateau in the time series of the phosphorylated STAT-5 between 10 and 30 mins. The estimated sojourn time of STAT-5 in the nucleus is $\tau = 6.4$ min. The estimates of the parameters k_3 and k_4 are consistent with each other in accordance with the expectation that nuclear influx and outflux should balance. Therefore we identify the parameters in the following.

Again, we considered the possibility that adding a back-reaction of the dimer to the monomer in the cytoplasm might improve the fit (model 4). The likelihood ratio resulted in 0.7, corresponding to a p-value of 0.2 which indicated that this is not a significant improvement of the fit. The resulting fits are virtually nondiscriminatable from the fit of model 3 and therefore not displayed.

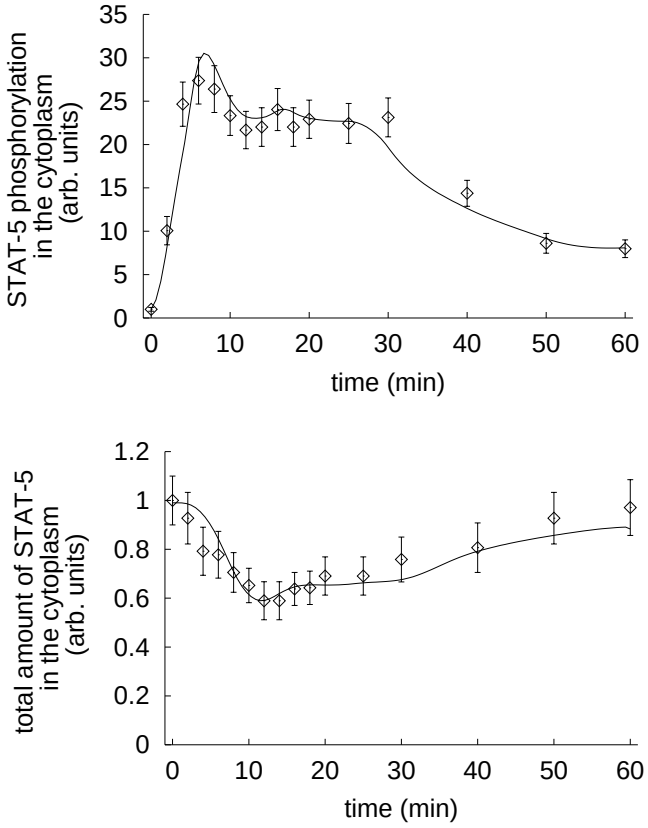


Fig. 5. Fit of model 3 (nuclearcytoplasmic cycling) to the measured time series of phosphorylated and total STAT-5 in the cytoplasm.

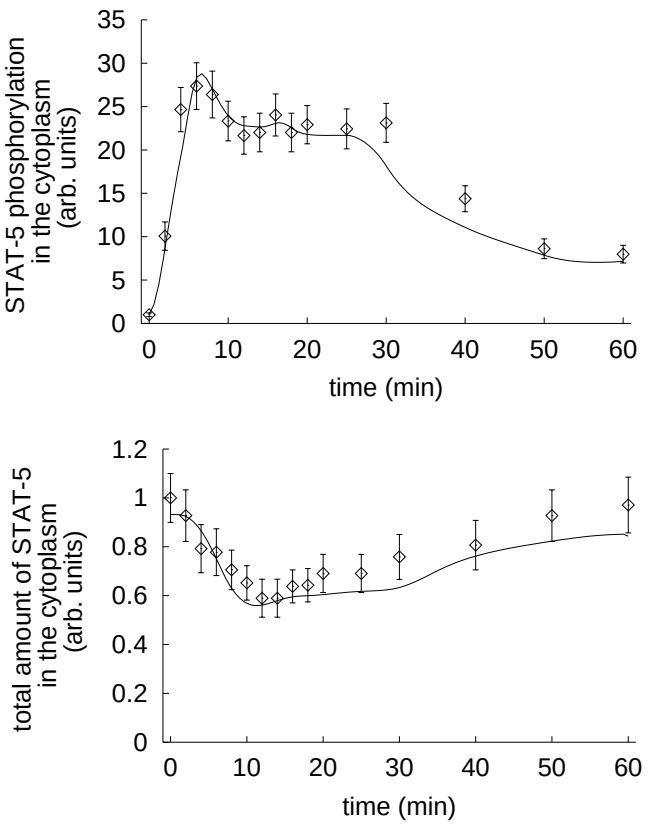


Fig. 6. Fit of model 5 (nuclearcytoplasmic cycling with delay distribution) to the measured time series of phosphorylated and total STAT-5 in the cytoplasm.

The assumption of a sharp sojourn time τ is certainly a simplified view. Thus, we investigated whether a distribution of delay times would significantly improve the fit (model 5). To render this approach feasible in the frame of parameter estimation, we assumed a Gaussian distribution of delay times, resulting in one additional parameter, the width of the distribution, for this generalization. The resulting likelihood ratio was 0.55, corresponding to $p = 0.23$, and states that this is not a significant improvement of the model. The resulting fit is shown in Fig 6, supporting the statistical analysis.

Finally, we investigated the description of the dynamics of STAT-5 in the nucleus by an effective compartment (model 6) compared to the delay model 3. Since these models are not nested, we applied the bootstrap procedure outlined in Sec. 3.3. We used 200 bootstrap time series. Figure 7 displays the cumulative distributions of the bootstrap test statistcs and the empirical value. Model 6 is rejected with $p = 0.007$, model 3 is consistent with the data ($p = 0.33$).

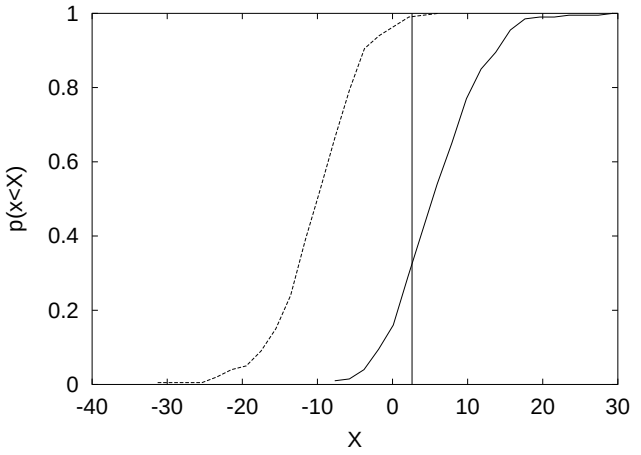
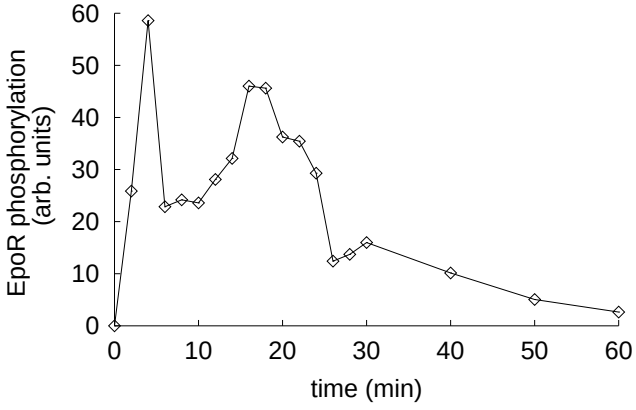


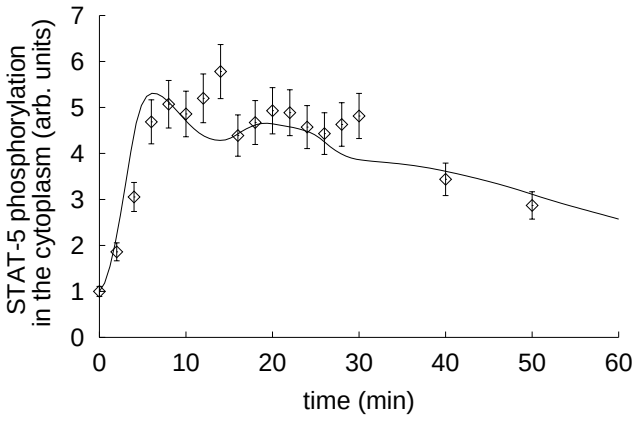
Fig. 7. Bootstrap distributions of $X = C_{12}^{*1} - C_{12}$ for model 3 (solid line) and model 6 (dashed line) and empirical value (horizontal line).

4.2. *Model validation*

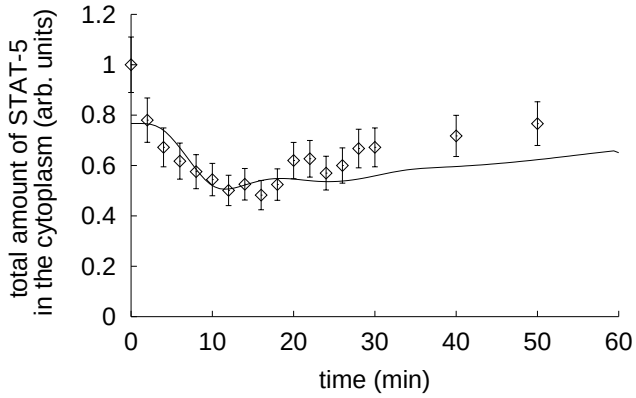
Statistical testing and visually convincing fits should only be regarded as necessary criteria for the judgement of a model. More challenging is the description of independent measurements by a



(a)



(b)



(c)

Fig. 8. Fit of independent data. Atypical activation of the Epo-receptor and fit of model 3 to the time series of phosphorylated and total STAT-5 in the cytoplasm.

before-fitted model. Therefore, we used the fitted model 3 from above and applied it to time series from a new experiment which showed an atypical activation of the Epo-receptor. Due to the atypical activation of the receptor the dynamics of the system visits regions in the phase space that were not seen during the previous optimization of the parameters. Note, that application of the model to data with a similar time course of the Epo-receptor would not serve for a validation of the model.

The dynamical parameters k_1–k_4 and τ were kept fixed and only the scaling parameters k_5–k_7 were estimated from the new data. Figure 8 displays the results that support the validity of our fitted model.

5. Discussion

We presented the derivation of a dynamical model for a cellular signalling pathway based on measured time series. The procedure uses parameter estimation given a certain model and statistical testing to decide between different models. Parameter estimation was performed by maximum likelihood estimation. On the one hand this results in efficient estimates [Cox & Hinkley, 1994], on the other it allows for statistical testing by likelihood ratio tests which are, at least asymptotically, maximal powerful.

For model selection we applied a forward selection strategy, i.e. starting with the simplest model we searched for generalizations until the quality of the fit did not increase furthermore. An alternative strategy would be backward selection, i.e. starting with the most general model and simplifying it until the model became insufficient. Theoretical considerations recommend the backward selection strategy [Mantel, 1970]. Unfortunately, the most general model in the present case would comprise so many parameters that the model would be non-identifiable. Therefore, we had to follow the forward selection strategy.

In the first steps of the model selection procedure we showed that the established model of a feed-forward cascade is not consistent with the measured time series and the feed-back that allows for nuclearcytoplasmic cycling of STAT-5 should be included in the model. A biological interpretation of this finding is first that the cell has to perform an optimal use of a limited pool of STAT-5 molecules. A second reason might be that cycling allows for a continuous monitoring of the receptor activity by the nucleus.

The fact that statistical testing preferred modeling of the nuclearcytoplasmic STAT-5 dynamics by a delay term (model 3) to modeling by an ordinary differential equation (model 6) is consistent with biochemical knowledge that STAT-5 binds for

a minimum time to the DNA promotor region. For model 6 there is no lower limit of the sojourn time of STAT-5 in the nucleus which seems to be detected by the statistical test. As in other cases of modeling biological systems by delay differential equations, e.g. the Mackey–Glass system [Mackey & Glass, 1977], the choice of a fixed sojourn time of STAT-5 in the nucleus is surely a rough simplification. Although statistical testing does not advocate a more complex model with a distribution of sojourn times, we believe that this is mainly due to the small amount of data and we expect that this part of the model will have to be refined in the presence of more detailed future measurements.

The fitted model allows for *in silico* experiments that predict the outcome of new biochemical experiments or even for experiments that cannot be performed biochemically. Such quantitative understanding of biochemical pathways opens the field of clinical application. Apart from the discovery of therapeutic targets [Nicholson, 2000] it can help to understand clinical markers [Simpson & Dorow, 2001].

Here, we discussed the analysis of a signalling pathway but the approach of data driven modeling also applies to other biochemical systems, e.g. metabolic pathways [Mendes & Kell, 1998]. This suggests that biochemistry offers interesting phenomena to be explained by applied nonlinear dynamics.

Acknowledgments

We thank R. Kehlenbach and B. Stenkamp for helpful discussions and U. Baumann, A. Walker and A. Geist for technical help. T. G. Müller has received financial support from the Deutsche Forschungsgemeinschaft (DFG).

References

Akaike, H. [1973] "Information theory and an extension of the maximum likelihood principle," *2nd Int. Symp. Information Theory*, eds. Petrov, B. & Csaki, F. (Akademiai Kiado, Budapest), pp. 267–281.

Akaike, H. [1974] "A new look at the statistical model identification," *IEEE Trans. Autom. Contr.* **AC-19**, 716–723.

Alon, U., Surette, M., Barkai, N. & Leibler, S. [1999] "Robustness in bacterial chemotaxis," *Nature* **397** 168–171.

Atkinson, A. [1981] "Likelihood ratios, posterior odds and information criteria," *J. Econometrics* **16**, 15–20.

Barkai, N. & Leibler, S. [1997] "Robustness in simple biochemical networks," *Nature* **387**, 913–917.

Bauer, P., Pötscher, B. & Hackl, P. [1988] "Model selection by multiple test procedures," *Statistics* **1**, 39–44.

Bhalla, U. & Iyengar, R. [1999] "Emergent properties of networks of biological signal pathways," *Science* **283**, 381–387.

Bock, H. [1981] "Numerical treatment of inverse problems in chemical reaction kinetics," in eds. Ebert, K., Deuflhard, P. & Jäger, W., *Modelling of Chemical Reaction Systems*, Vol. 18 (Springer, NY), pp. 102–125.

Bock, H. [1983] "Recent advances in parameter identification for ordinary differential equations," in *Progress in Scientific Computing*, Vol. 2, eds. Deuflhard, P. & Hairer, E. (Boston, Birkhäuser), pp. 95–121.

Boman, B., Fields, J., Bonham-Carter, O. & Rumquist, O. [2001] "Computer modeling implicates stem cell overproduction in colon cancer initiation," *Cancer Res.* **61**, 8408–8411.

Campbell, P. [1999] "Can physics deliver another biological revolution?" *Nature* **397**, p. 89.

Cox, D. & Hinkley, D. [1994] *Theoretical Statistics* (Chapman & Hall, London).

Darnell, Jr., J. [1997] "STATs and gene regulation," *Science* **277**, 1630–1635.

Downward, J. [2001] "The ins and outs of signalling," *Nature* **411**, 759–762.

Editorial [2000] "The changing face of biomedical research," *Nature Med.* **6**, p. 113.

Edwards, J., Ibarra, R. U. & Palsson, B. [2001] "*In silico* prediction of *escheria coli* metabolic capabilities are consistent with experimental data," *Nature Biotech.* **19**, 125–130.

Endy, D. & Brent, R. [2001] "Modelling cellular behavior," *Nature* **409**, 391–395.

Fussenegger, M., Bailey, J. & Varner, J. [2000] "A mathematical model of caspase function in apoptosis," *Nature Biotech.* **18**, 768–774.

Hall, P. & Wilson, S. [1991] "Two guidelines for bootstrap hypothesis testing," *Biometrics* **47**, 757–762.

Haspel, R., Salditt-Georgieff, M. & Darnell, Jr., J. [1996] "The rapid inactivation of nuclear tyrosine phosphorylated STAT-1 depends upon a protein tyrosine phosphatase," *EMBO J.* **15**, 6262–6268.

Haspel, R. & Darnell, Jr., J. E. [1999] "A nuclear protein phosphatase is required for the inactivation of STAT-1," *Proc. Nat. Acad. Sci.* **96**, 10188–10193.

Hegger, R., Kantz, H., Schmüser, F., Diestelhorst, M., Kapsch, R. & Beige, H. [1998] "Dynamical properties of a ferroelectric capacitor observed through nonlinear time series analysis," *Chaos* **8**, p. 727.

Horbelt, W., Timmer, J., Bünner, M., Meucci, R. & Ciofini, M. [2001] "Identifying physical properties of a CO_2 laser by dynamical modeling of measured time series," *Phys. Rev.* **E64**, 016222.

KEGG, http://www.genome.ad.jp/kegg/.

Kholodenko, B., Demin, O., Moehren, G. & Hoek, J. [1999] "Quantification of short term signalling by the epidermal growth factor receptor," *J. Biol. Chem.* **274**, 30169–30181.

Koshland, Jr., D. E. [1998] "The era of pathway quantification," *Science* **280**, 852.

Kremling, A. & Gilles, E. [2000] "The organization of metabolic reaction networks II. Signal processing in hierarchical structured functional units," *Metab. Engin.* **3**, 138–150.

Ljung, L. & Glad, T. [1994] "On global identifiability for arbitrary model parametrizations," *Automatica* **30**, 265–276.

Loomis, W. & Sternberg, P. [1995] "Genetic networks," *Science* **269**, p. 649.

Mackey, M. & Glass, L. [1977] "Oscillation and chaos in physiological control systems," *Science* **197**, 287–289.

Mantel, N. [1970] "Why stepdown procedures in variable selection," *Technometrics* **12**, 621–625.

McAdams, H. & Shapiro, L. [1995] "Circuit simulation of genetic networks," *Science* **269**, 650–656.

Mendes, P. & Kell, D. [1998] "Non-linear optimization of biochemical pathways: Application to metabolic engineering and parameter estimation," *Bioinformatics* **14**, 869–883.

Neyman, L. & Pearson, E. [1933] "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc.* **A231**, 289–337.

Nicholson, D. [2000] "From bench to clinic with apoptosis-based therapeutic agents," *Nature* **407**, 810–816.

Pellegrini, S. & Dusanter-Fourt, I. [1997] "The structure, regulation and function of the Janus kinase (JAK) and the signal tranducers and activatiors of transcription (STATs)," *Eur. J. Biochem.* **248**, 615–633.

Schittkowski, K. [1994] "Parameter estimation in systems of nonlinear equations," *Numer. Math.* **68**, 129–142.

Self, S. G. & Liang, K. Y. [1987] "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions," *J. Am. Stat. Ass.* **82**, 605–610.

Simpson, R. & Dorow, D. [2001] "Cancer proteomics: From signaling networks to tumor markers," *TRENDS Biotech.* **19**, S40–S48.

Swameye, I., Müller, T., Timmer, J., Sandra, O. & Klingmüller, U. [2003] "Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling," *Proc. Natl. Acad. Sci.* **100**, 1028–1033.

Teräsvirta, T. & Mellin, I. [1986] "Model selection criteria and model selection tests in regression models," *Scand. J. Stat.* **13**, 159–171.

Timmer, J. & Klein, S. [1997] "Testing the Markov condition in ion channel recordings," *Phys. Rev.* **E55**, 3306–3310.

Timmer, J., Müller, T. & Melzer, W. [1998] "Numerical methods to determine calcium release flux from calcium transients in muscle cells," *Biophys. J.* **74**, 1694–1707.

Timmer, J., Rust, H., Horbelt, W. & Voss, H. [2000] "Parametric, nonparametric and parametric modelling of a chaotic circuit time series," *Phys. Lett.* **A274**, 123–134.

Tomita, M. *et al.* [1999] "E-CELL: software environment for whole-cell simulation." *Bioinformatics* **15**, 72–84.

Vajda, S., Godfrey, K. & Rabitz, H. [1989] "Similarity transformation approach to identifiability analyis of nonlinear compartmental models," *Math. Biosci.* **93**, 217–248.

von Dassow, G., Meir, E., Munro, E. & Odell, G. [2000] "The segment polarity network is a robust developmental module," *Nature* **406**, 188–192.

Vuong, Q. H. [1989] "Likelihood ratio tests for modelselection and non-nested hypotheses," *Econometrica* **57**, 307–333.

Zheng, T. & Flavel, R. [2000] "Death by numbers," *Nature Biotech.* **18**, 717–718.