Bioinformatics, 36(6), 2020, 1848–1854
doi: 10.1093/bioinformatics/btz838
Advance Access Publication Date: 14 November 2019
Original Paper

Systems biology

Estimating chain length for time delays in dynamical systems using profile likelihood

Adrian L. Hauber [1,*], Raphael Engesser[1,2], Joep Vanlier[1] and Jens Timmer[1,2]

[1]Institute of Physics and [2]Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Freiburg 79104, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on July 22, 2019; revised on October 24, 2019; editorial decision on November 5, 2019; accepted on November 13, 2019

Abstract

Motivation: Apparent time delays in partly observed, biochemical reaction networks can be modelled by lumping a more complex reaction into a series of linear reactions often referred to as the linear chain trick. Since most delays in biochemical reactions are no true, hard delays but a consequence of complex unobserved processes, this approach often more closely represents the true system compared with delay differential equations. In this paper, we address the question of how to select the optimal number of additional equations, i.e. the chain length (CL).

Results: We derive a criterion based on parameter identifiability to infer CLs and compare this method to choosing the model with a CL that leads to the best fit in a maximum likelihood sense, which corresponds to optimizing the Bayesian information criterion. We evaluate performance with simulated data as well as with measured biological data for a model of JAK2/STAT5 signalling and access the influence of different model structures and data characteristics. Our analysis revealed that the proposed method features a superior performance when applied to biological models and data compared with choosing the model that maximizes the likelihood.

Availability and implementation: Models and data used for simulations are available at https://github.com/Data2Dynamics/d2d and http://jeti.uni-freiburg.de/PNAS_Swameye_Data.

Contact: adrian.hauber@fdm.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Data-based quantitative dynamical modelling is playing an important role in gaining mechanistic insights into biological systems (Kitano, 2005). Parameter values, such as reaction rate constants or initial values, can be estimated from experimental data by optimization techniques, e.g. maximum likelihood methods.

Because of the sheer complexity of biological systems, the inclusion of all known mechanisms and species is often not feasible. Rather, effective descriptions and coarse-grained models are used to capture the relevant processes while neglecting unnecessary details and therefore allow an efficient study of a given system. Conversely, if an effective model is able to describe the data adequately, it is understood that the assumptions made for coarse-graining hold true (Peifer *et al.*, 2014).

However, time delays can arise from the use of effective models. An example for such delays in dynamical systems is that of Mackey and Glass (1977), where the introduction of a time delay into a physiological control system results in chaotic oscillations. Time delays are also used in infection models (Culshaw and Ruan, 2000; Nelson and Perelson, 2002), population models (Gurney and Nisbet, 1980; May, 2001), gene regulation (Bratsun *et al.*, 2005; Jensen *et al.*, 2003) and signalling pathways (Srividhya *et al.*, 2007; Swameye *et al.*, 2003).

The linear chain trick (MacDonald, 1978) is considered a valuable alternative to the modelling approach of utilizing delay differential equations and provides a model closer to the underlying processes which cause the delay (Bachmann *et al.*, 2011; Müller *et al.*, 2013a, b; Raue *et al.*, 2009; Sobotta *et al.*, 2017). By introducing additional states as a linear chain, a time delay between input and output signal is generated.

Previous work has not addressed the question of how many intermediate states have to be used in the linear chain trick. Here we compare two methods for estimating the number of states in the linear chain from experimental data: the approach of using the model with the chain length (CL) corresponding to maximal likelihood values and a criterion to access CL information based on the techniques of profile likelihood (Raue *et al.*, 2009) and model reduction (Maiwald *et al.*, 2016).

We first show that both strategies are suitable to reproduce the correct CL from simulated data and provide a consistent estimator. After that, we estimate CLs from biological data for the model of Epo-induced JAK/STAT signalling by Bachmann *et al.* (2011), which yields reasonable results. However, many biological models in the literature were established and calibrated with data mostly not suitable to infer CLs in terms of large noise, sparse time resolution or few observed species, which we show through several

simulation studies. Experimental design makes it possible to identify the most informative data points that are needed to infer CLs. We apply both methods to a number of biological models and infer CLs.

2 Materials and methods

In this section, we first introduce methods used for dynamical modelling and parameter estimation, as well as criteria on identifiability and likelihood for recovering the CL from data.

2.1 Dynamical modelling and parameter estimation

Biochemical reaction networks can be condensed into mathematical rate equations comprising the molecular species x, external stimuli u and the parameters θ:

$$\dot{x} = f\left(x, u, \theta\right) \tag{1}$$

The initial conditions $x(t=0)$ can be treated as additional parameters if they cannot be specified beforehand. Because of measurement-related scaling and offsets as well as measurement error, the states x are in general not directly accessible to experiments. Furthermore, systems are often only partially observed, i.e. only certain species or combinations of them can be measured. The observation function links the states to the measurements:

$$y = g\left(x, \theta\right) + \epsilon \tag{2}$$

The measurement errors ϵ are usually assumed to represent additive Gaussian noise, i.e. $\epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right),\ \epsilon_i \in \epsilon$. Multiplicative noise can be accounted for by considering the model and data in log-space (Kreutz *et al.*, 2007).

Calibration of the model is usually performed using the maximum likelihood method. For n_d data points $\{y_i^D\}$, the optimal parameter values are acquired by minimizing the negative log-likelihood function

$$\mathcal{L}\left(\theta\right) = -2\log\ell\left(\theta\right) = \sum_{i=1}^{n_d}\left[\frac{\left(y_i^D - y\left(\theta\right)\right)^2}{\sigma_i^2} + \log 2\pi\sigma_i^2\right] \tag{3}$$

One technique for accessing the confidence interval of an estimated parameter θ_i in non-linear models is calculating the profile likelihood (Raue *et al.*, 2009):

$$PL(\theta_i) = \text{argmin}_{j\neq i} L\left(\theta\right) \tag{4}$$

Confidence intervals are defined by the inequality (Kreutz *et al.*, 2012)

$$PL(\theta_i) - L\left(\theta\right) \leq Q\left(\chi_1^2,\ 1-\alpha\right), \tag{5}$$

where $Q(\chi_1^2,\ 1-\alpha)$ is the $(1-\alpha)$-quantile of the χ_1^2-distribution. Throughout this work, we will use $\alpha = 0.05$. Non-finite confidence intervals indicate a non-identifiable parameter. In general, a parameter with a flat profile likelihood is considered to be *structurally non-identifiable* (Raue *et al.*, 2009). If the profile likelihood is not entirely flat but neither crosses the significance threshold in at least one direction of the parameter axis, the parameter is called *practically non-identifiable*.

2.2 Linear chain trick

Consider a reaction network comprising a time delay between the states x_0 and x_n. When utilizing the linear chain trick, this delay is replaced by a linear chain of states $x_1, x_2, \ldots, x_{n-1}$, which can be translated into an ODE model:

$$\dot{x}_i = k_{\text{delay}} x_{i-1} - k_{\text{delay}} x_i, \quad i = 1,\ 2, \ldots, n \tag{6}$$

This approach is justified because the linear chain leads to the same expression for x_n as a continuous delay,

$$\dot{x}(t) = f\left(x(t), \int_{-\infty}^{t} \mathrm{d}\tau\ K(t-\tau)\ x(\tau)\right) \tag{7}$$

with a gamma-type delay kernel K (Smith, 2011). To further analyse Equation (6), it is convenient to apply the Laplace transform $L(\cdot)$, which yields (Supplementary Section S1):

$$L(x_n) = \left(\frac{k_{\text{delay}}}{s+k_{\text{delay}}}\right)^n L(x_0) \tag{8}$$

Different delays can be accounted for by tuning the length of the chain n and the reaction rate constant k_{delay}.

2.3 Estimating CLs

2.3.1 Optimizing the Bayesian information criterion

In order to find the CL associated with the delay in experimental data, one approach is to compare likelihood values of fits of model with different CLs. The CL that leads to maximum likelihood can be interpreted as optimally describing the true delay. Since increasing the CL does not alter the number of parameters in the model and we perform the comparison on the same dataset, this corresponds to optimize the Bayesian information criterion (BIC). Note that it is not possible to conduct a likelihood ratio test because the models for different CLs are not nested, i.e. they do not emerge from each other by introducing parameter constraints.

2.3.2 The identifiability criterion

Consider data comprising a time delay, i.e. data for x_n generated from the model in Equation (6). To recover the true CL, it is necessary to use an auxiliary model that describes a linear chain with states $y_1, y_2, \ldots, y_m$, where one state transition is assigned a distinct reaction rate k_{skip}. Equation (8) changes accordingly:

$$L(y_m) = \frac{k_{\text{skip}}}{s+k_{\text{skip}}}\left(\frac{k_{\text{delay}}}{s+k_{\text{delay}}}\right)^{m-1} L(x_0) \tag{9}$$

Since it comprises an additional parameter, this model has more freedom when fitting data than the model in Equation (6):

- When trying to fit the model for $m = n$, it is possible to exactly recover the model in Equation (6) by setting $k_{\text{skip}} = k_{\text{delay}}$. This should provide the best possible fit and therefore any changes in parameters should induce a significant worsening, which translates into k_{delay} being identifiable.
- By construction of the auxiliary model, it is also possible to exactly recover the true model if $m = n+1$.When $k_{\text{skip}} \to \infty$, Equation (9) becomes

$$L(y_{n+1}) \to \left(\frac{k_{\text{delay}}}{s+k_{\text{delay}}}\right)^n L(x_0) = L(x_n), \tag{10}$$

  which can be interpreted as a shortening of the chain by one step. Fit quality increases when increasing k_{skip} which makes the parameter practically non-identifiable.

For $m < n$, a further shortening does typically not improve the fit –k_{skip} is identifiable. For $m > n+1$, a shortening of the chain by one step would not be sufficient to recover the true model. Instead, the second timescale provided by k_{skip} when the chain is not shortened can be employed to achieve the best possible fit. No statement on identifiability can be made in this case.

These considerations allow us to derive a criterion for recovering the CL from data: *The true CL is the smallest CL for which the auxiliary model yields a non-identifiability of* k_{skip}. The particular case of short true CLs $n = 1, 2$ has to be considered separately. If the data originate from a model with $n = 1$, the true CL can be

recovered by shortening the chain of all but one step. This is achieved by setting $k_{\text{delay}} \to \infty$ and $k_{\text{skip}} = k_{\text{delay}}$:

$$L(y_m) \to \frac{k_{\text{skip}}}{s + k_{\text{skip}}} L(x_0) = \frac{k_{\text{delay}}}{s + k_{\text{delay}}} L(x_0) \tag{11}$$

In this case, the non-identifiability occurs in k_{delay}. For $m = 2$, k_{skip} can become non-identifiable too as those two parameters are equivalent and can be interchanged in this setting.

2.4 Experimental design for recovering CLs

Certain features of experimental data can impede recovering the true CL, e.g. a poor choice of sampling time points or high noise levels that mask the dynamics of a system. Those data characteristics can induce a practical non-identifiability on k_{delay} or the auxiliary parameter k_{skip} of the auxiliary model discussed in Section 2.3.2, which renders the inference of CLs infeasible.

To test whether the data are suitable for CL inference, one can perform a simulation study using realistic parameter values acquired from fitting the data with the same sampling time points as in the experimental data. By utilizing the likelihood or the identifiability criterion, one can infer CLs from simulated data. Comparing the CL of the model used for simulating data with the CL recovered from that data allow to access whether the properties of the data are suitable.

If the data are not suitable, the techniques of experimental design can be employed. For example, one can add sampling time points carrying information on k_{skip} by identifying the regions in which the model predictions along the parameter profiles show the largest spread (Steiert *et al.*, 2012). Ideally, the CLs inferred from applying the criterion on simulated data and the CL of the model used for simulation should be identical.

3 Results

In this section, we evaluate the performance of the two methods to recover the CL from simulated data of a model with given CL. After that, we apply both methods to different biological models of the JAK-STAT signalling pathway (Bachmann *et al.*, 2011; Merkle *et al.*, 2016; Swameye *et al.*, 2003), as well as to the model of IL-6-induced JAK1-STAT3 signalling of Sobotta *et al.* (2017) with their respective measured biological data. For that we use the open-source MATLAB toolbox Data2Dynamics (Raue *et al.*, 2015; Steiert *et al.*, 2019).

3.1 Simulation study: linear chain trick

We start by analysing a simple model in which species x_0 (Input) activates the delayed production of species x_n (Output), which is degraded afterwards. The delay is accounted for by a linear chain, whose first state is activated by x_0 and whose last state is converted to x_n (Fig. 1A). The system of ODEs describing this model is equal to Equation (6).

The initial conditions are represented by the steady state, which is $x(t = 0) = 0$. We simulated 25 datasets for the observables x_0 and x_n for a CL of $n = 3$. We chose an exponentially decaying function for the input function x_0. A typical data realization is depicted in Figure 1B. We fitted models with different CLs to estimate the true delay via optimizing the BIC and the identifiability criterion.

The profile likelihood (Fig. 1C) shows that k_{skip} is identifiable for $m \leq n = 3$. The multiple optima are a consequence of the additional auxiliary parameter. For $m = n + 1 = 4$, k_{skip} is mostly non-identifiable, i.e. the chain is shortened. The identifiability criterion therefore is able to recover the true CL in 20 out of the 25 data realizations. For $m = n + 2 = 5$, k_{skip} is identifiable in the vast majority of realizations. An insufficient shortening of only one step is not realized. Instead, the second timescale introduced with k_{skip} is utilized to achieve the best possible fits (Fig. 1D).

The BIC was able to recover the true CL in only 16 out of the 25 cases (Fig. 1E). In a substantial portion of data realizations, the CL is slightly overestimated which leads to the overall estimate being biased in this setting.

3.1.1 Varying noise levels

In order to determine how sensitive the two methods are to noise in the data, we looked at four different situations in the range from

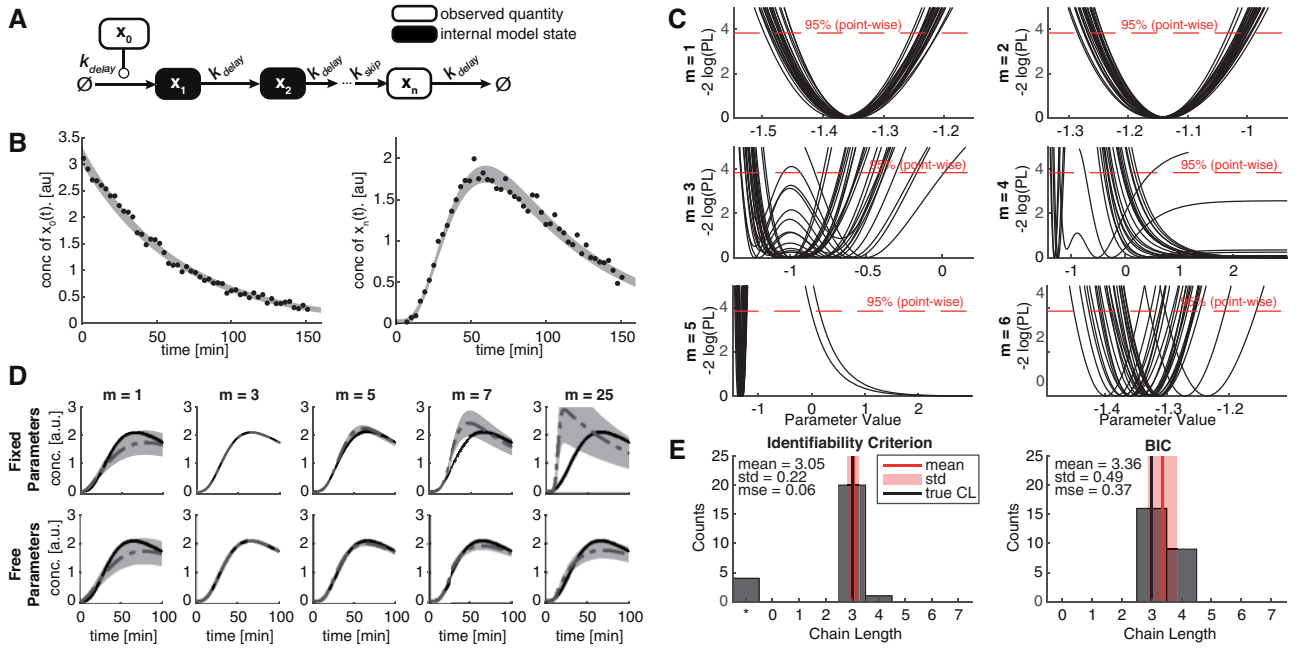

Fig. 1. (**A**) Reaction scheme of the (auxiliary) linear chain model. (**B**) A typical realization of data simulated with the linear chain model for the observables x_0 and x_n. Shaded areas represent one standard deviation of the error model around the model trajectory the data were simulated with. (**C**) Profile likelihood of the auxiliary parameter k_{skip} for fits of the auxiliary model for different chain lengths m. All profiles were shifted to the same baseline. The dashed line indicates the threshold for the 95% confidence interval which is given by the quantile function of the χ^2 distribution in Equation (5). (**D**) The top row shows densely sampled data (black) and fits (dashed-dotted) with respective error models (grey) with $k_{\text{skip}} = 10^3$ 1/min for $m > n$, i.e. a shortening is enforced for models with overly long chains. The bottom row depicts fits without such restrictions. (**E**) Comparison of chain length results for the linear chain model using the BIC and the identifiability criterion. *Cases in which the auxiliary parameter did not become non-identifiable for any chain length and therefore no statement on the chain length is possible

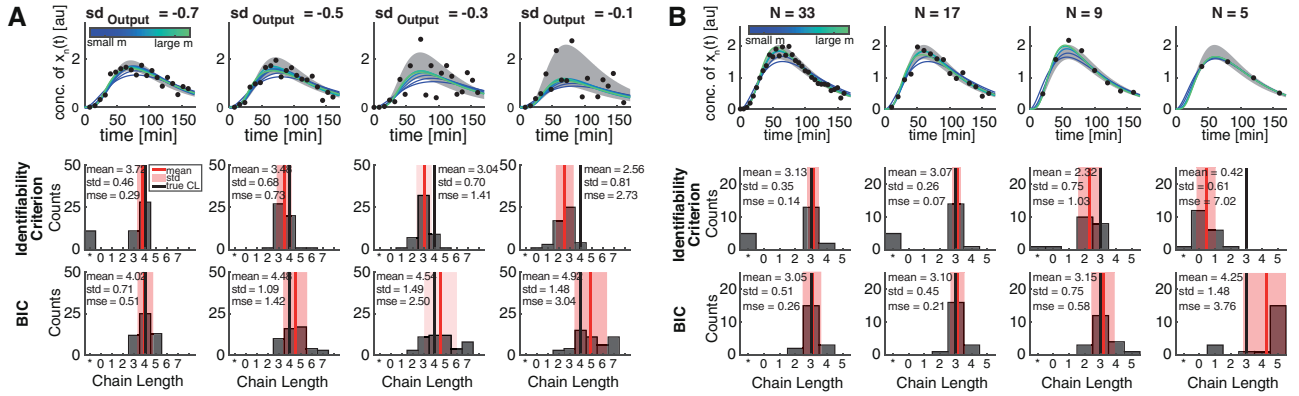

Fig. 2. Evaluation of the performance of the two methods in different data settings. The top row shows a typical data realization together with the respective fits for models with different chain lengths. Histograms of estimated chain lengths are shown below with the black bar representing the chain used for simulations. The mean and standard deviation of estimated chain lengths are denoted by a red bar and a shaded red area, respectively. (**A**) Variation the level of relative noise. (**B**) Variation of the number of data points N in one realization. (Color version of this figure is available at *Bioinformatics* online.)

moderate relative noise of $10^{-0.7} \approx 0.2$ to very large relative noise of $10^{-0.1} \approx 0.8$ and simulated 50 datasets each for a CL of $n = 4$. We then applied both criteria to recover the CL from the data.

For small noise levels, both methods perform equally well. The identifiability criterion produces a result which has a slightly smaller variance but comes with the trade-off of being biased (Fig. 2A). While the bias is only marginal in this setting, it becomes substantial when going to larger noise levels. Additionally, in certain data realizations, the identifiability criterion is not able to infer any CL at all because the auxiliary parameter remains identifiable for all tested CL. The CLs associated with maximum likelihood provide an unbiased estimator which at the same time has a larger variance. The mean squared error, defined by $\mathrm{MSE(CL)} = \langle(\hat{\mathrm{CL}} - \mathrm{CL})^2\rangle$, where $\hat{\mathrm{CL}}$ is an estimate of the CL, is lower for the identifiability criterion. The reduced variance outweighs the bias. However, both methods result in the correct value for the CL length for small noise levels.

3.1.2 Varying the amount of data

We evaluated the influence of the number of data points on the performance of both methods. CLs were estimated from $N \in \{5,\ 9,\ 17,\ 33\}$ data points with the same error model in 20 simulated datasets for each N.

The analysis shows that for the given noise, $N = 17$ data points are sufficient to obtain the best possible CL estimate with both methods (Fig. 2B). When data sparsity increases, the identifiability criterion again shows a bias towards small CLs and has a reduced variance when compared with the BIC. The latter features a lower mean squared error except for the scenarios with extreme data sparsity. Trajectories of fits of models with different CLs differ mostly in the region of their maximum. Because for $N = 5$, there are no data points in this specific region, the data as a whole does not carry enough information to infer any CL. Therefore, while longer chains apparently lead to slightly better fits, these changes are regarded as not significant by the identifiability criterion. With a larger number of data points both criteria lead to the correct CL. This asymptotic correctness when decreasing noise and data sparsity implies statistical consistency for both methods. We also analysed the behaviour when using very large CLs in Supplementary Section S2.

3.2 Simulation study: translation model

Consider the following simple model of protein translation (Fig. 3A): an input function mRNA0 activates the production of mRNA, which in turn activates the production of a protein in a delayed reaction. Both mRNA and protein become degraded afterwards. The delay is modelled by introducing a linear chain of mRNA states:

$$\begin{aligned}
\dot{\mathrm{mRNA}}_1 &= k_{\mathrm{in}} mRNA_0 - k_{\mathrm{delay}} mRNA_1 \\
\dot{\mathrm{mRNA}}_i &= k_{\mathrm{delay}} mRNA_{i-1} - k_{\mathrm{delay}} mRNA_i \ \text{ for } 2 \le i \le n-1 \\
\dot{\mathrm{mRNA}}_n &= k_{\mathrm{delay}} mRNA_{n-1} - k_{\mathrm{out}} mRNA_n \\
\dot{\mathrm{Protein}} &= k_{\mathrm{prod}} mRNA_n - k_{\mathrm{deg}} Protein
\end{aligned} \tag{12}$$

Typically, data are available for the protein itself and the input mRNA_0, which can represent the output of any upstream reaction network, but not for mRNA. Since mRNA is not actually converted from input or to protein, the scales of their concentration levels are independent. mRNA concentrations can be arbitrarily high or low for a given protein concentration if the protein production rate k_{prod} is tuned accordingly (Fig. 3B). Due to this fact, it is not possible to infer mRNA concentrations levels with only input and protein concentrations measured. These levels are governed by the non-identifiable parameters k_{in} and k_{prod} (Fig. 3C), which can be seen from the Laplace transform of the protein state:

$$L(\mathrm{Protein}) = \frac{k_{\mathrm{prod}}}{s + k_{\mathrm{deg}}} \frac{k_{\mathrm{in}}}{s + k_{\mathrm{out}}} \left(\frac{k_{\mathrm{delay}}}{s + k_{\mathrm{delay}}}\right)^{n-1} \tag{13}$$

Because the Laplace back-transform is linear, k_{in} and k_{prod} appear as multiplicative factors in the untransformed concentration function and therefore do not alter the dynamics but only the concentration scale. A change of k_{in} in one direction, which leads to different mRNA levels, can always be compensated for by a change in rate k_{prod} in the other direction so that the experimentally measured outcomes of input and protein are not altered, which makes both parameters non-identifiable.

We simulated 20 densely sampled datasets with small noise for the initial conditions $x(t = 0) = 0$ and applied the BIC as well as the identifiability criterion to access CL information. The identifiability criterion slightly underestimates the CL (Fig. 3D). Optimizing the BIC yields estimates which are closer to the true CL on average but have a much larger variance. The mean squares error is much lower for the identifiability criterion.

3.3 Application study: model of JAK2/STAT5 signalling by Bachmann *et al.*

To evaluate the performance of the two methods on biological models, we applied them to the mathematical model of JAK2/STAT5 signalling (Bachmann *et al.*, 2011). STAT5 phosphorylation, which is activated by Erythropoietin, is inhibited by the two proteins CIS and SOCS3 (Fig. 4A). Production of both proteins is activated by nuclear phosphorylated STAT5 (npSTAT5) in a delayed mechanism. Bachmann *et al.* used the linear chain trick with five intermediate states to model both delays (Fig. 4B).

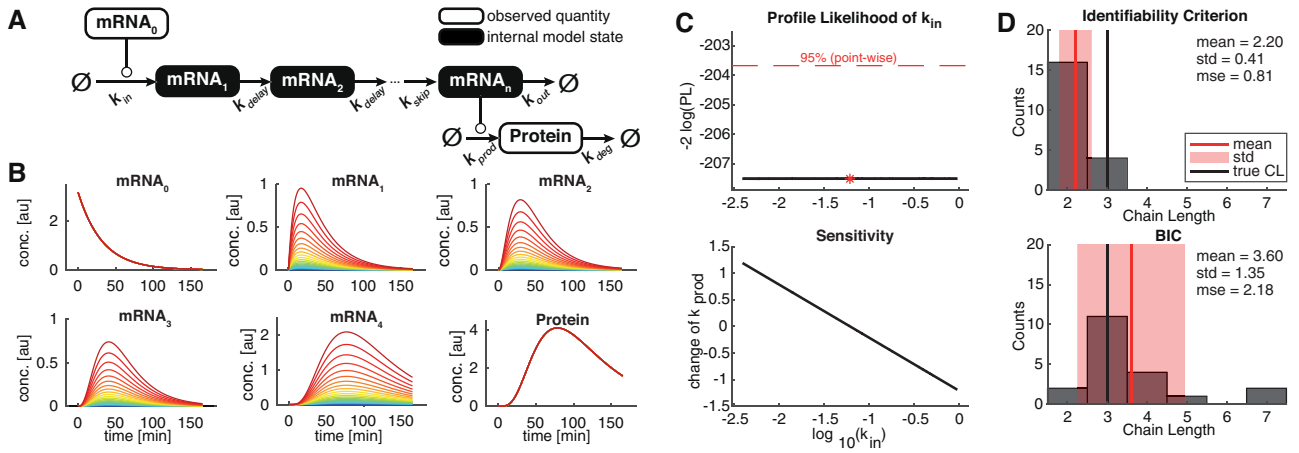

Fig. 3. (**A**) A simple model of protein transcription: the input $mRNA_0$ activates the production of mRNA, which in turn activates the production of a protein in a delayed reaction. Both, mRNA and protein become degraded afterwards. The delay is modelled by introducing a linear chain of mRNA states. (**B**) Model prediction trajectories ($m = 4$) compatible with the likelihood threshold in the parameter profile (representative data realization, $n = 3$). Different colours indicate the different values of k_{in}. (**C**) Profile likelihood of k_{in} and sensitivity of k_{prod} with respect to changes in k_{in}. (**D**) Comparison of chain length results for the translation model for both methods. (Color version of this figure is available at *Bioinformatics* online.)

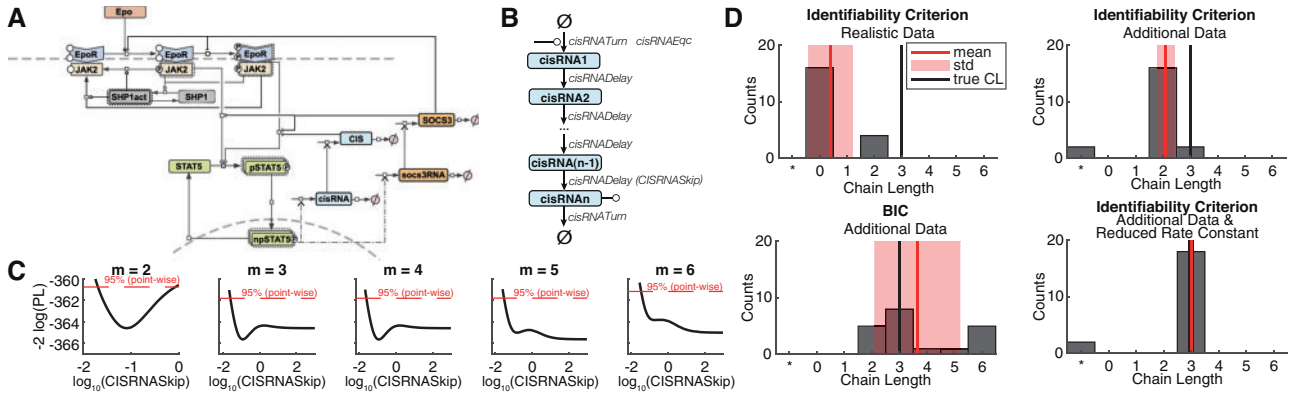

Fig. 4. (**A**) Reaction scheme of the model of JAK2/STAT5 signalling (originally published in Bachmann *et al.*, 2011). Dashed-dotted lines indicate delayed reactions. (**B**) Detailed representation of the delay chain in the production of CIS. The auxiliary model includes the distinct reaction rate *CISRNASkip*, as indicated by the parentheses. (**C**) Likelihood profiles of the auxiliary parameter *CISRNASkip* for fitting of biological data. (**D**) Performance of both criteria when applied to a biological model with simulated data

The equations governing the delay are similar to those of the linear chain model discussed in Section 3.1, but with distinct in- and outgoing reaction rates, as described in Supplementary Section S3. npSTAT5 acts as the input function x_0, while the last state of the linear chain x_n activates the production of the respective protein, CIS or SOCS3, which is degraded afterwards. Auxiliary models that are needed to recover the CL via the identifiability criterion comprise the additional parameters *CISRNASkip* and *SOCS3RNASkip*.

3.3.1 Experimental data

Bachmann *et al.* collected data for multiple observables using quantitative immunoblotting, mass spectrometry and dose–response experiments. Fitting of the auxiliary models described above for different CLs and performing an identifiability analysis allows inferring the CL from experimental data using the identifiability criterion. We also conducted a CL analysis with the BIC.

The two delay chains were analysed separately, i.e. a length of five was assumed for the respective other chain. With the identifiability criterion, we found that the CLs for the delayed production of proteins can be reduced from five to two for CIS (Fig. 4C) and zero for SOCS3 (see Supplementary Section S4). Using the BIC, CLs are estimated to four for CIS and at least six for SOCS3. Only a lower bound for the CL of SOCS3 can be provided because likelihood values decreased with increasing CL and we did not analyse chains with lengths larger than six. We repeated the analysis for both delay chains incorporating these results for the respective other chain, which resulted in the same CLs. We therefore conclude that our assumption is justified.

3.3.2 Artificial data

The analysis in the previous section showed that the experimental data does not give rise to a delay chain in production of SOCS3. In order to test whether the true CL can be inferred in the given data setting, we performed a simulation study. We simulated data from a model with the CL $n = 3$, which mimics the characteristics of the experimentally collected data: simulations were conducted for the same observables and measurement timepoints. Standard deviations of simulated data were obtained by the error parameters of the best fit. This procedure can also be interpreted as a bootstrap approach to quantify the uncertainty of the CL estimate.

The results show that the identifiability criterion does not correctly identify the true CL but underestimates it in all of 25 data realizations (Fig. 4D, Realistic Data). This fact implies that inferring CLs is not possible in the given experimental setting. However, biological a priori knowledge suggests the existence of such a chain that represents mRNA in the process of translation. Therefore, we proceeded with an experimental design approach: time points were added to the observables that are most sensitive to changes in *CISRNASkip* and *CISRNADelay*. This approach led to a CL of two in most cases. Accessing the CL via optimizing the BIC yields far less

Table 1. Comparing the performance of both criteria on published biological models

Model	Bachmann		Swameye	Merkle		Sobotta
Chain	CIS	SOCS3	STAT	CIS	SOCS3	SOCS3
CL in literature	5	5	5	5	5	5
CL from identifiability criterion	2	0	1	3	0	5
CL from optimizing the BIC	4	6+	1	3	5+	3

Note: Cases where the identifiability criterion was not fulfilled even for the model with the largest analysed chain length n are denoted with $n+$.

informative results (Fig. 4D, Additional Data). While the bias is smaller, the variance is considerably larger which results in a much larger mean squared error. This renders the CL estimate much less useful when dealing with a single dataset. The rate parameter governing mRNA production and degradation, *CISRNATurn*, was estimated to its upper bound at 10^3 1/min, which corresponds to a shortened delay chain (Supplementary Section S3). Therefore, the artificial data were effectively simulated for $n = 2$, which renders the result of estimated CLs reasonable. Using a smaller value of $CISRNATurn = 10^{-1}$ 1/min for simulations results in CLs that closely resemble $n = 3$ (Fig. 4D, Additional Data and Reduced Rate Constant).

3.4 Additional biological models

Apart from the model of JAK2/STAT5 signalling (Bachmann *et al.*, 2011), we analysed the linear chains in a number of models (Merkle *et al.*, 2016; Sobotta *et al.*, 2017; Swameye *et al.*, 2003) included in the benchmark collection (Hass *et al.*, 2019), which all contain at least one linear chain used to model a time delay in protein production (Supplementary Section S5). We applied both criteria to infer CLs. The lengths of the linear chains were assumed to be independent. The inclusion of the auxiliary parameter is analogous to what is described in detail in Supplementary Section S3. Results suggest that CLs used in published models mostly are overly large (Table 1).

4 Conclusion

Apparent time delays in partly observed biochemical reaction networks can be modelled by lumping a more complex reaction into a series of linear reactions often referred to as the linear chain trick. However, the number of additional states, i.e. the length of the linear chain, remains unclear. Accessing the CL by taking a model that is associated with the maximal likelihood value leads to a CL that optimally describes the true delay present in the data (optimizing the BIC). Additionally, we derived a criterion for recovering the CL from data utilizing a profile-likelihood-based identifiability analysis inspired by model reduction techniques with an auxiliary model including an additional auxiliary parameter (identifiability criterion). The proposed method provides the shortest CL which cannot be rejected by a likelihood ratio test compared with a larger model and therefore leads to the simplest model possible.

We evaluated the performance of both criteria with different toy models. Both methods were able to correctly recover the true CL from simulated data. We continued by analysing the effects of data properties such as noise levels and the number of sampling time points. The performance of both methods for large amounts of simulated data and low noise levels in providing a narrow distribution of CLs centred around the true CL used for simulations implies statistical consistency. While the identifiability criterion provides more accurate results with a smaller variance for well-behaved data, it suffers from bias when noise levels go up or the sample size goes down. However, the majority of simulation studies we conducted showed that it yields results with a lower mean squared error than the BIC.

We considered the model of JAK2-STAT5 signalling (Bachmann *et al.*, 2011) as a candidate for a typical biological model. When applied to biological data, both methods provide reasonable results. The linear chain governing the production of CIS can be shortened to two steps. No delay in the production of SOCS3 can be inferred from the data. While biological knowledge implies the existence of such a chain as a consequence of various mRNA states occurring in the process of protein translation, this is not portrayed by the data. We analysed whether it is possible to recover the true CL from realistically simulated data which turned out not to be the case. Using the principles of experimental design, we added informative time points to the relevant observables which rendered CL estimation feasible. Again, the mean squared error was lower for the identifiability criterion. Additionally, we examined several more biological models from the benchmark collection (Hass *et al.*, 2019).

Both methods, optimizing the BIC as well as the identifiability criterion, were able to recover the correct CLs from simulated data. In realistic settings this becomes more demanding because of less observed species, sparsely sampled data and larger noise levels. While the simulation studies show a bias towards shorter chains for the identifiability criterion, the estimated CL is derived from a likelihood-ratio test, i.e. it represents the smallest one that cannot be rejected when compared with a larger model. Therefore, this method leads to the smallest and simplest model possible, which makes the model more easy and understandable and facilitates comprehending the internal mechanics and characteristic features of the system. Especially in the case of realistic data with a biological model, the identifiability criterion showed to be more suitable to recover the correct CL.

When interpreting the CLs acquired by either method, one has to keep in mind that the linear chain trick assumes the delay to be caused by a series of reaction with equal rate constants, which is not a natural assumption. Whether or not this fact has substantial effects on the number of states inferred is subject to further research. However, in order to find the smallest model possible, these methods may be applied without biologically interpreting the results.

The simulation study has shown that both methods are suitable for recovering CL information from data and provide a consistent estimator. The drawbacks of the identifiability criterion are a potentially biased estimation for inappropriate data with high noise levels and few informative data points. The BIC on the other hand produces result with a much larger variance. Overall, the identifiability criterion features a superior performance as it leads to results with a lower mean squared error. An application to the model of JAK2-STAT5 signalling of Bachmann *et al.* showed that typical experimental data quality is not sufficient to reliably estimate CLs. However, with new measurement techniques such as mass spectrometry or imaging methods allowing the collection of high-quality data in terms of error or time resolution, this criterion provides a promising technique for inferring CL information on various systems. As the field of systems biology increasingly turns to large-scale models, reducing unnecessary complexity becomes more and more important. The identifiability criterion presented in this work might be an appropriate tool to accomplish this goal if the data quality is sufficient.

Funding

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – EXC-2189 – Projektnummer 390939984. This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2189 – Project ID: 390939984. The authors acknowledge support for high-performance computing by the state of Baden-Württemberg through the bwHPC initiative, which is supported by the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG.

Conflict of Interest: none declared.

References

Bachmann,J. *et al.* (2011) Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, 7, 516.

Bratsun,D. *et al*. (2005) Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. USA*, **102**, 14593–14598.
Culshaw,R.V. and Ruan,S. (2000) A delay-differential equation model of HIV infection of CD4(+) T-cells. *Math. Biosci*., **165**, 27–39.
Gurney,W.S. and Nisbet,R.M. (1980) Age- and density-dependent population dynamics in static and variable environments. *Theor. Popul. Biol*., **17**, 321–344.
Hass,H. *et al*. (2019) Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, **35**, 3073–3082.
Jensen,M.H. *et al*. (2003) Sustained oscillations and time delays in gene expression of protein Hes1. *FEBS Lett*., **541**, 176–177.
Kitano,H. (2005) International alliances for quantitative modeling in systems biology. *Mol. Syst. Biol*., **1**, 2005.0007.
Kreutz,C. *et al*. (2007) An error model for protein quantification. *Bioinformatics*, **23**, 2747–2753.
Kreutz,C. *et al*. (2012) Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Syst. Biol*., **6**, 120.
MacDonald,N. (1978) *Time Lags in Biological Models*. Springer, Berlin.
Mackey,M.C. and Glass,L. (1977) Oscillation and chaos in physiological control systems. *Science*, **197**, 287–289.
Maiwald,T. *et al*. (2016) Driving the model to its limit: profile likelihood based model reduction. *PLoS One*, **11**, e0162366.
May,R.M. (2001) *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton.
Merkle,R. *et al*. (2016) Identification of cell type-specific differences in erythropoietin receptor signaling in primary erythroid and lung cancer cells. *PLoS Comput. Biol*., **12**, e1005049.
Müller,K. *et al*. (2013a) A red/far-red light-responsive bi-stable toggle switch to control gene expression in mammalian cells. *Nucleic Acids Res*., **41**, e77.
Müller,K. *et al*. (2013b) Multi-chromatic control of mammalian gene expression and signaling. *Nucleic Acids Res*., **41**, e124.
Nelson,P.W. and Perelson,A.S. (2002) Mathematical analysis of delay differential equation models of HIV-1 infection. *Math. Biosci*., **179**, 73–94.
Peifer,M. *et al*. (2014) Analyzing effective models: an example from JAK/STAT5 signaling. arXiv: 1410.6341 [physics, q-bio].
Raue,A. *et al*. (2015) Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, **31**, 3558–3560.
Raue,A. *et al*. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929.
Smith,H. (2011) *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer Verlag, New York.
Sobotta,S. *et al*. (2017) Model based targeting of IL-6-induced inflammatory responses in cultured primary hepatocytes to improve application of the JAK inhibitor ruxolitinib. *Front. Physiol*., **8**., 775.
Srividhya,J. *et al*. (2007) The effects of time delays in a phosphorylation-dephosphorylation pathway. *Biophys. Chem*., **125**, 286–297.
Steiert,B. *et al*. (2012) Experimental design for parameter estimation of gene regulatory networks. *PLoS One*, **7**, e40052.
Steiert,B. *et al*. (2019) Recipes for analysis of molecular networks using the Data2Dynamics modeling environment. In: Hlavacek,W.S. (ed.) *Modeling Biomolecular Site Dynamics*. Springer: New York, pp. 341–362.
Swameye,I. *et al*. (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. USA*, **100**, 1028–1033.