the FEBS Journal

MINIREVIEW

Systems biology: model based evaluation and comparison of potential explanations for given biological data

Gunnar Cedersund[1] and Jacob Roll[2]

[1]Department of Cell Biology, Linköping University, Sweden
[2]Department of Electrical Engineering, Linköping University, Sweden

Keywords
data analysis; explanations; hypothesis testing; mathematical modeling; statistical testing; systems biology

Correspondence
G. Cedersund, Department of Cell Biology, Linköping University, SE58185 Linköping, Sweden
Fax: +46 (0)13 149403
Tel: +46 (0)702 512323
E-mail: gunnar@ibk.liu.se

(Received 8 April 2008, revised 23 November 2008, accepted 8 December 2008)

doi:10.1111/j.1742-4658.2008.06845.x

Systems biology and its usage of mathematical modeling to analyse biological data is rapidly becoming an established approach to biology. A crucial advantage of this approach is that more information can be extracted from observations of intricate dynamics, which allows nontrivial complex explanations to be evaluated and compared. In this minireview we explain this process, and review some of the most central available analysis tools. The focus is on the evaluation and comparison of given explanations for a given set of experimental data and prior knowledge. Three types of methods are discussed: (a) for evaluation of whether a given model is sufficiently able to describe the given data to be nonrejectable; (b) for evaluation of whether a slightly superior model is significantly better; and (c) for a general evaluation and comparison of the biologically interesting features in a model. The most central methods are reviewed, both in terms of underlying assumptions, including references to more advanced literature for the theoretically oriented reader, and in terms of practical guidelines and examples, for the practically oriented reader. Many of the methods are based upon analysis tools from statistics and engineering, and we emphasize that the systems biology focus on acceptable explanations puts these methods in a nonstandard setting. We highlight some associated future improvements that will be essential for future developments of model based data analysis in biology.

Introduction

It is open to debate as to whether the new approaches of systems biology are the start of a paradigm shift that will eventually spread to all other fields of biology as well, or whether they will stay within a subfield. Without a doubt, however, these approaches have now become established alternatives within biology. This is demonstrated, for example, by the fact that most biological journals now are open to systems biology studies, that several new high-impact journals are solely devoted to such studies [1], and that much research funding is directly targeted to systems biology [2].

Although the precise definition of systems biology is still debated, several characteristic features are widely acknowledged [3–5]. For example, the experimental data should reflect the processes of the intact system rather than that of an isolated component. Of more focus in this minireview, however, are features related to the interpretation of the data. Advanced data interpretation is often conducted using methods inspired by other natural sciences, such as physics and engineering,

Abbreviations
AIC, Akaike information criterion; BIC, Bayesian information criterion; IR, insulin receptor.

even though such methods usually need to be adopted to the special needs of systems biology. These methods, which usually involve mathematical modeling, allow one to focus more on the explanations deduced from the information-rich data, rather than on the data itself.

The strong focus on the nontrivially deduced explanations in a systems biology study is in close agreement with the general principles of scientific epistemology. However, as we will argue in several ways, this focus is nevertheless a feature that distinguishes systems biology from both more conventional biological studies and from typical hypothesis testing studies originating in statistics and engineering.

The general principles of scientific epistemology have been eloquently formulated by Popper and followers [6–8]. Importantly, as pointed out by Deutsch, Popper's principle of argument has replaced the need for a principle of induction [8]. Basically, the principle of argument means that one seeks the 'best' explanation for the currently available observations, even though it is also central that explanations can never be proved, but only rejected. The problem of evaluating and comparing two or several explanations for a given set of data and prior knowledge, so as to identify the best available explanation(s), is the focus of this minireview.

The basic principles of Popper *et al.* are more or less followed also in conventional biological studies. Nevertheless, in a systems biology study, more effort is devoted to the analysis of competing nontrivial explanations, based on information that is not immediately apparent from the data. For example, in the evaluation of the importance of re-cycling of STAT5 [9–11], a primary argument for the importance of this recycling was based on a model based analysis of the information contained in the complex time-traces of phosphorylated and total STAT5 in the cytosol. A more conventional biological approach to the same problem would be to block the recycling experimentally and compare the strength of the response before and after the blocking [9]. Generally, one could say that a conventional biological study typically seeks to find an experimental technique that directly examines the differences between two competing explanations, but that a systems biology study may distinguish between the two explanations without such direct experimental tests, using mathematical modeling. In other words, the emphasis in systems biology is on the explanations rather than on the available experimental techniques and the data themselves.

Similarly, even though the methods for hypothesis testing in statistics are based on the principles of Popper *et al.*, it could be argued that systems biology focuses even more on the explanations *per se*. As we review below, statistical testing is primarily oriented around the ability of an explanation to make predictions, and the central questions concern those explanations that would be expected to give the best prediction in a future test experiment. In a systems biology study, on the other hand, the best explanation should also fulfil a number of other criteria. In particular, the explanation should be based on the biological understanding for the system, and all its deduced features should be as realistic as possible, given what is known about the system from other sources than those included in the given data sets. In other words, the structure of the model should somehow reflect the underlying mechanisms in the biological system. We denote such a model a mechanistic model. Nevertheless, the theories and methods from statistics are very useful also in a systems biology context because they directly fit into the framework of mathematical modeling, which is the framework in which competing explanations typically are evaluated.

The most central question in this minireview is therefore 'What is the best explanation(s) to the given data and prior knowledge?'. We suggest and discuss methods for analysing this question through a number of related sub-problems. Possible results from these methods are outlined in Fig. 1. We start off by reviewing how a potential explanation (i.e. a hypothesis) can be reformulated into one or several mathematical models. Then we review methods from statistical testing that examine whether a single model can be rejected based on a lack of agreement with the available data alone. After that, we review methods for comparison of the predictive ability of two models, and finally suggest a scheme for the general comparison of two or more models. In the subsequent sections ('Rejections

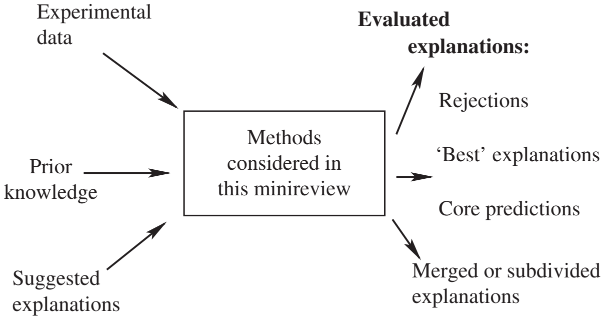

Fig. 1. The kind of methods reviewed in the present minireview: analysis of given explanations for a given set of experimental data and prior knowledge.

]

based on a residual analysis' and 'Rejection because another model is significantly better'), which are the most theory intensive sections, we start by giving a short conceptual introduction that is intended for people with less mathematical training (e.g. biologists/experimentalists). Also, following this idea, we will start with a short example serving as conceptual introduction to the whole article.

Introductory example

The example is concerned with insulin signaling, and is inspired by the developments in [12]. Insulin signaling occurs via the insulin receptor (IR). The IR signaling processes may be inspected experimentally by following the change in concentration of phosphorylated IR (denoted IR·P), and a typical time-series is presented as vertical lines (which gives one standard deviation, with the mean in the middle) in Fig. 2. As is clear from the figure, the degree of phosphorylation increases rapidly upon addition of insulin (100 nM at time zero), reaches a peak value within the first minute, and then goes down again and reaches a steady-state value after 5–10 min. This behavior is referred to as an overshoot in the experimental data. These data are one of the three inputs needed for the methods in this minireview (Fig. 1).

The second input in Fig. 1 is prior knowledge. For the IR subsystem this includes, for example, the facts that IR is phosphorylated much more easily after binding to insulin and that the phosphorylation and dephosphorylation occurs in several catalysed steps. It is also known that IR may leave the membrane and enter the cytosol, a process known as internalization. The internalization may also be followed by a return to the membrane, which is known as recycling.

The final type of input in Fig. 1 concerns suggested explanations. In systems biology, an explanation should both be able to quantitatively describe the experimental data, and do so in a way that does not violate the prior knowledge (i.e. using a mechanistic model). However, it is important to note that a mechanistic model does not have to explicitly include all the mechanisms that are known to occur. Rather, modeling is often used to achieve a characterization of which of these mechanisms that are significantly active, and independently important, and which mechanisms are present but not significantly and/or uniquely contributing to the experimentally observed behavior. For example, it is known that there is an ongoing internalization and recycling, but it is not known whether this is significantly active already during the first few minutes in response to insulin, and it is only the first few minutes that are observed in the experimental data. Therefore, it is interesting to consider explanations for these data that contain recycling and then to compare these with corresponding explanations that do not include recycling. Examples of two such alternative suggested explanations are given in Fig. 3.

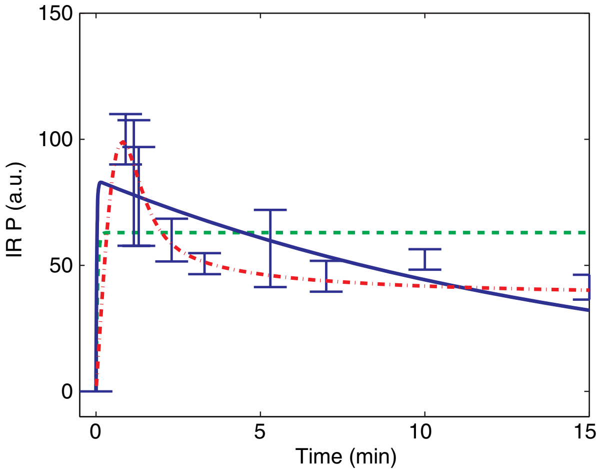

Fig. 2. Experimental data and simulations corresponding to the introductory example. This minireview deals with methods for a systematic comparisons between such experimental and simulated data series. The result of these methods is an evaluation and comparison of the corresponding explanations. Importantly, this allows for mechanistic insights to be drawn from such experimental data that would not be obtained without modeling.

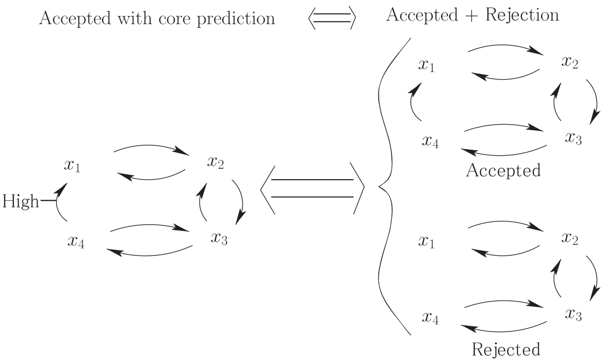

Fig. 3. To the right, two of the models for the insulin signaling example in the introductory example are depicted. The top one includes both internalization and recycling after dephosphorylation, but not the lower one. The figure to the left corresponds to a discussion on core predictions in the section 'A general scheme for comparison between two models'. It depicts a model with internalization and recycling, where the core prediction shows that the recycling must have a high (nonzero) rate; this of course corresponds to the rejection conclusion to the right. x_1 and x_2 corresponds to unphosphorylated and phosphorylated IR, respectively, and x_3 and x_4 corresponds to internalized phosphorylated and dephosphorylated IR, respectively.

With all inputs established, the methods in this review can be applied to achieve the outputs displayed in Fig. 1. The first step is to translate the graphical drawings in Fig. 3 to a mathematical model ('Reformulation of a hypothesis into a mathematical model'). This is the step that allows for a systematic, quantitative, and automatic analysis of many of the properties that are implied by a suggested explanation. The second step ('Rejection because another model is significantly better') evaluates whether the resulting models are able to describe the experimental observations in a satisfactory manner. This is typically carried out by evaluating the differences between the model predictions and the experimental data for all time-points (referred to as the residuals) and there are several alternatives for doing this. For the present example, such an analysis shows that the given explanation with both internalization and recycling cannot be rejected (Fig. 2, red, dash-dotted line). The analysis also shows that sub-explanations lacking the internalization can not display the overshoot at all (green, dashed), and that the resulting model with internalization but without recycling can not display an overshoot with a sufficiently similar shape (blue, solid) [12]. Nevertheless, the hypothesis with internalization but without recycling is not completely off, and is therefore interesting for an alternative type of analysis as well ('Rejection because another model is significantly better'). This type of analysis analyses whether the slightly better model (here, the one with both internalization and recycling) is significantly better than a worse one (here, the one without recycling). The final step analyses the surviving explanations, and decides how to present to results. This step is presented in the penultimate section ('A general scheme for comparison between two models'), which also includes a deeper discussion of how the methods in this minireview can be combined.

Reformulation of a hypothesis into a mathematical model

As mentioned in the Introduction, the main focus of this article is to evaluate competing explanations for a given data set and prior knowledge. We will now introduce the basic notation for this data set, and for the mathematical formulation of the potential explanation. The most important notation has been standardized in this and the two accompanying reviews, and is summarized in Table 1.

The data set consists of data points, which are distinguished according to the following notation:

$$y_i(t_j) \tag{1}$$

where t_j is the time the data point was collected, and i is the index vector specifying the other details of the measurement. This index vector could for example

Table 1. Overview of mathematical symbols that are shared in all three minireviews [present review, 17, 56].

Meaning	Symbol	Comment
Dynamic state variables	x	Typically, x correspond to concentrations
Time dependency of state variables	$\dot{x} = f(x, p, u)$	The dynamics is described via ordinary differential equations
Parameters	p	p_x and p_y are common subsets, and they are concerned with the state dynamics and the measurements, respectively
Estimated parameters	$\hat{p}$, $\hat{p}_x$, $\hat{p}_y$	Typically generated by minimizing a cost function, V
Input function	u,	External perturbations on the studied system
Observational function	$g(x,p,u)$	Link in the model between dynamic states and experimental observations
Model prediction after parameter estimation	$\hat{g}$, $g(x, \hat{p}, u), \hat{y}$	
Measurements, data points	y	Typically, we assume that $y = g(x,p) + \varepsilon$ if the model structure and the parameters are 'true'
Measurement noise	ε	Typically, we assume that $\varepsilon = y - g(x, \hat{p}, u)$ (i.e. that there is no noise in the dynamic equations)
Noise standard deviation	σ	The variance is denoted by σ^2
Residual	e	Typically, $e = y - g(x, \hat{p}, u)$
Model structure	$\mathcal{M}$	
Time	t	
Total number of measurements	N	
Cost functions	V	This represents the total difference between the model predictions and the data + prior knowledge
Statistical expectation	$\langle \ldots \rangle$, $E(\cdot)$	Expected value for random variables

contain information about which signal (e.g. concentration of a certain substance) that has been measured, which experiment the measurement refers to, or which subset of data (e.g. estimation or validation data) that the measurement point belongs to. In many cases, some indexes will be superfluous and dropped, simplifying the notation $y(t)$. The N data points are collected in the time series vector Z^N. Finally, it should be noted that some traditions uses the concept 'data point' to denote all the data that have been collected at a certain time point [13].

Now consider a potential explanation for this data set. Let the explanation be denoted $\mathcal{M}$. We will sometimes refer to such a 'potential explanation' as a 'hypothesis'. These two expressions can be used interchangeably, but the first option will often be preferred because it highlights the fact that a successful hypothesis must not only be able to mimic the data, but also be able to provide a biologically plausible explanation with respect to the prior knowledge about the system. A potential explanation $\mathcal{M}$ must also be able to produce predicted data points corresponding to the experimental data points in Z^N. Note that this is a requirement that typically is not fulfilled by a conventional biological explanation, which often is comprises verbal arguments, or nonquantitative interaction maps, etc. A predicted data point corresponding to (1) and the hypothesis $\mathcal{M}$ will be denoted:

$$\hat{y}_i^{\mathcal{M}}(t_j, p) \tag{2}$$

where the symbol p denotes the parameter vector. Generally, a model structure is a mapping from a parameter set to a unique model (i.e. to a unique way of predicting outputs). A hypothesis $\mathcal{M}$ that fulfils (2) is therefore associated with a model structure, which also will be denoted $\mathcal{M}$. A specific model will be denoted $\mathcal{M}(p)$.

The problem of formulating a mathematical model structure from a potential biological explanation has been treated in many text books [4,14], and will not be discussed in depth here. All the examples we consider below will be dynamic, where the model structure will be in the form of a continuous-time deterministic state-space model:

$$\dot{x} = f(x, p, u) \tag{3a}$$

$$\hat{y} = g(x, p, u) \tag{3b}$$

$$x(0) = x_0 \tag{3c}$$

where x is the n-dimensional state vector (often corresponding to concentrations), $\dot{x}$ is the time-derivative of this vector, $x(t)$ is the state at time t, and f and g are vectors of smooth nonlinear functions. The symbol u denotes the external input to the system. The inputs may be time-varying, and can for example correspond to a ligand concentration. Note that the inputs are, just like the parameters, not themselves effected by the dynamic equations. Note also that the parts of the potential explanation that refer to the biological mechanisms are contained in f, and that the parts that refer to the measurement process are contained in g. Note, finally, that the parameter vector x_0 is a part of the parameter vector p.

Finally, one important variation is the replacement of time-variation for steady state. There is no major difference between these cases. This can be understood by choosing time-points for t_i that are so large that the transients have passed. Therefore, almost all results and methods presented in this minireview are applicable to steady-state data and models as well.

Rejections based on a residual analysis

Conceptual introduction

We now turn to the problem of evaluating a single hypothesis $\mathcal{M}$ with respect to the given data Z^N. From the introduction of $\mathcal{M}$ above, an obviously important entity to consider for the evaluation of $\mathcal{M}$ is the difference between the measured and predicted data points. We denote such a difference e:

$$e^{\mathcal{M}}(t, p) := y(t) - \hat{y}^{\mathcal{M}}(t, p)$$

and it is referred to as a residual. Residuals are depicted in Fig. 4. If the residuals are large, and especially if they are large compared to the uncertainty in the data, the model does not provide a good explanation for the data. The size of the residuals is tested in a χ^2 test, which is presented in a subsequent section. Likewise, if a large majority of the residuals are similar to their neighbours (e.g. if the simulations lie on the same side of the experimental data for large parts of the data set), the model does not explain the data in an optimal way. This latter property is tested by methods given in a subsequent section. The difference between the two types of tests is illustrated in Fig. 4. Tests such as the χ^2 test, which analyses the size of the residuals, would typically accept the right part of the data series, but reject the left one, and correlation-based methods such as the whiteness or run test, would typically reject the left part, but accept that to the right.

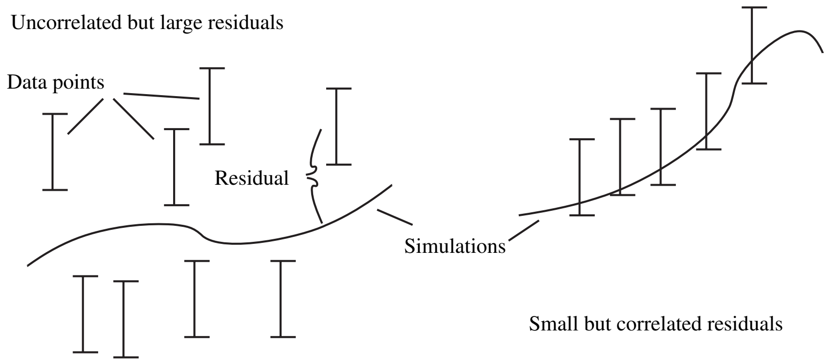
Uncorrelated but large residuals
Data points
Residual
Simulations
Small but correlated residuals

Fig. 4. Two sections of experimental data series and simulations. The data points y are shown with one standard deviation. As can be seen on the left, the simulations lie outside the uncertainty in the data for all data points. Nevertheless, they lie on both sides of the simulation curve, and with no obvious correlation. Conversely, the second part of the data series shows a close agreement between the data and simulations, but all data points lie on the same side of the simulations. Typically, situations like that on the left are rejected by a χ^2 test but pass a whiteness test, and situations such as that on the right pass a χ^2 test but would be rejected by a whiteness test.

The null hypothesis: that the tested model is the 'true' model

We now turn to a more formal treatment of the subject. A common assumption in theoretical derivations [13] is that the data has been generated by a system that behaves like the chosen model structure for some parameter, p^0, and for some realization of the noise $\varepsilon(t)$:

$$y(t_i) = \hat{y}^{\mathcal{M}}(t_i, p^0) + \varepsilon(t_i) \quad \forall i \in [1, N] \tag{4}$$

If the $\varepsilon(t)$s are independent, they are sometimes also referred to as the innovations because they constitute the part of the system that never can be predicted from past data. It should also be noted that the noise here is assumed to be additive, and only affects the measurements. In reality, noise will also appear in the underlying dynamics, but adding noise to the differential equations is still unusual in systems biology.

The assumption of Eqn (4) can also be tested. According to the standard traditions of testing, however, one cannot prove that this, or any, hypothesis is correct, but only examine whether the hypothesis can be rejected [6,15]. In a statistical testing setting, a null hypothesis is formulated. This null hypothesis corresponds to the tested property being true. The null hypothesis is also associated with a test entity, $\mathcal{T}$. The value of $\mathcal{T}$ depends on the data Z^N. If this value is above a certain threshold, $\delta_{\mathcal{T}}$, the null hypothesis is rejected, with a given significance α_δ [15]. Such a rejection is a strong statement because it means that the tested property with large probability does not hold, which in this particular case means that the tested hypothesis $\mathcal{M}$ is unable to provide a satisfactory explanation for the data. On the other hand, if $\mathcal{T} < \delta_{\mathcal{T}}$, one simply says that the test was unable to reject the potential explanation from the given data, which is a much weaker statement. In particular, one does not claim that failure to reject the null hypothesis means that it is true, (i.e. that $\mathcal{M}$ is the best, or correct, explanation). Nevertheless, passing such a test is a positive indication of the quality of the model.

Identification of $\hat{p}$

Below, we introduce the probably two most common ways for testing Eqn (4): a χ^2 test and a whiteness test. Both of these two tests evaluate the model structure $\mathcal{M}$ at a particular parameter point, $\hat{p}$. This parameter point corresponds to the best possible agreement between the model and the part of the data set chosen for estimation, Z^N_{est}, according to some cost function V, which measures the agreement between the model output and the measurements. The $\hat{p}$ vector thus serves as an approximation of p^0. A common choice of cost function is the sum of the squares of the residuals, typically weighted with the variance of the experimental noise, σ^2. This choice is motivated by its equivalence to the method of maximum likelihood [if $\varepsilon(t) \in N(0,\sigma^2(t))$], which has minimum variance to a unbiased parameter estimate and many other sound properties [13]. The likelihood function is very central in statistical testing; it is denoted $\mathcal{L}$, and gives a measure of the likelihood (probability) that the given data set should be generated by a given model $\mathcal{M}(p)$.

Another important concept regarding parameter estimation is known as regularization [15]. Regularization is applicable (e.g. if one has prior knowledge about certain parameter values), but can also be used as a way of controlling the flexibility of the model. Certain regularization methods [15,16] can also be used for regressor selection. The main idea of regularization is to add an extra term to the cost function, which penalizes deviations of the parameters from some given nominal values. Together with a quadratic norm cost function, the estimation criterion takes the form:

$$\widehat{p} := \arg\min V(p) \tag{5}$$

$$V(p) := \frac{1}{N} \sum_{i \in Z_{\text{est}}^N} \sum_j \frac{(y_i(t_j) - \widehat{y}_i^M(t_j))^2}{\sigma_i^2(t_j)} + \sum_k \alpha_k h_{\text{pen}}(p_k - p_k^g) \tag{6}$$

Here, p_k^g is the nominal value of p_k, and $h_{\text{pen}}(\cdot)$ is a suitable penalizing function [e.g., $h_{\text{pen}}(p) = p^2$ (ridge regression) or $h_{\text{pen}}(p) = |p|$] and the α_ks are the weights to the different regularization terms. Further information about the identification process is included in a separate review in this minireview series [17].

Testing the size of the residuals: the χ^2 test:

With all the notations in place, Eqn (4) together with the hypothesis that $p^0 = \widehat{p}$ can be re-stated as:

$$e^{\mathcal{M}}(t_j, \widehat{p}) \text{ follows the same distribution as } \varepsilon(t_j) \quad \forall t \in [1, N] \tag{7}$$

which is a common null hypothesis. The most obvious thing one can do to evaluate the residuals is to plot them and to calculate some general statistical properties, such as maximum and mean values, etc. This will give an important intuitive feeling for the quality of the model, and for whether it is reasonable to expect that Eqn (7) will hold, and that $\mathcal{M}$ is a nonrejectable explanation for the data. However, for given assumptions of the statistical properties of the experimental noise $\varepsilon(t)$, it is also possible to construct more formal statistical tests. The easiest case is the assumption of independent, identically distributed noise terms following a zero mean normal distribution, $\varepsilon(t) \in N(0,\sigma^2(t))$. Then, the null hypothesis implies that each term $(y(t) - \widehat{y}(t,p))/\sigma(t)$ follows a standard normal distribution, $N(0,1)$, and this in turn means that the first sum in Eqn (6) should follow a χ^2 distribution [18]; this sum is therefore a suitable test function:

$$\mathcal{T}_{\chi^2} = \sum_{i,j} \frac{(y_i(t_j) - \widehat{y}_i^M(t_j))^2}{\sigma_i^2(t_j)} \in \chi^2(d) \tag{8}$$

and it is commonly referred to as the χ^2 test. The symbol d denotes the degrees of freedom for the χ^2 distribution, and this number deserves some special attention. In case the test is performed on independent validation data, the residuals should be truly independent, and d is equal to N_{val}, the number of data points in the validation data set, Z_{val}^N [19,20]. Then the number d is known without approximation.

A common situation, however, is that one does not have enough data points to save a separate data set for validation (i.e. that both the parameter estimation and the test are performed on the same set of data, Z^N). Then one might have the problem of over-fitting. For example, consider a flexible model structure that potentially could have $e = 0$ for all data points in the estimation data. For such a model structure, $\mathcal{T}_{\chi^2}$ could consequently go to zero, even though the chosen model might behave very poorly on another data set. This is the problem of over-fitting, and it is discussed further later in this minireview. In this case, the residuals cannot be assumed to be independent. In summary, this means that if $Z_{\text{test}}^N = Z_{\text{est}}^N$, one should replace the null hypothesis of Eqn (7) by Eqn (4), and find a distribution other than $\chi^2(N_{\text{val}})$ for the χ^2 test if Eqn (8).

If the model structure is linear in the parameters, and all parameters are identifiable, each parameter that has been fitted to the data can be used to eliminate one term in Eqn (8), i.e. one term [e.g. $(y_1(t_4) - \widehat{y}_1(t_4))^2/\sigma_1^2(t_4)$] can be expressed using the other terms and the parameters. When all parameters have been used up, the remaining terms are again normally distributed and independent. This means that the degrees of freedom can then be chosen as:

$$d = N - r \quad \text{where } r = \dim(p) \tag{9}$$

This result is exact and holds, at least locally, also for systems that are nonlinear in the parameters, such as Eqn (3) [19,20]. Note that this compensation with r is performed for the same reason as why the calculation of variance from a data series has a minus one in the denominator, if the mean value has been calculated from the data series as well.

However, Eqn (9) does not hold for unidentifiable systems (i.e. where the data is not sufficient to uniquely estimate all parameters). This is especially the case if some parameters are structurally unidentifiable [i.e. if they can analytically be expressed as a function of the other parameters without any approximation of the predicted outputs $\widehat{y}(t,p)$]. The number of parameters that is superfluous in this way is referred to as the transcendence degree [21]. We denote the transcendence degree by t_M, which should not be confused with the index notation on the time-vector. With this notation, we can write a more generally applicable formula for d as:

$$d = N - (r - t_M) \tag{10}$$

This compensation for structural unidentifiability should always be carried out, and is not a matter of design of the test. However, when considering practical identifiability, the situation is more ambiguous [19,20]. Practical identifiability is a term used for example by Dochain and Vanrolleghem [22], and it is concerned with whether parameters can be identified with an acceptable uncertainty from the specific given data set, given its noise level and limited number of data points, etc. Practical unidentifiability is very common for systems biology problems; this means that there typically are many parameters that do not uniquely contribute to the estimation process, even after eliminating the structurally unidentifiable parameters. If this problem leads to a large discrepancy between the number of practically identifiable parameters and r–t_M, and especially if N–(r–t_M) is approximately equal to the number of data points, Eqn (10) in Eqn (8) results in an unnecessarily difficult test to pass. A more fair test would then include a compensation of the number of practically identifiable parameters (i.e. the effective number of parameters, $A_{\mathcal{M}}$). One way to estimate this number is through the following expression [15]:

$$A_{\mathcal{M}} = \sum_k \frac{\lambda_k}{\lambda_k + \alpha_k} \tag{11}$$

where λ_i is the ith eigenvalue to the Hessian of the cost function, and where the α_is are the regularization weights for ridge regression, or some otherwise chosen cut-off values. The best expression for d in Eqn (8) applied to a systems biology model, where $Z^N_{\text{val}} = Z^N_{\text{est}}$, is thus probably given by:

$$d = N - A_{\mathcal{M}} \tag{12}$$

Note, however, that this final suggestion is not exact, and includes the design variables α_k.

Example 1

To illustrate the various choices of d, and especially to illustrate the potential danger of only considering structural unidentifiability, we first consider the simple, but somewhat artificial, model structure in Fig. 5. Assuming mass action kinetics, and that all the initial mass is in states x_1 and $x_{2,1}$, the corresponding set of differential equations are:

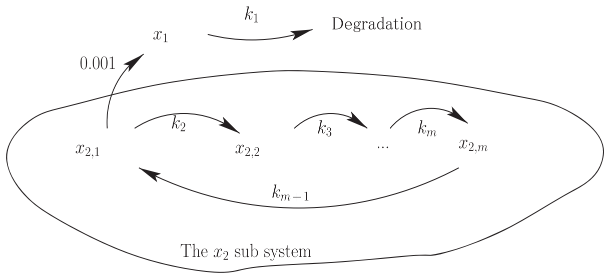

Fig. 5. The model structure examined in Example 1. The key property of this system is that all parameters are structurally identifiable (after fixing one of them to a specific value), but that only one parameter, k_1, is practically identifiable.

$$\dot{x}_1 = -k_1 x_1 + 0.001 x_{2,1} \tag{13a}$$

$$\dot{x}_{2,1} = -k_2 x_{2,1} + k_{m+1} x_{2,m} - 0.001 x_{2,1} \tag{13b}$$

$$\dot{x}_{2,2} = -k_3 x_{2,2} + k_2 x_{2,1} \tag{13c}$$

$$\vdots$$

$$\dot{x}_{2,m} = -k_{m+1} x_{2,m} + k_m x_{2,(m-1)} \tag{13d}$$

$$y = x_1 \tag{13e}$$

$$x(0) = (10, 10, 0, 0, \ldots) \tag{13f}$$

Here m is a positive integer, determining the size of the x_2 subsystem. This means that m also determines the number of parameters, and thus, in some ways, the complexity of the model structure. Note, however, that the x_2 subsystem only exerts a very small effect on the x_1 dynamics, which is the only measurable state.

Let us now consider the result of estimating and evaluating this model structure with respect to the data in Fig. 6. The results are given in Table 2 for the different options of calculating d. The details of the calculations are given in the MATLAB-file Example1.m, except for the calculations of the transcendence degree which are given in the Maple file Example1.mw, using the Sedoglavic' algorithm [21] (see Doc. S1). In the example, the data have been generated by the tested model structure, which means that the model should pass the test. However, when calculating d according to Eqn (9) or Eqn (10), the test erroneously rejects the model structure, and does so with a high significance. This follows from the fact that all parameters in the x_2 subsystem are practically unidentifiable, even though they are structurally identifiable ($t_M = 0$), and the fact that the r–t_M is approximately equal to the number of data points N.

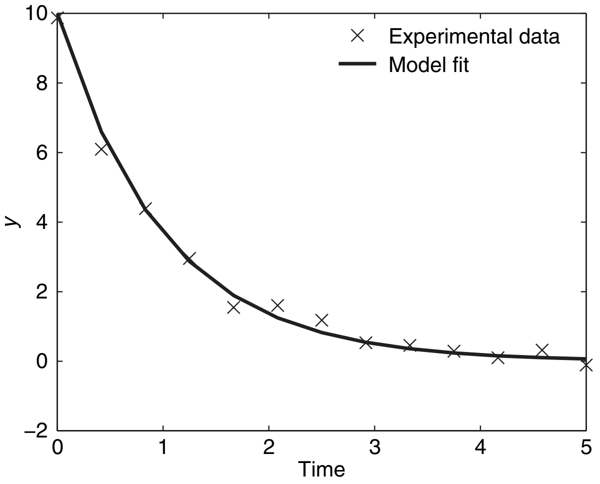

Fig. 6. The data used in Example 1. The whole data set is used for both estimation and validation/testing.

Table 2. The values from Example 1 illustrating the importance of choosing an appropriate *d* in Eqn (8).

d-formula	N	m	d value	$\delta_{\chi^2}(95\%)$	$\mathcal{T}$	Pass?
N	13	11	13	22.36	8.15	Yes
$N - r$	13	11	1	3.84	8.15	No
$N - (r - t_M)$	13	11	1	3.84	8.15	No
$N - A_{\mathcal{M}}$	13	11	12	21.02	8.15	Yes

In this example, it is straightforward to see that the parameters in the x_2 subsystem have no effect on the observed dynamics, and thus are practically unidentifiable; it is apparent from the factor 0.001 in Eqn (13a). However, the situation highlighted by this example is common. As another example one could consider the models of Teusink *et al*. [23] or Hynne *et al*. [24] for yeast glycolysis. They are both of a high structural identifiability ($t_M < 10$), even when only a few states can be observed, but have many parameters ($r > 50$) and only a handful of them are practically identifiable with respect to the available *in vivo* measurements of the metabolites [25,26]. Therefore, if one does not have access to a large number of data points (especially if $N < 50$), a χ^2 test would be impossible to pass, using $d = N{-}(r{-}t_M)$, even for the 'true' model. Note, however, that this problem disappears when N is large compared to $r{-}t_M$.

Testing the correlation between the residuals

Although the χ^2 test of Eqn (8) is justified by an assumption of independence of the residuals, it primarily tests the size of the residuals. We will now look at two other tests that more directly examine the correlation between the residuals.

The first test is referred to as the run test. The number of runs R_u is defined as the number of sign changes in the sequence of residuals, and it is compared to the expected number of runs, $N/2$ (because it is assumed that the mean of the uncorrelated Gaussian noise is equal to zero) [22]. An assessment of the significance of the deviation from this number is given by a comparison of:

$$\frac{R_u - N/2}{\sqrt{N/2}}$$

and the cumulative $N(0, 1)$ distribution for large N and a cumulative binomial distribution for small N [22].

The second test is referred to as a whiteness test. Its null hypothesis is that the residuals are uncorrelated. The test is therefore based on the correlation coefficients $\mathcal{R}(\tau)$, which are defined as:

$$\mathcal{R}_i(\tau) := \frac{1}{N_i} \sum_{j=1}^{N_i} e_i(t_j) e_i(t_{j-\tau})$$

where N_i is the number of data points with index i. Using these coefficients, one may now test the null hypothesis by testing whether the test function $\mathcal{T}_{\mathrm{white}}$ follows a χ^2 distribution [22]:

$$\mathcal{T}_{\mathrm{white}} := \frac{N}{\mathcal{R}(0)^2} \sum_{\tau=1}^{M} \mathcal{R}(\tau)^2 \in \chi^2(M)$$

Rejection because another model is significantly better

Conceptual introduction

In the previous section, we looked at tests for a single model. These tests can of course be applied to several competing models as well. Because models will typically result in different test values, these already mentioned test functions can in principle be used to compare models. However, it would then not be known whether a model with a lower test value is significantly better, or whether the difference lies within the difference in test values that would be expected to occur also for equally good models. We will now review some other statistical tests that are especially developed for the model comparison problem.

As demonstrated above, the sum of the normalized residuals can be expected to follow a χ^2 distribution.

This insight lead to a very straightforward χ^2 test, which simply compares the calculated sum with the threshold value for the appropriate distribution. This is easy because the distribution is known analytically. A similar distribution has been derived for the difference between the sums of two such models. It also follows a χ^2 distribution. A very straightforward test is therefore to simply calculate this difference, and compare it with an appropriate χ^2 distribution. This is the basis behind the likelihood ratio test described below.

However, in the derivation of the likelihood ratio test, a number of conditions are assumed, and these conditions are typically not fulfilled. Therefore, a so-called bootstrap-based approach is advisable, even though it is much more computationally expensive. The basic principle behind this approach is depicted in Fig. 7. Here, each green circle corresponds to the cost (i.e. sum of residuals) for both the models, when the data have been generated under the assumption that model 1 is correct, and when both models have been fitted to each generated data set. Likewise, the blue Xs corresponds to the costs for both models, when the data have been generated under the assumption that model 2 is correct. As would be expected, model 1 is always fitting the data well (i.e. there is a low cost) when model 1 has generated the data, but model 2 is less good at fitting to these data, and vice versa. Now, given these green and blue symbols, the following four situations can be distinguished for evaluation of the model costs for the true data (depicted as a red square). If the square ends up in the upper right corner, none of the models appear to be able to describe the data in an acceptable manner, and both models should be rejected. If the square ends up in the lower right or upper left corner, model 1 or model 2 can be rejected, respectively. Finally, if the red square ends up in the lower left corner, none of the models can be rejected. In Fig. 7, these four scenarios can be distinguished by eye but, for the general case, it might be good to formalize these decisions using statistical measures. This is the conceptual motivation for developing the approaches below, and especially the bootstrap approach described in a later section.

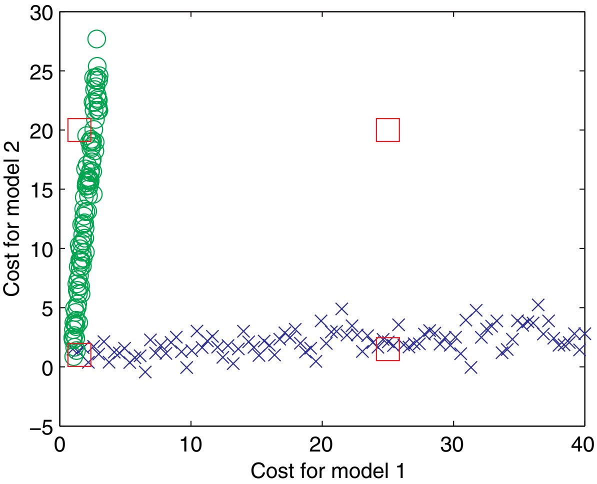

Fig. 7. The conceptual idea behind many model comparison approaches, especially those in the sections 'The *F* and the likelihood ratio test' and 'Bootstrap solutions'. The green circles correspond to the distribution under the hypothesis that model 1 is true, and the blue Xs correspond to the corresponding distribution under the hypothesis that model 2 is correct. The red squares correspond to the cost for four different scenarios, rejecting one, both, or none of the models. Adapted from Hinde [44].

The classical objective of statistical testing: minimization of the test error

Let us now turn to a more formal treatment of the subject of model comparison. The central property in statistical testing is the test error, *Err*. This is the expected deviation between the model and a completely new set of data, referred to as test data [15]. Ideally, one would therefore divide the data set into three separate parts: estimation data, validation data and test data (Fig. 8). Note that the test data are different from the validation data (strictly this only means that the data points are different, but the more fundamental and large these differences are, the stronger the effect of the subdivision). The reason for this additional subdivision is that the validation data might have been used as a part of the model selection process. In statistical testing, it is not uncommon to compare a large number of different models with respect to the same validation data, where all models have been estimated to the same estimation data. In such a case, it is apparent that $V(Z^N_{\mathrm{val}})$can be expected to be an under-estimation of the desired $Err = E(V(Z^N_{\mathrm{test}}))$, where E is the expectation operator. However, the same problem is to some extent also present if only two models are compared in this way.

Quite often, however, one does not have enough data to make such a sub-division. Then the test error *Err* has to be estimated in some other way, quite often based on the estimation data alone. In that case, it is

Estimation	Validation	Test

Fig. 8. Ideally, one should divide the given data set, Z^N, in three parts: one part Z^N_{est} for estimation, one part Z^N_{val} for validation, and one part Z^N_{test} for testing.

even more important that one does not equate *Err* with $V(Z^N) = V(Z^N_{\mathrm{est}})$, due to the problem of over-fitting. Over-fitting is most common when using highly flexible model structures because, in principle, they can give $V(Z^N_{\mathrm{est}}) = 0$ but still have a very large true *Err*. Because flexibility usually increases upon increasing model complexity, over-fitting is therefore also a problem of model selection.

One can also explain the problem of over-fitting by studying the trade-off between variance and bias. Then the test error is subdivided in its components [15]:

$$Err = Err_{\mathrm{irr}} + Err_{\mathrm{bias}} + Err_{\mathrm{var}} \tag{14}$$

In this equation, Err_{irr} denotes the irreducible part of the test error (i.e. the part that is due to the innovation component in the test data, Z^N_{test}). Thus, if $y_i(t_j) = y_i(t_j,p^0) + \varepsilon_i(t_j)$, where the $\varepsilon_i(t_j)$ are uncorrelated with zero mean and standard deviation $\sigma_i(t_j)$, we have:

$$Err_{\mathrm{irr}} = \frac{1}{N}\sum_{i,j} \sigma_i(t_j)^2. \tag{15}$$

and where the sum is taken over all i,j such that $y_i(t_j) \in Z^N$ test. The second term, Err_{bias}, is the square of the bias of the error (i.e. the square of the average difference between our estimated predictions and the true measurements). Expressed more formally, using the same assumptions as for Eqn (15), we have:

$$Err_{\mathrm{bias}} = \frac{1}{N} \sum_{y_i(t_j)\in Z^N_{\mathrm{test}}} \left([E\widehat{y}_i(t_j,\widehat{p}) - y_i(t_j,p^0)] \cdot [E\widehat{y}_i(t_j,\widehat{p}) - y_i(t_j,p^0)]\right)$$

The third term, Err_{var}, is the variance estimated predictions (i.e. a measure of how much the predictions would vary if the estimation data were collected again). Expressed more formally, with the same assumptions as for Eqn (15), we have:

$$Err_{\mathrm{var}} = E\left(\frac{1}{N} \sum_{y_i(t_j)\in Z^N_{\mathrm{test}}} ([\widehat{y}_i(t_j,\widehat{p}) - E(\widehat{y}_i(t_j,\widehat{p}))] \cdot [\widehat{y}_i(t_j,\widehat{p}) - E(\widehat{y}_i(t_j,\widehat{p}))])\right)$$

The important thing with respect to the subdivision of Eqn (14) is the dependency of the three terms Err_{irr}, Err_{bias} and Err_{var} on the complexity of the model. Typically, Err_{bias} decreases monotonously with model complexity, whereas Err_{var} increases with model complexity. Consequently, there is a model complexity where *Err* is minimal, even though the model agreement increases with increasing complexity; this insight is the other way of motivating the over-fitting problem.

There are two final concepts from the statistical testing tradition that need to be mentioned. The first is the concept of nested models. Two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, are nested if one can be obtained as a special case of another. This can be written as $\mathcal{M}_1 \subset \mathcal{M}_2$ or $\mathcal{M}_2 \subset \mathcal{M}_1$, if $\mathcal{M}_1$ or $\mathcal{M}_2$ is the smaller model, respectively, and typically the dependency can be formulated as a constraint on the parameters, which always is fulfilled for $\mathcal{M}_1$, but not necessarily for $\mathcal{M}_2$. For example, $\mathcal{M}_1$ could correspond to a model with a specific reaction described through an irreversible reaction, which, in $\mathcal{M}_2$, is described through reversible kinetics (all other parts are equal). Another example of nested models is given by the upper right and lower right model structures in Fig. 3. Most of the derivations for model comparison in the statistical testing tradition are derived for the case of nested models.

The other concept is referred to as in-sample error. This is the error *Err* for the special case of the test data being collected using the exact same 'external conditions' as for the estimation data. Specifically, this means that the data are collected at the same time-points, and that the controlled perturbations of the systems are performed in an identical manner [15]. The in-sample error is a convenient measure for model comparison, even though it is the extra-sample error that describes the future usage of the model in most cases. It is therefore common that one calculates the in-sample error, and uses this to approximate *Err* on a generic data set. This is the case, for instance, for the Akaike information criterion (AIC).

AIC and Bayesian information criterion (BIC) tests

There are many approaches to compare two or more models, with the attempt to identify the model that has the smallest expected test error *Err*. The perhaps most well-known of these methods is due to Akaike [27,28], and is often based on the following function:

$$\mathrm{AIC} = V(\widehat{p}) + \sigma^2 \frac{2d_p}{N} \tag{16}$$

where V is the quadratic norm cost function, σ^2 is the variance of the experimental noise, and where N is the number data points used for the test. The final symbol, d_p, represents model complexity, and, in the simplest cases, can be given by the dim (p) directly, but, for the more general case (nonlinear models, minimization using more regularization, unidentifiable systems, etc.), d_p should be replaced by some measure of the effective number of parameters, A_p; Eqn (11). Interestingly, the first term in Eqn (16) represents the cost function in the in-sample test error, Err_b, and the second term is referred to as the optimism, which thus represents the difference between the true in-sample test error and the cost function. It is important to note that there are several variations of AIC; for example, see the accompanying minireview on experimental design [56], and see also Doc. S2, specifying the relation between these expression.

A similar test entity, but that is derived in a Bayesian framework, is the [13]:

$$\mathrm{BIC} = V(\widehat{p}) + \frac{\log(N)}{N} d_p \tag{17}$$

where the same notations are used as for AIC.

For both AIC and BIC, the model with the lowest criterion value is the chosen model because this is the model that is expected to give the lowest test error. There is no guarantee that AIC and BIC will prefer the same model, and for $N > 7$, AIC has a bias towards more complex model structures [15].

The *F* and the likelihood ratio test

There exists many other tests similar to AIC and BIC, using more or less related test expressions. Some important examples include the minimum description length, Vapnik–Chervonenkis dimension, the final prediction error, and the general information criterion [15,22]. A shared problem among all these tests, however, is that they will only choose one single model as the preferred one, even though the compared models might perform similarly for all practical purposes (i.e. even though the difference between the models is insignificant). That means that these methods are primarily useful if one simply needs a single model to make a prediction, as in an engineering problem.

A test that does attach a significance to its choices is the likelihood ratio test. The test function, $\mathcal{T}_{lr}$, and the corresponding distribution under standard conditions is given by:

$$\mathcal{T}_{lr} = 2(l_1 - l_2) \in \chi^2(d_1 - d_2) \tag{18}$$

where l_i is the logarithm of the likelihood function for model $\mathcal{M}_i(\widehat{p}_i)$, and where d_i is given by $\dim(p_i){-}t_{\mathcal{M}_i}$ for $i = 1,2$.

The standard conditions for the likelihood ratio test are rather general, at least compared to the $\mathcal{T}_{\chi^2}$ test. The two most severe assumptions are that the models are assumed to be nested and that N, the number of data points, is assumed to be large [29–31]. If these two assumptions are fulfilled, the remaining assumptions are probably nonproblematic. For example it is, assumed that the estimated parameters follow a Gaussian uncertainty distribution, and this holds asymptotically for all likelihood minimizations under very general constraints (i.e. for sufficiently large N) [32]. Note that it is not necessary the measurement noise to be normally distributed or white, or that for the likelihood function to be given by any specific type of expression.

Despite this generality, the assumptions are still typically not fulfilled. For example, an estimated parameter might lie close to a boundary (i.e. 0), and thus making the distribution non-Gaussian. For this violation, if the other assumptions still are fulfilled, one may still obtain an analytical expression for the distribution, which is then given by a linear combination of other χ^2 expressions. The specific linear combination for a given problem is derived using the geometrical arguments developed previously [33,34]. A more severe problem than the possible vicinity to boundaries is the fact that the number of data points often is limited. This means that practical identifiability becomes a real problem [i.e. that d_i typically is lower than $\dim(p_i){-}t_{\mathcal{M}_i}$], and that the parameter distributions no longer are Gaussian. Furthermore, it is not uncommon that the tested model structures are non-nested. This problem was first considered by Cox [35,36] who obtained some asymptotic results, which have been developed further [31]. For the general situation of limited data, the likelihood ratio test function, $\mathcal{T}_{lr}$, may still be used, but the distribution to which it should be compared is no longer possible to obtain analytically. It may, however, be obtained using simulation based approaches such as bootstrapping, which we describe below.

Another important test that should be mentioned is the *F*-test. It also provides a significance to its comparison, and the test and the corresponding distribution are given by [13,22]:

$$\mathcal{T}_{\mathrm{F}} = \frac{V(\widehat{p}^1) - V(\widehat{p}^2)}{V(\widehat{p}^2)} \frac{N - d_2}{d_2 - d_1} \in \mathcal{F}_{N-d_1,d_2-d_1} \tag{19}$$

where F is the F-distribution, and the indices specify the degrees of freedom. The test is asymptotically equal to the likelihood ratio test, but has been shown to have less power for fewer data points [37].

Bootstrap solutions

Bootstrapping is a general method to estimate the distribution of almost any property that has been estimated from experimental data. Historically, simulation-based precursors to bootstrapping have had the reputation of being empirical, and nonstringent, compared to for example the exact analytical solutions described above. However, subsequent to some groundbreaking studies [38–40] clarifying the theoretical motivations for bootstrapping, bootstrapping has been considered as another mathematically valid approach to statistical problems. Actually, as is clear from the comments in the previous sections, the commonly used analytical solutions are also burdened with severe problems of validity, due to underlying assumptions that typically are not fulfilled. Bootstrapping approaches may often be based on fewer such assumptions, with the compensation of a higher computational cost for calculating the sought distributions [41].

The basic idea is to estimate the distribution of a property θ by generating new data sets b^i from the given data set Z^N (Fig. 9). The most straightforward approach to bootstrapping is probably the nonparametric bootstrap, which is as resampling with replacement [41]. Here, each bootstrap is solely based on picking samples from the given data series, Z^N, where each data point is returned to the pool of data before each new point is picked. With this procedure, and $N = 5$, three bootstraps could be given by:

$$b^1 = \{y(t_2), y(t_3), y(t_3), y(t_4), y(t_5)\}$$
$$b^2 = \{y(t_1), y(t_2), y(t_2), y(t_5), y(t_5)\}$$
$$b^3 = \{y(t_1), y(t_2), y(t_3), y(t_4), y(t_5)\}$$

Note that data points might appear in more than one place in a single bootstrap; in fact, this is what allows the bootstraps to vary.

Common in all bootstrap approaches is that each bootstrap corresponds to a 'new version' of the original time-series Z^N (Fig. 9). These new versions should share some critical properties with the original time-series, but the bootstraps taken together should also give a representation of variations that might occur (e.g. if the experiment was conducted again). In the nonparametric approach mentioned above, the shared features are the total number and the values of the data points themselves, and the variation is given by the number of times the data points appear.

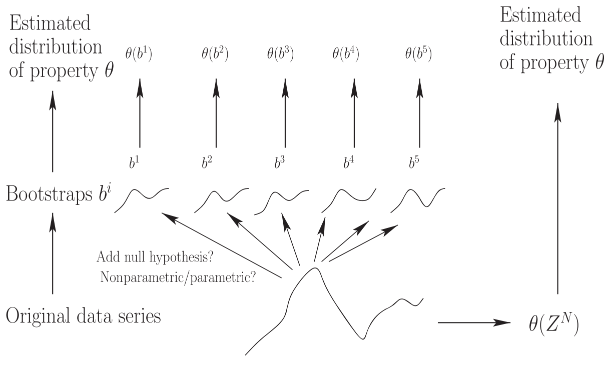

Fig. 9. Graphical depiction of the idea behind bootstrapping. First bootstraps are generated that are similar, but not identical, to the original data set, Z^N. Then the property of interest deduced from the data set, which we denote $\theta(Z^N)$, is calculated for all the bootstraps, and the resulting set of values serves as an empirical distribution with which $\theta(Z^N)$ can be compared.

Another type of bootstrap is based on a model $\mathcal{M}_1$, and on analysis of the corresponding residuals. In such residual-based bootstrapping, each new bootstrap is generated by the simulated curve, which is the best fit of $\mathcal{M}_1$ and Z^N, to which a new realization of the estimated of the estimated noise distribution (or a resampling of the residuals) is added. The noise distribution is not necessarily estimated from the residuals $e^{\mathcal{M}_1,}$ but may be estimated from the residuals of another low-bias model, or from a part of the time-series where the noise is believed to be the only reason for the fluctuations [22]. Model-based bootstrap generation is typically referred to as a parametric bootstrap, even if there is a gray-zone between nonparametric and parametric bootstraps. A general basic introduction to bootstrap approaches is provided elsewhere [39,41,42] and a more theoretically advanced alternative is also available [43].

A simulation-based approach to likelihood ratio distribution estimation was first proposed by Wlliams [38]. The proposed method for evaluating the differences between two non-nested nonlinear model structures $\mathcal{M}_f$ and $\mathcal{M}_g$ with respect to a limited data set can essentially be summarized as [38,44]:

(a) Fit models $\mathcal{M}_f$ and $\mathcal{M}_g$ to obtain parameters $\widehat{p}^f$ and $\widehat{p}^g$, and calculate the observed likelihood ratio, $\mathcal{T}_{lr}$, according to (18).

(b) Simulate B bootstraps based on the fitted outputs $\widehat{y}^f(t, \widehat{p}^f)$ corresponding to fitted model $\mathcal{M}_f(\widehat{p}^f)$. Fit both models to each bootstrap to obtain $\widehat{p}^{f,fr}$, $\widehat{p}^{g,fr}$, and calculate $\mathcal{T}_{lr}^{*,fr} = 2(l_f(\widehat{p}^{f,fr}) - l_g(\widehat{p}^{g,fr})), r = 1, \ldots, B$.

(c) Simulate B bootstraps based on the fitted outputs $\widehat{y}^g(t, \widehat{p}^g)$ corresponding to fitted model $\mathcal{M}_g(\widehat{p}^g)$. Fit both models to each bootstrap to obtain $\widehat{p}^{f,gr}$, $\widehat{p}^{g,gr}$, and calculate $\mathcal{T}_{lr}^{*,gr} = 2(l_f(\widehat{p}^{f,gr}) - l_g(\widehat{p}^{g,gr})), r = 1, \ldots, B$.

The value $\mathcal{T}_{lr}$ is then compared with the simulated sets of values $\mathcal{T}_{lr}^{*,fr}$ and $\mathcal{T}_{lr}^{*,gr}$ to indicate support for one or the other of the models, inability to choose between them, or possible evidence against both models. In practice, it is often convenient to replace the log-likelihood function by the sum of residuals, typically normalized with the variance of the noise, and to drop the factor 2 in all places. Finally, significance levels can be obtained by formulae such as:

$$\widehat{\alpha} = \frac{\#(\mathcal{T}_{lr}^{*,gr} < \mathcal{T}_{lr})}{B} \tag{20}$$

where the # symbol indicates the number of $\mathcal{T}_{lr}^{*,gr}$s that fulfils the criterion $\mathcal{T}_{lr}^{*,gr} < \mathcal{T}_{lr}$. This bootstrap approach has been described and used to some extent in econometrics studies [45–47] and, in some modified forms, also in bioinformatics [48,49] and a few systems biology studies [10,11]. It should, however, be noted that there is currently no consensus about exactly what to use as a test function, or how to calculate the distribution [10,37,50]; furthermore, the asymptotic validity is, at least in some situations, still disputed [51].

A general scheme for comparison between two models

Measuring the difference between two models

It might often happen that several explanations pass all the quality tests described in the section 'Rejections based on a residual analysis', and that none of these explanations provide a significantly better predictor than another according to the tests described in the section 'The F and the likelihood ratio test'. Then these explanations can be analysed further, because other properties of the models might lead to rejection of some of the explanations anyway. Similarly, it is also interesting to examine the models' characteristic similarities and differences because this will also provide crucial information on how to relate to the remaining explanations.

The first and most straightforward option is to visually inspect the two models (e.g. by comparing the biochemical interpretations of their interaction graphs, or by comparing their behaviors in specific simulations). Note that the studied behaviors now also can include the response of the models to new inputs or operating conditions, or the behavior of other states, compared to those examined in the earlier tests.

Typically, some states or properties that have not been measured in the given data set, Z^N, are of especially large interest. Denote these output variables y_o and assume that they are given by some function h as:

$$y_o = h(x, p) \tag{21}$$

where x and p are the states and parameters specified in Eqn (3). An obvious entity to consider is the difference between these outputs for different model structures, $y_o^1 - y_o^2$, where the superscript i as usual denotes that the model prediction corresponds to $\mathcal{M}_i$. These differences may also be mapped to a more formal distance measure, $\mathcal{D}$, between the two models; for example, by integrating over time:

$$\mathcal{D}_{ij} = \int_t \|y_o^i - y_o^j\| \tag{22}$$

where $\|$ denotes some suitable norm.

Core predictions

When identifying the interesting model outputs, y_o, one should not only consider whether a particular output is biologically interesting, but also the quality of that part of the model. Ideally, the identification step [17] should not only produce an identified parameter set $\widehat{p}$, but also an uncertainty in the model predictions. Because over-parametrization and unidentifiability is common and usually quite substantial in systems biology models, many predictions made by a systems biology model will be highly uncertain. For many predictions, the uncertainty can be so large that almost any value could be produced, while still allowing for a good agreement with Z^N [25]. On the other hand, there are also model predictions that must be fulfilled if that particular model structure is to describe the given data set. Such uniquely identified predictions with a high quality tag (low uncertainty) were given the name core predictions [25] (G. Cedersund, J. Roll, T. Pettersson, H. Tidefelt & P. Strålfors, unpublished data), and they are obviously interesting candidates to qualify as interesting model outputs y_o.

Core predictions may be identified in various ways. One way is to first determine the uncertainty of the estimated parameters, $\Delta\widehat{p}$; for example, by using the Hessian of the cost function [22,25], or by using modifications of global searches for the optimization step [53] (G. Cedersund, J. Roll, T. Pettersson, H. Tidefelt & P. Strålfors, unpublished data). These uncertainty regions in the parameter space can then be sampled, and subsequently simulations can be used to translate the parameter uncertainty to a corresponding uncertainty in specific model predictions [25] (G. Cedersund, J. Roll, T. Pettersson, H. Tidefelt & P. Strålfors, unpublished data). The model predictions that are highly similar for all sampled parameter values could be taken as core predictions. This is a good way of identifying potential core predictions but, ideally, they should then also be specifically tested. This can be carried out as follows. Assume that a candidate for core prediction is denoted $y_c(t, \widehat{p})$ and that its values are given by $c_y(t)$. Then one can form the following constrained optimization problem:

$$\max_p \int_t \|y_c(t,p) - c_y(t)\| \quad \text{subject to} \quad V(p) < \delta \tag{23}$$

where V is the cost function describing the quality of the model, and where δ can be chosen according to a

5% significance threshold from some of the tests described in the Section 'Rejections based on a residual analysis'. Note that even though solving Eqn (23) is difficult, it follows the standard formulation of a constrained optimization problem, and there are advanced optimization algorithms that can tackle such problems, both locally [54] and globally [55].

Finally, for many systems biology models, the dimension of the parameter space is so large that searching it becomes a serious problem, especially for the more advanced optimization algorithms. This is one of the main reasons why smaller models may be useful, especially for identification of core predictions. For nested model structures, this is possible because a smaller model structure is equal to the larger model, with some of its parameter values set to constant values (typically to zero). Also in the non-nested case, a smaller model may give information about a larger model if the models are related to each other in some other comprehensible way. For example, a state (or reaction) in the smaller model could correspond to a lump of many states (or reactions) in the larger model. In all such cases, an analysis of the smaller model will give information about the larger model. Note that this is information that, in principle, is possible to extract from an analysis of the larger model directly, but that, in practice, is impossible to extract due to the high dimensional of the larger model's parameter space. In that case, testing and comparing of different submodels may be a feasible alternative for drawing conclusions; for example concerning which parts of a larger model that may, may not, and must be active, if the larger model should explain the data. (see also the discussion on the choice of model size in the Discussion).

Summary of the central questions and steps to be taken

We have now introduced the most important methods and tests in this minireview; let us finally see how these relate to each other, and suggest how they can be combined to achieve a complete analysis of a given set of data, prior knowledge and proposed explanations.

First, however, it should be stressed that the construction of a formal division of the analysis process in specific substeps is virtually impossible. First of all, analysis is, just like modeling in general, an iterative process, which requires human reasoning that cannot be fully automated. For example, earlier steps and analyses may have to be revisited due to new insights and suggestions. Furthermore, each problem is unique, and requires its specific approach and combination of methods, possibly also including methods that have not been proposed in this minireview. It is also important to have a clear understanding of what the purpose of the analysis is. As we have stressed repeatedly, the purpose of a systems biology problem is generally different from that of a classical engineering and statistical problem, but this might not always be the case, and there are certainly large variations between different systems biology problem settings. Nevertheless, with all these comments made, we would like to discuss the structure of the overall problem by making a subdivision of the data analysis process into three major steps (Fig. 10).

The first step is the reformulation and formalization of the available data, prior knowledge, and suggested explanations into formal data sets and model structures. We have not dealt with this step extensively in this minireview because it is dealt with in many text books and exemplified in many modeling articles [4,9,24,25]. However, it should be stressed that the choices regarding which model structures to consider as different cases of a super model structure (containing all of them as special cases), and which model structures to consider separately, is not always treated in such texts, because model comparison is not an explicit part of all modeling works. Furthermore, this division problem is a highly nontrivial issue, and much of the following analysis will provide further insights into whether there are other, better, subdivisions.

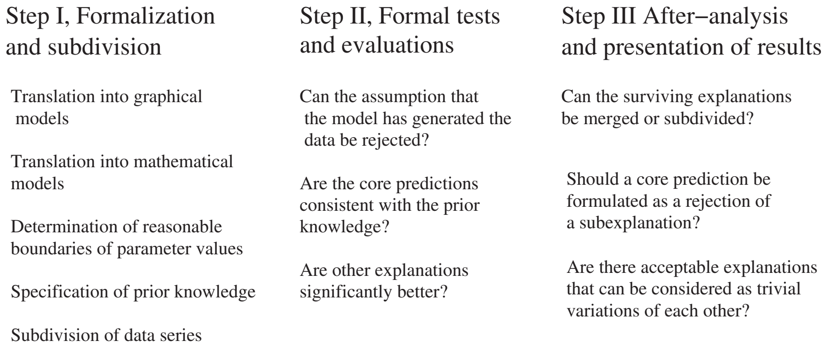
Step I, Formalization and subdivision
Translation into graphical models
Translation into mathematical models
Determination of reasonable boundaries of parameter values
Specification of prior knowledge
Subdivision of data series
Step II, Formal tests and evaluations
Can the assumption that the model has generated the data be rejected?
Are the core predictions consistent with the prior knowledge?
Are other explanations significantly better?
Step III After–analysis and presentation of results
Can the surviving explanations be merged or subdivided?
Should a core prediction be formulated as a rejection of a subexplanation?
Are there acceptable explanations that can be considered as trivial variations of each other?

Fig. 10. The three main steps in a model-based data analysis and explanation evaluation process. Also, common substeps and questions are suggested.

The second step is the most central step in this review because it contains the actual tests and quality evaluations. The overall question is whether an explanation is acceptable. The first type of statistical tests that we have considered test whether the null hypothesis that the model has generated the data Z^N can be rejected. Such methods were reviewed in the section 'Rejections based on a residual analysis'. It should here be added that one also should test the quality of the explanation from a biological point of view, and in the light of the quality tags. It might, for example, be the case that a core prediction (i.e. that is, a property that must be fulfilled for a particular explanation to be able to mimic the data) is biologically unrealistic. This will not be seen by the tests described in the section 'Rejections based on a residual analysis', but is still a related question because it is the result of an analysis of the quality of an individual explanation. The second type of rejection concerns comparisons between explanations. An explanation that passes the first type of tests may still be significantly worse than another explanation. Two important such methods were reviewed in the section 'The F and the likelihood ratio test', but it should again be stressed that this analysis should be complemented by the results from the quality tag analysis, combined with the prior biological knowledge.

The third and final step is concerned with the surviving explanations (i.e. those that have not yet been rejected). Basically, this step deals with the presentation of the results. This, however, also involves a revisiting of the subdivision decisions taken in the first step. This revisiting is a good idea, for example, when the core predictions have shed some new light on the issue of subdivisions. Consider for example, that a core prediction shows that a particular reaction rate in a model structure must have a high value (i.e. that small values are excluded). That is the same as rejecting the submodel to the original model that lacks this particular reaction rate (Fig. 3). The final result could therefore be presented as a rejection result, but with a different subdivision in competing explanations. Conversely, some surviving explanations might also benefit from being presented as a merged super-model containing the individual models as special cases. This could be the case if none of its tested submodels may be rejected, and when this is not judged to be an interesting insight in itself. Note, however, that the submodels might give different core predictions, which are experimentally testable. In such a case, the submodels could be presented differently, and the result could serve as a guide for future experimental design. However, experimental design and the iteration of the above mentioned analyses with the experimental data gathering phase is the topic of an accompanying minireview [56], and is outside the scope of the present one.

Software

An efficient and user-friendly software option for statistical testing is MATLAB, for which there are at least two toolboxes (http://www.potterswheel.de and http://www.sbtoolbox.org) [57] targeting the systems biology community, which both have some basic statistical testing functionalities. However, neither of them have implemented all the methods reviewed here, although this might be improved within the near future. There are also rather well-developed statistical environments with several ready-to-use tests in both MATHEMATICA and MAPLE. Some more statistically oriented softwares are given by R (http://www.r-project.org) and S-PLUS (http://www.insightful.com). However, none of these other generic software environments provide toolboxes for the systems biology community.

Discussion

The focus of this minireview is the problem of evaluating and comparing two or several explanations for a given set of data and prior knowledge, so as to identify the best available explanations. We have reviewed methods that evaluate a single explanation with respect to the data directly, methods for evaluating whether one explanation is significantly better than another, and put them together into a general framework for comparison and evaluation of suggested explanations.

Most of the presented methods are based on statistical and engineering methods, which have a slightly different epistemological setting than that of systems biology (i.e. the type of knowledge that is sought is different). These differences are important, and also are important to clarify and agree upon if systems biology is to mature as a research field. To contribute to this process, we will now seek to clarify some of these epistemological differences.

In a systems biology setting, the focus is on the understanding of the underlying biological mechanisms, and not just on achieving an optimal predictor of a given system output. We have highlighted this difference through the usage of the term explanation, rather than the term hypothesis, which is the typical choice in a statistical hypothesis testing setting.

An important concept in relation to this is instrumentalism, as are its various opposing concepts. Instrumentalism is the view that a model is only used as an instrument (as a means) to obtain a certain prediction [8,58]; hence, view is often the case in an engineering setting. One opposite of instrumentalism is direct realism [58]. According to this view there is a one-to-one correlation between a 'perfect' model and the real system. This means that the 'perfect' model will not only be able to give accurate predictions of the measurable system output y, but also will provide an accurate description of all the components and processes involved in the generation of this output. This view could certainly be ascribed to many theoretical physicists, which aspire to find a final theory describing reality as it really is [8,59]. A more moderate view is referred to as critical realism [58]. According to this view, it is acknowledged that a model yielding good predictions on a wide variety of data could be expected to contain some degree of correlation between its components and mechanisms and the components and mechanisms of the real system. However, a model is still viewed as a simplification of the true system, which only captures some of its aspects, and one therefore has to be careful when drawing conclusions about which these aspects might be. Of these three options (i.e. instrumentalism, direct realism and critical realism), we argue that it is the last option that describes the best view for systems biology.

One final issue regarding the differences between the underlying modeling philosophies concerns the ideal size of a model. A classical engineering principle for choosing the size of a model is known as Occam's razor. According to this principle, one should not add any unnecessary details to the model (i.e. one should choose the smallest model that does the job). Another reason for not choosing overly complex problems is that the variance term Err_{var} in Eqn (14) increases with complexity (i.e. the over-fitting problem). However, in a systems biology setting, the situation is different. First, the purpose with the model is to provide an explanation. That means that 'doing the job' could mean including all the known details of the system, apart from being able to produce good predictions $\hat{y}$. Furthermore, biological model structures are typically of a limited flexible (i.e. they will not be able to describe the data better than a certain agreement, even if more mechanistic details within the given explanation are added). To stress this difference between the size of the model and its flexibility, one sometimes uses measures of model complexity other than the number of parameters or states. One such measure is the effective number of parameters, $A_{\mathcal{M}}$, in Eqn (11). That $A_{\mathcal{M}}$ and $\dim(p){-}t_{\mathcal{M}}$ typically are widely different in a systems biology model means that unidentifiability typically is a severe problem; however, it also means that the variance increase (the over-fitting problem) typically is a less pronounced problem. Therefore, a systems biology model can often benefit from adding more mechanistic details, and thus providing a 'better' explanation, without suffering from the problem of over-fitting or variance increase. Note that this could still be considered as being consistent with the principle of Occam's razor if the additional mechanistic details are considered as a part of the data that should be explained by the model (Fig. 11). Finally, these insights do not mean that systems biology models always should be large. By contrast, as described in the section 'A general scheme for comparison between two models', finding small models that can and cannot explain the data is a highly useful way of identifying core predictions (i.e. of learning crucial information about the available explanations).

It is important when using statistical tests to evaluate a potential explanation to achieve a sound

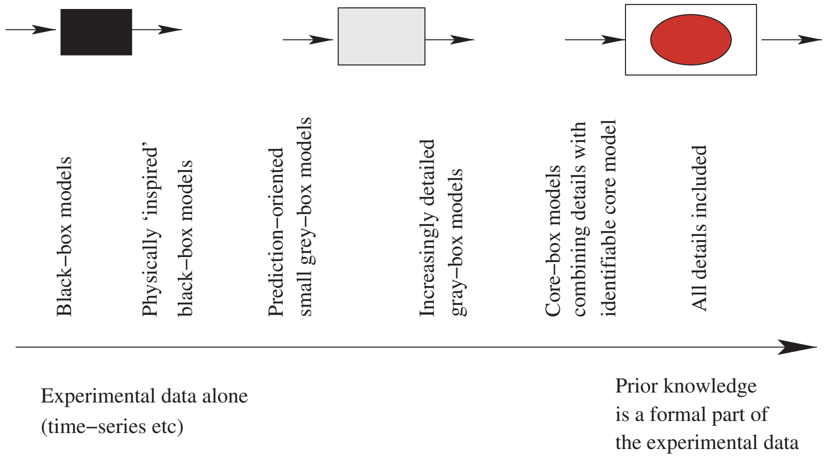
Black–box models
Physically 'inspired' black–box models
Prediction–oriented small grey–box models
Increasingly detailed gray–box models
Core–box models combining details with identifiable core model
All details included
Experimental data alone (time–series etc)
Prior knowledge is a formal part of the experimental data

Fig. 11. Symbolic scheme illustrating the differences in ideal model size between different traditions as a difference in whether the prior knowledge is a part of the data set that should be predicted/explained or not. The more that emphasis is laid on the prior knowledge, the more that mechanistic details may have to be included, and the larger the models become.

criticism of the statistical results. Statistical results should always be used as support for a decision, and not as the final decision itself. This is especially the case in a systems biology setting because there is so much prior knowledge that the explanation should be evaluated with respect to. It is also important to be critical of the actual test (i.e. 'to test the test'). This is especially the case if a nonstandard version of a test has been constructed especially for a particular problem. Testing the test can be carried out by constructing relevant test problems where the answer is known, and where the behavior of the test entity can be evaluated. Another important aspect of testing the test is to examine the underlying assumptions. This could concern the assumptions regarding the noise in the system. Perhaps it is possible to examine further what the true noise distribution is, and then to modify the test accordingly. This was conducted, for example, by Kreutz *et al*. [60], where it was shown that western blot noise typically is multiplicatively normal, and that it is possible to make it additively normal (i.e. the standard assumption) by a simple modification of the image analysis. However, when it is impossible to modify the test procedure so that the theory can be fulfilled, it might still be possible to make an estimate of how much the assumptions are violated, and to estimate the qualitative and quantitative implications of this violation [10].

Finally, let us add some short comments regarding important future work within this field. The classical systems biology situation of nonlinear, dynamical, non-nested models has received little attention in the statistical literature, compared to many other modeling situations, and few results and methods are actually valid for this situation. It will be important to develop methods for this specific situation, and to further evaluate the implications of violating the assumptions of the currently available methods in these specific ways. Currently, the most general way of comparing whether one model is significantly better than another is probably the likelihood ratio test, where the corresponding distribution is generated using some kind of parametric bootstrap [10]. However, this approach has been used relatively little in the systems biology discipline, and it must generally be examined further, both theoretically and in practical situations [51]. The approach is also somewhat computationally expensive. Therefore, a feasible but highly useful future goal could be to implement a more mature version of that approach in a public software platform associated with a powerful computer cluster, where one can submit systems biology models for testing of significant differences.

Acknowledgements

GC was supported by the BioSim Network of Excellence, and wishes to thank Jens Timmer (University of Freiburg, Germany) for helpful discussions and references.

References

1 http://www.ietdl.org/IET-SYB; http://www.nature.com/msb; http://compbiol.plosjournals.org/

2 http://cordis.europa.eu/fp7/cooperation/health_en.html

3 Kitano H (2002) Computational systems biology. *Nature* **420**, 206–210.

4 Klipp E, Herwig R, Kowald A, Wierling C & Lehrach H (2005) *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, Weinheim.

5 Di Ventura B, Lemerie C, Michalodimitrakis K & Serrano S (2006) From in vivo to in silico biology and back. *Nat Reviews* **443**, 527–533.

6 Popper K. (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, UK.

7 Kuhn TS (1962) *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

8 Deutsch D (1998) *The Fabric of Reality: The Science of Parallel Universes and Its Implications*. Penguin, London (chapter 9).

9 Swameye I, Müller TG, Timmer J, Sandra O & Klingmüller U (2003) Identification of nucleucytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *Proc Natl Acad Sci USA* **100**, 1028–1033.

10 Müller TG, Faller D, Timmer J, Swameye I, Sandra O, Klingmüller U (2004) Tests for cycling in a signalling pathway. *J Royal Stat Soc C* **53**, 557–568.

11 Timmer J, Müller TG, Swameye I, Sandra O and Klingmüller U (2004) Modeling the nonlinear dynamics of cellular signal transduction. *Int J Bif Chaos* **14**, 2069–2079.

12 Cedersund G, Roll J, Ulfhielm E, Tidefelt H, Danielsson A & Strå lfors P (2008) Model based hypothesis testing of key mechanisms in initial phase of insulin signaling. *PLoS Comp Biol* **4**, e1000096.

13 Ljung L (1999) *System Identification Theory for the User, 2nd edn*. Prentice-Hall inc., Upper Saddle River, NJ.

14 Segel IH (1975) *Enzyme Kinetics*. John Wiley & Sons, New York, NY.

15 Hastie T, Tibshirani R & Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, Berlin.

16 Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J Royal Statist Soc B* **58**, 267–288.

17 Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA & Blom JG (2009) Systems biology: parameter estimation for biochemical models. *FEBS J* **276**, 886–902.

18 Gut A (1991) *An Intermediate Course in Probability*. Springer-Verlag, New York, NY.

19 Sheskin DJ (1997) *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC Press, New York, NY.

20 Kanji GK (1994) *100 Statistical Tests*. SAGE Publications Ltd, Chennai.

21 Sedoglavic A (2002) A probabilistic algorithm to test local algebraic observability in polynomial time. *J Symb Comput* **33**, 735–755.

22 Dochain D & Vanrolleghem PA (2001) *Dynamic Modeling and Estimation in Wastewater Treatment Processes*. IWA Publishing, London, UK.

23 Teusink B, Passarge J, Reijenga CA, Esgalhado E, der Weijden CC, Schepper M, Walsch MC, Bakker B, van Dam K, Westerhof H *et al.* (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem* **267**, 5313–5329.

24 Hynne F, Danø S & Sørensen PG (2001) Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Bioph Chem* **94**, 121–163.

25 Cedersund G (2006) Core-box modelling – theoretical contributions and appliciations to glucose homeostasis related systems. Ph.D. dissertation, Dept. Elect. Eng., Chalmers, Gothenburg, Sweden.

26 Anguelova M, Cedersund G, Johansson M, Franzen C-J & Wennberg B (2007) Conservation laws and unidentifiability of rate expressions in biochemical models. *IET Syst Biol* **1**, 230–237.

27 Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* **AC-19**, 716–723.

28 Akaike H (1981) Modern development of statistical methods. In *Trends and Progress in System Identification* (Eykhoff P, ed). Pergamon Press, Elmsford, NY.

29 Chernoff H (1954) On the distribution of the likelihood ratio. *Ann Math Stat* **25**, 573–578.

30 Chant D (1974) On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika* **61**, 291–298.

31 Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

32 Miller JJ (1977) Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann Stat* **5**, 746–762.

33 Shapiro A (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72**, 133–144.

34 Self S & Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Soc* **82**, 605–610.

35 Cox DR (1961) Tests of separate families of hypotheses. *Proc Fourth Berkeley Symp on Mathem Stat Prob* **1**, 105–123.

36 Cox DR (1962) Further results on tests of separate families of hypotheses. *J Roy Stat Soc B* **24**, 406–424.

37 Müller TG (2002) Modelling complex systems with differential equations. Ph.D. disseration, Freiburg University, Germany.

38 Williams DA (1970) Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell culturues. *Biometrics* **28**, 23–32.

39 Efron B (1979) Bootstrap methods: another look at the Jackknife. *Ann Stat* **7**, 1–26.

40 Efron B (1987) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

41 Efron B & Tibshirani RJ (1993) *An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC Press, New York, NY.

42 Davison AC & Hinkley DV (1997) *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

43 Hall P (1992) *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, Berlin and Heidelberg GmbH & Co KG.

44 Hinde J (1992) Choosing between nonnested models: a simulation approach. In *Advances in GLIM and Statistical Modelling. Proceedings of the Glim92 Conference* (Fahrmeir L *et al.*, eds). Springer-Verlag, Munich.

45 Kim S, Shephard N, Chib S (1998) Stochastic volatility: likelihood ratio inference and comparison with ARCH models. *Rev Econ Studies* **65**, 361–393.

46 Pesaran MH, Weeks M (2001) Non-nested testing: an overview. In: *Companion to Theoretical econometrics* (Baltagi, B, ed), pp. 279–309. Basil Blackwell, Oxford.

47 Winkelmann R (2003) *Econometric Analysis of Count Data*. Springer-Verlag, New York, NY.

48 Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**, 652–670.

49 Schork N (1992) Bootstrapping likelihood ratios in quantitative genetics. In *Exploring the limits of the bootstrap* (LePage R, Bilard L, eds). Wiley, New York, NY.

50 Hall P & Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762.

51 Godfrey LG (2007) On the asymptotic validity of a bootstrap method for testing nonnested hypotheses. *Econ Lett* **94**, 408–413.

52 Reference withdrawn.

53 Pettersson T (2008) Modified global searches for identification of core predictions. M.Sc. Thesis, Linköping University, Sweden.

54 Nocedal J & Wright SJ (1999) *Numerical Optimization*. Springer-Verlag, New York, NY.

55 Wah BW & Wang T (2000) Tuning strategies in constrained simulated annealing for nonlinear global optimization. *Int J AI Tools* **9**, 3–25.

56 Kreutz C & Timmer J (2009) Systems biology: experimental design. *FEBS J* **276**, 923–942.

57 Schmidt H, Jirstrand M (2006) Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics* **22**, 514–515

58 Barbour IG (2002) *Religion and Science: Historical and Contemporary Issues*. HarperCollins Publishers, New York, NY.

59 Hawking S (1988) *A Brief History of Time*. Bantam Books, New York, NY.

60 Kreutz C, Bartolome Rodriguez MM, Maiwald T, Seidl M, Blum HE, Mohr L & Timmer J (2007) An error model for protein quantification. *Bioinformatics* **23**, 2747–2753.

Supporting information

The following supplementary material is available:
Doc. S1. Simulation files used for the calculations in the examples.
Doc. S2. Summary of standard formulae for model comparison, such as AIC, BIC, including a more explicit statement of the underlying assumptions and their correspondence to the different versions of these formulae appearing in the three minireviews in this series.

This supplementary material can be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.