Part Three
Analysis of Gene Networks
9

What if the Fit is Unfit? Criteria for Biological Systems

Estimation Beyond Residual Errors

Eberhard O. Voit

9.1

Introduction

The analysis of biological networks has made enormous strides in recent years. In the context of static networks, which do not change over short periods of time, new biological techniques have begun to permit the characterization of very large interaction maps (e.g., among proteins), and computational graph theory has been the tool of choice for analyzing and interpreting what these maps entail. Complementing these activities has been the exploration of dynamically changing, regulated biological systems. On the experimental side, these efforts have enormously benefitted from the astounding advances in high-throughput biology at the genomic, proteomic, metabolic, and physiological levels. On the analytical side, the procedures and results of this new field of experimental systems biology have first been supported by a rapidly expanding repertoire of bioinformatics tools that permit the storage, retrieval, and analysis of very large datasets. More recently, the bioinformatics tools have become tightly interwoven with analytical and simulation techniques that are at the heart of the emerging field of computational systems biology.

At this point in time, it is no longer a real challenge to simulate large linear or even nonlinear systems in the form of algebraic or differential equations. It has also become feasible to simulate hybrid systems that contain continuous and discrete events, stochastic effects, and delays (e.g., [1, 2]). The simplicity with which we can perform large-scale simulations is in stark contrast to the overwhelming challenges we face much earlier in any biological systems analysis, namely when the biological phenomenon of interest is to be translated into a mathematical or computational model. This translation task may be subdivided into three aspects. The first concerns the acquisition of data. While biology is producing high-quality data in large quantities, these data are not always of the type and completeness that elucidate all aspects of the biological phenomenon from sufficiently many angles to construct a mathematical model. For instance, models of dynamic processes in the brain are hampered by the extremely difficult access to specific, restricted neuronal areas in living organisms [3]. This aspect of data availability will without doubt continue to
improve throughout the foreseeable future and we will not discuss issues of data
generation in this chapter. The other two components of the translation from biology
into a computational construct are distinct, but closely related to each other. The first
is the determination of suitable mathematical descriptions for all relevant details of
the biological phenomenon, while the second is the identification of numerical values
for the parameters in these descriptions [4]. These two fundamental tasks of
computational systems biology are the focus of this chapter.

Before we discuss details, challenges, methods, and pitfalls associated with the
construction of biological systems models, we should ask why such an effort appears
to be worth our while. One might begin by pre-empting a widespread critique of
modeling, namely that models merely recreate, often in a much abstracted and
simplified fashion, what “real” biologists had known all along and in greater detail. So
what, if a model produces results similar to those observed? Modelers are sometimes
stunned by this critique, because it is certainly not a trivial matter to write computer
code that fits a large collection of biological data well. The truth behind this (mis-)
conception is that a well-fitting model is a necessary but not sufficient condition for
greater things to come. Indeed, without further analysis, exploration, explanation, or
prediction, an accurate fit by itself does not earn the modeler much more than
bragging rights.

In generic terms, the construction of a model is worthwhile if the model is able to
answer specific questions or helps decide between acceptance and rejection of a
hypothesis. Such a hypothesis may take many different forms. It may be qualitative in
a sense that one is primarily interested in whether some key variable in the system
responds to a specific input with an increase or a decrease. It may be semi-quantitative
if one is interested in the rough extent of the response and it is quantitative or
numerical if the model is supposed to show the correct value of the affected key
variable. What level of accuracy is needed in a model result depends on the questions
asked, on the effort one is willing to invest in the modeling effort, and on many issues
associated with the biological phenomenon itself and with the model.

9.2
Model Design

The challenge of converting a biological system into a mathematical structure
requires the specification of functions that describe all pertinent processes, as well
as the identification of suitable parameter values. The selection of process descrip-
tions is by no means trivial. Granted, there are situations where a function can be
inferred from the type and mechanism of the process. For example, there is good
reason to choose an exponential function for the description of the growth of a small
bacterial population, because the process is biologically driven by repeated cell
doubling. However, such cases of mechanism-based model selection are actually
rare and even traditional choices like a Michaelis–Menten rate function for an
enzyme-catalyzed reactions are not without troubling questions, because deep
underlying assumptions like homogeneity of the medium and free movement of
enzymes and substrates are usually not satisfied \textit{in vivo} [5]. Physics affords us with a rich repertoire of proven formulations for fundamental characteristics like forces and energy, but in biology these fundamental aspects are convoluted and often mixed together in a complicated manner. As an example, consider the process of gene expression, which involves the opening of the DNA strands, the right spatial and temporal availability and action of transcription factors, and the complex process of transcription into RNA. It is simply impossible to reduce this collective event into mechanistic pieces that permit elementary, physics-based representations.

Faced with similar challenges, engineers typically resort to linear approximations. These are very convenient, because there are stringent rules for their design, as well as for their analysis. Indeed, the repertoire of analytical and computational methods for linear systems is huge. The problem with linear approaches in biology is that most phenomena are genuinely nonlinear. They saturate or oscillate in a stable fashion, show switches, and sometimes appear to be chaotic. Reducing their dynamics to linear functions would not permit a proper analysis of these features. At the same time, the number of nonlinear functions is infinite and there are no guidelines as to which of these might be optimal or even appropriate descriptions of biological processes. A useful alternative is a nonlinear approximation. The first idea presumably coming to mind might be a second-order (quadratic) approximation, but this choice actually turns out to be rather inconvenient for later analyses [6]. Instead, it has proven beneficial to approximate biological processes with linear functions in logarithmic coordinates. This procedure is mathematically sound, as it directly adheres to the tenets of Taylor’s theory and leads to nonlinear descriptions that can capture all types of responses, including different types of oscillations and chaos [7, 8]. Besides, these representations have desirable properties for mathematical and computational analysis. The concept of linearization in logarithmic coordinates is the core of biochemical systems theory (BST) [9, 10], which has been documented in several hundred articles and book chapters; book-length descriptions include [11–14].

BST comes in two main variants. In the generalized mass action (GMA) formulation, every process is represented with one product of power-law functions. For instance, in the simple branched pathway with two feedback signals that is shown in Figure 9.1, the equation for $X_2$ can be formulated directly as:

$$\dot{X}_2 = \gamma_{21} X_1^{f_{21}} - \gamma_{22} X_2^{f_{22}} - \gamma_{23} X_2^{f_{23}} X_4^{f_{24}}$$ (9.1)

where the $\gamma$ parameters denote rate constants, which can take any non-negative values, while $f_{21}, f_{22}, f_{23}$, and $f_{24}$ are kinetic orders that may take any real values. As $X_1$ is the substrate of the production reaction, $f_{21}$ is positive. By contrast, $f_{24}$ is negative, because it represents the inhibitory signal exerted by $X_4$. In general, a GMA system always has the format:

$$\dot{X}_i = \gamma_i \prod_{j=1}^{n+m} X_j^{f_{ij}} \pm \gamma_{12} \prod_{j=1}^{n+m} X_j^{f_{1j}} \pm \cdots \pm \gamma_n \prod_{j=1}^{n+m} X_j^{f_{nj}} \pm \cdots \quad i = 1, \ldots, n$$ (9.2)

In addition to the dynamically changing variables, the system may also contain independent variables, $X_{n+1}, \ldots, X_{n+\nu}$, which affect the system, but are not
affected by the system. In many cases, these variables are constant during a given mathematical experiment.

In the alternative S-system formulation, all processes entering a variable or pool are collectively represented with a single product of power-law functions that contains all variables affecting the collection of fluxes. Similarly, all processes leaving a variable or pool are collectively represented with a single product of power-law functions that contains all variables affecting the collection of fluxes. Revisiting the pathway in Figure 9.1, the only difference to the GMA formulation occurs for the degradation of $X_2$, which is now represented with only one term that contains both $X_2$ and $X_4$. Using the conventional parameter names for S-systems, the equation for $X_2$ is therefore:

$$X_2 = \alpha_2 X_1^{b_2} - \beta_2 X_2^{h_2} X_4^{h_4}$$  \hspace{1cm} (9.3)

All other equations are the same as before, with the minor deviation of traditionally different names for the parameters. Accounting again for independent variables, the generic S-system format is:

$$X_i = \alpha_i \prod_{j=1}^{n+m} X_j^{b_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \quad i = 1, \ldots, n$$  \hspace{1cm} (9.4)

BST models have a number of advantages over ad hoc formulations. First, in order to formulate the system equations one does not have to know the true mechanisms governing the phenomenon of interest. As long as it is clear how the system is “connected” (i.e., which component affects which other components), either as a source of material or as the source of a regulatory signal, it is a straightforward procedure to set up a symbolic model. Here, “symbolic” means that we formulate the equations of the model, but that we do not know what the values of their parameter are; examples are Equations 9.1–9.4. The second advantage of BST models is that each parameter, even if it does not yet have a specific value, has a clearly defined meaning. An example is $f_{X_4}$, which exclusively represents the inhibition by $X_4$ of the conversion between $X_2$ and $X_4$. Similarly, it is immediately clear into which parameter a particular feature of a biological system has to be translated. This one-to-one relationship between biological aspects and parameters is very helpful both during the model design and the interpretation of results from the model analysis. The third advantage of the BST formulation is that the particular format has convenient mathematical properties. The main disadvantage of BST is that there is no guarantee that the model will capture all relevant features of the biological system of interest with sufficient accuracy. While the Taylor approximation is mathematically guaranteed to match any
9.3 Concepts and Challenges of Parameter Estimation

data close to a chosen operating point, this guarantee extends only over a very small range. Outside this range, which is of unknown size, the model may or may not fit well. BST shares this feature with all other models in biology.

BST models are just one option for nonlinear representations, but they are especially powerful as initial default models when not much is known about the details of the biological system. Specifically, since the construction of the symbolic equations is essentially automatic, the model design challenge is reduced to determining optimal parameter values with which the model matches the observed data well. This determination, while still difficult, is to be seen in comparison with the task of setting up a model with unspecified functions that are supposed to capture the dynamics of complex and ill-characterized processes. Many methods are available for parameter estimation purposes (for a recent review, see [4]), but none of them works perfectly or even satisfactorily in cases of moderately large biological systems.

9.3 Concepts and Challenges of Parameter Estimation

Methods of parameter estimation for biochemical systems fall into two broad categories, which are directly tied to the available types of data (e.g., see chapter 5 in [14]). In the past, the data almost always consisted of kinetic features associated with a particular step in the biochemical pathway system. Such features included the $K_{m}$ of the enzyme, sometimes a flux rate or $V_{max}$, a dissociation constant, or some other kinetic characteristic. Also, once in a while steady-state values for the variables or fluxes of the system were available. Given such data, the “bottom-up” strategy of parameter estimation consisted of formulating each step symbolically and optimizing the parameter values such that they matched the alleged shape of the reaction step. Subsequently, all “local” descriptions of individual reaction steps were merged into a system of differential equations that described the entire pathway, the equations were integrated, and the numerical solutions of the system were compared against additional observation data. As the comparison typically led to numerical inconsistencies, one had to go back to the individual process representations many times and adjust functions of parameter values. This iterated reformulation and data matching could easily take months if not years. The vast majority of all existing biochemical models have been estimated with methods of this type; a very detailed example is [15].

Recent advances in molecular biology have begun to offer an alternative “top-down” approach to parameter estimation. The data here consist of measurements of metabolite concentrations at many subsequent time points, usually following some stimulus. The estimation now occurs in the “opposite” direction. Namely, one attempts to determine parameter values such that the solutions of the differential equations match the observed time-series data. The local parameters, which earlier were the starting point, are now the result of the top-down estimation. This type of an estimation task is known in mathematics and computing as an inverse problem, for
which many algorithms are available. However, these algorithms tend to fail for moderately complicated biological systems [4].

As the identification of parameter values is a severe bottleneck of systems biology, enormous efforts have been dedicated to the development of general and specific techniques. One crucial issue is the fact that essentially all estimation algorithms are iterative. They use some method to determine a candidate set of parameter values, solve the differential equations with these parameter values, compare the solution with the observed data, and evoke some method for improving the parameter set for the next iteration. As an immediate consequence, the set of differential equations must be solved thousands of times and while each solution may be relatively fast, their collection can become prohibitively long. To circumvent this particular issue, two independent groups [16, 17] proposed almost 30 years ago the smoothing of the raw data and the interpretation of the slopes of the time course for each metabolite as estimates for the differentials on the left-hand sides of the differential equations (Figure 9.2). Thus, according to this method, each differential equation is replaced with a set of $K$ algebraic equations, where $K$ is the number of time points where the metabolite concentrations and slopes are measured or estimated. Each of the algebraic equations contains on the left-hand side the estimated slope value $S(t_k)$ at a given time point $t_k$ and on the right-hand side the expression given by the differential equation and evaluated at $t_k$. Pursuing this strategy, estimating the parameters $p_1, \ldots, p_M$ in the differential equation:

$$\dot{X}_i = f_i(X_1, X_2, \ldots, X_n; p_1, \ldots, p_M)$$

becomes a task involving a system of $K$ algebraic equations of the form:

$$S_i(t_k) \approx f_i(X_1(t_k), X_2(t_k), \ldots, X_n(t_k); p_1, \ldots, p_M), \quad i = 1, \ldots, n, \quad k = 1, \ldots, K$$

As the estimation is based on observed time courses, all quantities $X_i(t_k)$ and $S_i(t_k)$ are known or estimated, and the only unknowns are the parameter values of $p_1, \ldots, p_M$.

![Figure 9.2](image.png)  
**Figure 9.2** Generic univariate dataset (blue dots), a smoothing function (red), and examples of estimated slopes (indicated by green triangles) along the time course. The slopes thus estimated are used as surrogates of the differentials in the differential equations of the systems model.
This slope estimation strategy has two significant advantages. First, the need for solving differential equations is eliminated and the estimation is done instead with a purely algebraic system of equations. This conversion is very consequential, because in excess of 95% of the estimation time for a system of differential equations is spent on numerically integrating the equations, and this percentage can approach 100% [18]. Second, each time course is “decoupled” from the others, because it can be addressed independently of all other time courses. Intriguingly, if this method of slope estimation and decoupling is applied to Lotka–Volterra models, the initially nonlinear parameter estimation task becomes a simple linear regression [19]. Similarly, applied to S-systems within BST, the nonlinear estimation can be converted into an iteration of linear regression tasks [20] (see also [21]).

Clearly, the computational speed-up of the slope estimation and decoupling strategy is very appealing, but one might wonder about its statistical rigor. For a long time, the method was seen as such a convenient shortcut that it would outweigh possible concerns of statistical bias. The argument was that the solution obtained with this method could at least be used as a starting point for more conventional and possibly less biased estimations. However, Brunel recently showed that the procedure is asymptotically normal, consistent, and indeed statistically sound [22].

As the optimization of parameters in algebraic equations is much easier than in differential equations, one might be tempted to assume that the estimation task is essentially solved. However, this is not always the case. The slope estimation and decoupling strategy consists of two key steps—the smoothing of the data, which is necessary for the estimation of slopes, and the parameter estimation of the system of nonlinear algebraic equations. The smoothing step is typically achieved with splines [23], although more sophisticated methods have also been proposed for this purpose [24, 25]. The choice of a spline or another smoother necessarily requires a decision regarding the degree of desired smoothness, and this decision cannot be made with total objectivity, because the smoothness of the data—or the lack thereof—depends on the processes governing the system from which the data were obtained, including experimental noise. At the same time it is quite evident that the degree of smoothing will affect the second step of slope estimation. No matter which method is chosen, smoothing incurs an approximation error, which is generally larger for smoother splines that are of lower order and consist of fewer pieces. If more spline pieces or higher-order splines are used, the approximation error is generally lower, but the appearance of the smoothing function is “bumpier” (Figure 9.3). This bumpiness may be due to time courses with many true ups and downs or to experimental noise in the data. If the latter is the case, a bumpy smoother simply tries to mimic the noise, the slopes along the smoothed time course increase and decrease to an unreasonable extent, and the subsequent parameter estimation results in a larger residual error. Thus, if the data contain even moderate noise, the smoothing error and the parameter estimation error are inversely related to each other, and the two must be weighed against each other. Addressing this issue with statistical rigor, Ramsay et al. developed algorithms that optimize weights associated with the two types of error [26]. A remaining question in this context is to what degree
Figure 9.3 Different degrees of smoothing incur different residual errors. A smooth interpolation (green) generally results in a larger residual error with respect to the data (blue dots), whereas the more detailed smoother (red) is associated with a smaller error but a "bumpier" appearance, which may cause problems with the slope estimation and decoupling technique.

different extents of smoothing may lead to a loss of biological information, especially if the true "nature" of the data is unknown.

Overall, the slope estimation and decoupling strategy makes the estimation task simpler and much faster, but it does not solve all problems. We discuss some remaining issues in the following.

9.3.1 Typical Parameter Estimation Problems

The problems typically encountered in the estimation of parameter values fall into two classes. In the first class, the algorithm simply does not find a suitable solution and, as a consequence, the residual error is unacceptably high. In the second class, an algorithm does find a solution, but something is not quite right with this solution. We discuss the different cases one by one.

9.3.1.1 Data Fit is Unacceptable

If no satisfactory fit can be found, the reasons may be manifold. In a relatively clear-cut case, the algorithm does not converge at all, reaches the maximally permitted number of iterations, or produces a fit that is obviously very different from the observed data. In such cases, the reasons (although not necessarily only) are likely of a technical nature. It might be that the computer- or user-suggested initial guesses for all parameters are simply so bad that the algorithm does not reach a basin of attraction surrounding the optimal solution. It is also possible that the algorithm is attracted to an unacceptable local minimum.

A distinctly different cause for not yielding a good solution may be that the alleged functions in the model are so far from the truth that the algorithm cannot determine a satisfactory solution. In contrast to purely technical issues, the result in this case is often "some" fit that, however, is clearly not optimal. For instance, if one attempts to fit a Michaelis–Menten function to a sigmoidal time course, it is clear that the initial
Figure 9.4. An algorithm tries to fit a sigmoidal Hill function (H, blue) with a Michaelis–Menten function (M, red) that simply does not have sufficient shape flexibility. As a consequence, no adequate fit can be reached.

phase cannot be matched appropriately, because the structure of the Michaelis–Menten model is not equipped to capture S-shaped datasets (Figure 9.4). In simple cases like the one described, the problem is easily detected and diagnosed. However, this analysis is not so readily accomplished if a high-dimensional parameter space. Of course, all combinations of the above causes may be encountered. Many studies have focused on these and other technical issues.

9.3.1.2 Differently Structured Candidate Models are Difficult to Compare

Within the realm of linear regression, methods have been developed for assessing the relative worth of an additional parameter. Specifically, objective criteria exist, based on residual errors and numbers of parameters and data points, for deciding the superiority of one of two candidate fits where one involves $M$ and the other one $M + 1$ parameters [27]. For nonlinear estimation tasks, such comparisons are much more difficult, especially if different model structures are involved. Surprisingly, there is even ambiguity in the number of parameters. As an example, consider a Hill function of the type:

$$V(S) = \frac{V_{\text{max}} S^2}{K_M + S^2} \quad (9.7)$$

which has a sigmoidal shape as shown in Figure 9.4. One will easily recognize $K_M$ and $V_{\text{max}}$ as parameters, but should the Hill coefficient (i.e., the power associated with $S$) be counted, even if it is a priori set equal to 2? After all, it clearly affects the shape of the function and if 2 did not fit, we could easily change it to a different value. It is difficult to find an objective criterion accounting for this issue. Thus, it is in general not a trivial matter to compare two fits, such as one with a Michaelis–Menten model (which is a special case with the Hill coefficient equaling 1) and one with a Hill model with a fixed or tunable Hill coefficient. Similarly, it is difficult to compare fits with a Hill model and a logistic model, which can both capture sigmoidal processes equally well but have different mathematical structures (Figure 9.5).
9.3.1.3 **Fit is Acceptable, But...**

In spite of all technical challenges, one often obtains a fit that is good, as judged by visual inspection and an acceptable sum of squared residual errors (SSE), and it might appear that the problem is solved. In some cases this is the case, but a good SSE should not be taken as the sole criterion.

The best-known situation is convergence to a local minimum. Even if an optimization algorithm returns an acceptable fit, there is no guarantee that the fitted model is truly the best option. It is easily possible that the solution corresponds to a local minimum and that other solutions, maybe far away, are even better. However, the algorithm may not necessarily find these superior solutions because they are separated from the current solution by domains of parameter sets that correspond to drastically inferior model fits. Thus, any time the algorithm attempts to move toward the global optimum, it hits the separating areas of high errors, deems the direction futile, and searches in other parts of the parameter space. A partial, but not always effective solution to the problem is repeated optimization with different initial guesses for some or all parameters.

A slightly different situation occurs if the identified optimum is surrounded by a large area of solutions with very similar errors. In the simplest cases, these “almost optimal” solution sets form slightly distorted ellipses in a higher-dimensional space, but this is not necessarily so. Recent years have seen quite a bit of attention dedicated to what is now called “sloppy” solutions (e.g., [28, 29]). One could argue that the optimized solution is still a tiny bit better than its neighbors in the sloppy set and that it should therefore be preferred over all other candidates. However, in some cases the residual errors within the sloppy set are so similar that a slight change in just one data point, which could easily correspond to experimental noise, would identify a different solution as optimal. Thus, one should not discard the range of parameters surrounding the optimized solution. In a way, sloppy solution sets are not necessarily a cause for concern, because the corresponding models are quite similar in their fits to the data. In fact, it might be possible to simplify the optimized solution by setting certain
parameters equal to zero, if the sloppy set includes this setting as a legitimate possibility. The SSE would very slightly increase, but arguments of simplicity might favor this increase in exchange for a simpler model structure. Intriguingly, if the sloppy set permits positive, zero, as well as negative values for a certain parameter, the interpretation of the corresponding models differs. For instance, within the context of BST, a negative kinetic order is to be interpreted as an inhibitory signal, while a positive kinetic order suggests an activating or augmenting influence and a value of zero corresponds to an unimportant role of the parameter. This result poses the interesting question of whether sloppy solutions of this type are computational artifacts or whether inter-individual variation could go so far as to allow activation in one organism and inhibition in another [30]. At present, this question cannot be answered with confidence. Finally, while Ockham’s razor might suggest that the simpler solution is to be preferred, biology has presented us in many cases with solutions that initially seem more complex than necessary. Later we may find that the more complex solutions are preferable because of improved robustness or other higher-order features.

Somewhat related to sloppy solutions is the situation where the data are not plentiful of informative enough to allow a precise determination of all parameter values. A very simple example arises if two parameters $p_1$ and $p_2$ always appear in the model in the same constellation, such as the ratio $p_1/p_2$. Clearly, by multiplying the two parameters with a nonzero factor, the ratio is unchanged. Turning the argument around, a search algorithm cannot find a unique solution in this situation, but only one where the ratio fits well. In these cases, the solution is “structurally non-identifiable,” which entails that infinitely many combinations of parameter values can yield solutions with the same SSE. It is recommended to remove these identifiability issues, for instance, with methods of model reduction [31].

Quite a different concern with an apparently good fit is the occasional identification of parameter values that are impossible or unreasonable from a biological point of view. The most obvious case is an optimized value that turns out to be negative, although the parameter must be positive. Examples include rate constants and $K_m$ values. If this situation occurs, the optimization should be redone with corresponding constraints on the parameters. A more subtle variation of this issue is a resulting model that is unstable or extremely sensitive. It may happen in this case that a very small percent change in a parameter value would lead to dramatic changes in an important system feature such as the steady state. It is rare that such a high sensitivity is realistic in biology. Whether the situation is caused by a faulty parameter or by the misidentification or omission of some process or regulatory signal cannot be said without further analysis.

Another common and generic issue with an apparently good fit is often overlooked at first and becomes an inconvenient surprise later. This issue is the frequent inability of the parameterized model to predict responses to new stimuli in an adequate fashion [32]. Thus, the original data are matched quite nicely and the model traverses the cloud of data points seemingly fine. It may also be able to predict responses of the model throughout a modest time period beyond the measured time points. However, if new observations are to be modeled, the model may woefully fail. How is that
possible? One explanation is that the fitted curve constitutes the intersection of two (usually curved) surfaces— one describing the model function with optimized parameter values and the other representing the true dynamics of the biological system (see Figure 9.6). If the data points are close to this intersection, the fit not only appears to be good, but actually is so. However, analyzing new data means moving along the true surface and away from the intersection, and it becomes significant that the two surfaces diverge. As a result, if the extrapolation differs substantially from the original data, the model is no longer able to capture it. Again, several root causes for this situation are possible. One is the compensation of error among different terms within a system of equations. As an example, consider an equation with two terms, of which one is modeled quite badly. It is not difficult to imagine that, at least in certain situations, the other term could compensate for the error. In fact by also being modeled badly, the overall results can be surprisingly good. At first glance, it appears that two wrongs could indeed make a right. However, the error compensation usually no longer holds for new situations where the system variables in the two terms of the equation change to different degrees. As consequence, the fit to the extrapolated dataset can be unacceptably bad. To some degree, the method of dynamic flux estimation can remedy the extrapolation problem [32]. However, this method requires ideal conditions that are seldom satisfied and therefore requires additional information, which is not always easy to obtain [33, 34].

Finally, a genuine challenge with the otherwise appealing slope estimation and decoupling technique is that dependencies among equations are a priori ignored. For instance, if the same parameter appears in two equations, the decoupling causes
9.3 Concepts and Challenges of Parameter Estimation

9.3.1.4 Needed: A Better Fit Or Not?
The previous sections have demonstrated that even an apparently good fit may not be the ultimate solution, because, for instance, it may not hold up in extrapolations. If so, what should be done? This is a complex question, and as in so many cases the answer is: it depends. The need to search for other solutions depends on the purpose of the model and the availability of appropriate data.

In some cases, even symbolic models, in which parameters are not at all numerically defined, can yield valuable insights. In fact, if insights can be gained for arbitrary parameter values, they are usually more general than results obtained for certain numerical parameter combinations. A beautiful example is the exploration of design principles [35]. In this line of research one asks what the role of a particular feature of a biological system is. For instance, one might observe a feedback inhibition signal in a pathway system and ask why it is there. According to the method of controlled mathematical comparisons (MCMC), one sets up two essentially identical systems models in parallel; however, one represents the observed signal and the other one does not [36]. The responses of these two models are compared with respect to performance criteria, such as robustness and response time. Studying a whole roster of such features and results, one system design is ultimately declared superior, either in general or within a certain environment. Many such comparisons have been performed without the specification of numerical parameter values, while other comparisons required the definition of relevant ranges of parameters. Using this MCMC strategy, Savageau proposed general rules for the regulation of bacterial gene circuits that have held up in all cases tested so far [37, 38]. These rules were independent of specific parameter values and identified the superior circuit designs primarily based on demands exerted by the environmental of the bacteria. Other examples included more complex gene circuit designs and a variety of other system structures [39–44]. In the context of signaling cascades, the structural design and performance demands even helped determine the ranges of effective parameter values [45].

Trusting in the observation that accurate parameter values are often not as critical as the correct model structure, Alves et al. exhaustively evaluated likely parameter ranges on discrete grid points, thereby yielding sufficiently good, although maybe not optimal solutions [46]. Parameter ranges can also be restricted by biological and clinical constraints, and coarse solutions can be refined by means of simulations. In spite of the rather uncertain nature of this process, the model results can be surprisingly strong. For example, in an effort to construct a model of dopamine metabolism in the human brain, Qi et al. collected semiquantitative input from clinicians, biochemists, and toxicologists regarding the relative concentrations of relevant metabolites and the flux split ratios at diverging branch points in the pathway system [3]. This information turned out to be sufficient for setting up complex pathway models with coarse parameter values. Even though these models did not capture the precise numerical features of the biological system, they helped explain
the inner workings of the system. For instance, Qi et al. used these models to perform
exhaustive analyses of root causes leading to Parkinson's disease and schizophrenia [3, 47, 48]. Models of this type can also indicate whether the responses to targeted
alterations occur in the right direction. Thus, even if the parameter values are coarse,
some types of valuable insights may be gained.

Alas, there are situations where the parameterization should be as good and
reliable as possible. For instance, it is becoming possible to develop health and
disease models, which could be used for designing specific, and maybe even
personalized, treatment strategies [49]. Clearly such strategies need to be quantitative
and accurate enough to allow at least modest extrapolations from the normal state.
These models should also permit reliable predictions of what might happen to an
individual if s/he is or is not treated for some abnormality. One might be tempted to
discard such predictions into the future, because our current models are simply too
inexact. However, even short-term predictions can be extremely beneficial. For
instance, a 5-min warning that a critical care patient is diverging from the normal
trajectory might be sufficient to initiate efficient countermeasures [50]. Such short
time horizon seems to be within reach of computational models, even if they are
based on relatively coarse approximations.

If the fit is good with respect to the SSE of one data set, but fails in extrapolations,
the situation is dire, because there is no general diagnostic tool for identifying where
the problem lies or whether the reason of failure is a combination of several
problems. To some degree, the modeling process needs to be restarted in such a
case. Of course, the process does not really start at the very beginning, because a
reasonable model for at least one situation is now available and this model may serve
as a starting point or as a constraint for further model development. One might also
expect both, greater reliability of the model and insight into the true nature of the
biological processes if more and ideally diverse datasets are available that cover a
greater portion of the space of variation in the system variables. Nevertheless, the
difficult question at this point is whether the chosen functions in the model are
appropriate or not. If they are, renewed parameter estimation with additional data
might lead to a better numerical model implementation, but if they are not, new
functions need to be determined. This task can be very difficult, because the structure
of these functions is a priori unclear and because one usually does not even have data
regarding the individual functions, but only regarding the entire dynamics of the
system, which is governed by numerous processes simultaneously [34].

One strategy is to use biological insights to find functions that might be appropriate
representations of individual processes. Accordingly, one designs a model with more
complex, biologically relevant functions, and fits it to the original data and to the new
data at the same time. The original, simpler model may be evoked as a constraint in a
sense that the new function should coincide with the original model function (and
possibly its slopes) with respect to the original data. In principle, this strategy seems
to make sense, but in reality it often requires considerable effort [34]. Ultimately, this
strategy leads to the question of when enough is enough. Is it really desirable to
develop a very complicated process description that now captures two datasets? It is
known that even slight overparameterization tends to lead to extrapolation problems
and the more parameters are involved in a functional description, the more severe the potential problem is.

An alternate strategy is an extension of the concept of generic approximations. Experience has shown that extensions toward higher-order Taylor terms become mathematically cumbersome and while they naturally fit better, the facility of their analysis is compromised [6]. Instead, one may employ a piecewise approximation [51]. This strategy is straightforward in principle, but realistically requires the determination of breakpoints, at which one approximation is substituted with another. For univariate functions, the breakpoints can almost be determined by visual inspection or with a simple algorithm, but the determination is much harder for multivariate functions. For linear and power-law functions, algorithms have been developed that automate this process [52, 53]. The result is an approximation, consisting of a certain number of pieces, throughout which the overall SSE is within desired bounds. The actual estimation of parameter values for these piecewise approximations requires considerably more data and of course the number of parameter values grows with each added piece. Nevertheless, this strategy is relatively unbiased and therefore may offer a first default in situations where biology does not suggest candidate functions or where such candidates are so complicated [54] that there is hardly a chance that parameters could be estimated without serious overfitting.

9.4 Conclusions

The goal of parameter estimation is clear – find numerical values that render a model optimal for the representation of a biological system. While clear in principle, the task is often convoluted in practice. In the past, most data were coarse and scarce, and simple model fits had to be considered adequate. With recent advances in molecular biology, the situation has changed and some data are so good now that it is difficult to excuse bad fits.

As parameter estimation is so important, many groups have devoted substantial effort to it. In most cases, these investigations focused on the substantial technical challenges associated with the task and on algorithmic improvements. In addition to distinctly different optimization methods, which include regression, simulated annealing, and numerous variants of genetic colony, and swarm algorithms, it has turned out that the estimation of slopes and the subsequent conversion of differential into algebraic equations is very beneficial. While these methodological and algorithmic improvements have made parameter estimation manageable in principle, they have not solved all problems. There are still issues with slow convergence or the trapping of algorithms in local minima.

Beyond the well-known technical issues, even apparently good fits should be subjected to additional muster. Criteria for such additional tests should be the reasonableness of the numerical values of all parameters, model stability, sensitivity, and robustness, and the ability to provide good fits in extrapolations. Possible causes for models to fail these criteria are plentiful. Particularly hideous among them is the
compensation of fitting errors within and among the system equations, because such compensation is not always easy to diagnose and remediate. The recent method of dynamic flux estimation [32] is able to help, but only under ideal conditions, and it will be necessary to expand such methods toward more realistic conditions.

If a fit with a good SSE is obtained, but extrapolations cause problems, the entire model structure may have to be revisited, including the choice of functions, the numbers and types of parameters, and the availability of data. Clearly, several replicates of time-series data and data obtained under different conditions will allow better estimations, along with their statistical analyses and interpretations. Ultimately, one must consider the purpose of the model and judge the real need for accurate parameter values. Indeed, a simple model with fewer parameters is often more robust and less sensitive to overparameterization, yet may yield as much, if not more insight.

Acknowledgments

The author expresses his gratitude to Yun Lee for help with one of the figures. This work was supported in part by a National Heart, Lung, and Blood Institute Proteomics Initiative (contract N01-HV-28181; D. Knapp, PI), a Molecular and Cellular Biosciences Grant (MCB-0517135; E.O. Voit, PI) from the National Science Foundation, and a Biological Energy Science Center grant from the Department of Energy (M. Keller, PI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the sponsoring institutions.

References


Abstract

The identification of functions and their parameter values for biological systems models is currently among the most severe bottlenecks in computational systems biology. While many papers and books have discussed the advantages and pitfalls of estimation algorithms of various types, including nonlinear regression, genetic algorithms, ant colony and swarm optimization, and simulated annealing, issues beyond the technical challenges of parameter estimation have seldom received much attention. This chapter discusses several situations where the computationally determined model fit to experimental data is satisfactory with respect to the sum of squared residual errors, but unsatisfactory or undesirable for other reasons. Examples include unrealistic parameter values, nonunique or sloppy solutions, and the inability of a fully parameterized model to make reliable predictions regarding so far untested scenarios.

Keywords: biochemical systems theory; extrapolation; inverse problem; model identification; parameter estimation; sloppiness.
Author Query

1. Update [19]?