Modelling metabolic networks using power-laws and S-systems

Eberhard O. Voit

Integrative BioSystems Institute, 313 Ferst Drive, Suite 4103, Georgia Institute of Technology, Atlanta, GA 30332-0535, U.S.A.

Abstract

Mathematical modelling has great potential in biochemical network analysis because, in contrast with the unaided human mind, mathematics has no problems keeping track of hundreds of interacting variables that affect each other in intricate ways. The scalability of mathematical models, together with their ability to capture all imaginable non-linear responses, allows us to explore the dynamics of complicated pathway systems, to study what happens if a metabolite, gene or enzyme is altered, and to optimize biochemical systems, for instance toward the goal of increased yield of some desired organic compound. Before we can utilize models for such purposes, we must define their mathematical structure and identify suitable parameter values. Because nature has not provided us with guidelines for selecting the best model design, the choice of the most useful model is not trivial. In the present chapter I show that power-law modelling within BST (Biochemical Systems Theory) offers guidance for model selection, construction and analysis that is otherwise difficult to find.

Introduction

A biomathematical model is a construct in the language of mathematics that represents something in biology. Modelling is the art of coming up with this representation and of squeezing new insights out of it.

1 eberhard.voit@bme.gatech.edu
Biomathematical modelling can be very challenging. It is apparently harder than rocket science, because we have successfully shot rockets to the moon and to predefined far-away locations in outer space, but we have not yet been able to develop reliable mathematical models of cancer, embryonic development or the human brain. Then again, rocket science has had a long history, at least if we trace it back to the ballistic studies that guided the production of cannons in the Middle Ages or to the relatively simple calculations that must have supported the advancement of Roman catapults, let alone the trials and errors of Stone Age hunters throwing rocks at deer and rabbits. In comparison, biomathematical modelling is a very young field indeed. So, for the time being, we might be well-advised to scale-down our modelling aspirations and realize that the main task of our generation may be the creation of a solid foundation from which cancer will be controlled one day, based on computational guidelines. Furthermore, not fully understanding the human brain should not imply that present-day modelling is not useful. It simply means that the insights that modelling has yielded so far have been more modest. Once we accept these caveats, modelling becomes more realistic and can be a very valuable companion of experimental work. Even in its simple implementations it can reveal aspects of biology that would be difficult if not impossible to gain otherwise. For instance, once a model is established, it is fairly easy to execute millions of simulations of what is probable, improbable or impossible. Such comprehensive studies are hardly possible with purely experimental means.

Modelling comes in a multitude of variations, and many books address methods, options and applications (e.g. [1–5]). The present chapter focuses crisply and narrowly on one specific type of modelling for networks and systems that is sometimes called canonical. The term implies that the process of setting up and analysing a model follows strict rules [6]. Of course, rules can be limiting, and one criticism of canonical modelling is that some options seen in other models are not available or discouraged. However, there are good reasons for adhering to the rules of this approach, at least initially. The main advantage of following a good set of guidelines is that they allow us to get started very quickly, as I will discuss in detail below.

For now, supporting our claims of the usefulness of canonical models with a double-negative argument, let’s ask ourselves what happens if we do not have such rules and guidelines. To answer this question, imagine being faced with the task of coming up with a mathematical model describing how the up-regulation of a gene leads to some physiological response. Making the task more specific, suppose it is our goal to model how a bacterium responds to the lack of glucose in the medium by activating transporters for other sugars. This is a very narrowly defined task, yet it is immediately clear that a fair number of ‘players’ are involved in this response, from glucose sensors to signal transduction systems, transcription factors, genes, mRNA, ribosomes, proteins, energy providers, and a lot of other cellular and subcellular components and
processes. In the scheme of nature, sugar uptake is just a minute system, but it is evident that developing a mathematical representation even for such a small system would be a real challenge. How can we even get started? This is where canonical modelling is most useful.

To be clear, canonical modelling is not the answer to all questions, but it provides enormous help setting up initial models with minimal effort, especially when we do not really know what mathematical functions optimally describe the processes taking place in the system. In physics, mathematical formulations are often directly derived from ‘first principles’ or ‘laws’ that very precisely describe basic phenomena like forces, velocity, voltage, cooling or light refraction at the surface of a prism. The mathematical descriptions of these elemental processes can then be used as building blocks for much more complex situations. In biology we do not have such laws, and borrowing laws from physics and chemistry becomes too cumbersome. Just imagine how many laws of physics and chemistry come into play when the presence of a transcription factor leads to the up-regulation of a gene! Sure, physics and chemistry govern biological processes, but accounting for every physical and chemical detail is simply too much, and also usually not needed. In response, biologists and biochemists have developed mathematical descriptions, such as the Michaelis–Menten rate law for enzyme-catalysed processes. These representations are not as fundamental as physical laws, but have become standard descriptions. However, in many cases these defaults are not applicable, or there are difficult questions about which of the traditional formulations might or might not be appropriate [7]. Canonical modelling circumvents these problems by using convenient approximations that are mathematically rigorous and guaranteed to be appropriate, at least if the variables don’t fluctuate too much [8]. Let’s see how that works.

Setting up canonical models of networks and systems

The definitions of networks and systems are blurry. For our purposes we consider a network as a collection of nodes (pools, players, metabolites, system components etc.) and edges (conversions, transport steps, reactions, processes etc.) connecting the nodes. A system is a network that is regulated. Thus, in addition to nodes and edges, a system contains signals that are sent out by players inside or outside the system and affect the magnitude, activity or speed with which material flows through an edge. As an analogy, we might imagine a network of water pipes under a city. This network is the cornerstone of the water distribution system of the city. However, water is not just distributed according to the size of each pipe, but in addition to the pipes the system contains valves that control how much water flows when and where. The difference between just the network of pipes and the functionality of the water distribution system is easy to see if we imagine that local, daily changing water demands, feedbacks from sensors indicating blockages or pipe breaks,
and also human operators control the valves, and thereby the flow distribution. In biology, such feedbacks and controls are often part of the system, with the result that the system ‘automatically’ adapts to a new situation by satisfying changing demands and diverting excesses toward other uses.

To exemplify the role of regulation more simply, let’s look at the simple pathway in Figure 1(A); the equations are given in 1(B). Notice that two parameters (α and f) are left unspecified in order to allow us a comparison between two scenarios. In the first scenario, X₃ has no effect on the influx. We set α=1 and f=0, which results in X₁ = 0. Thus simply corresponds to a constant influx with magnitude 1. In the alternative scenario (α=0.02 and f=-1), X₃ controls the influx as a feedback inhibitor; the more X₃ is present, the stronger the reduction of the influx. Let’s see what difference the inhibitor makes. Our test scenario is the following. It begins with the values X₁ = 200, X₂ = 50 and X₃ = 100. These were not arbitrarily chosen but constitute the steady state of the system, which means that material is flowing into and out of the system, but that the amounts of the three variables stay constant over time. We can see that this combination of values is indeed a steady state by looking at the first 100 time units of the graphs in Figures 1(C) and 1(D), which show that all three variables remain constant. At time t=100, we simulate a sudden drop from 200 to 120 in variable X₁. Such a drop could be due to a situation where X₁ is siphoned off for some other demand. In response, X₂ and X₃ also decrease, but subsequently return to the original steady state, which is more or less restored at approx. t=800. At time t=1000, some outside force causes X₁ to increase to 300. Again, X₂ and X₃ respond to this perturbation and the system is more or less ‘back to normal’ at the end of the experiment (t=2000).

Figure 1. Model of a simple pathway with or without inhibition

(A) Diagram on which the model is based. (B) Model equations. The free parameters α and f allow for a comparative analysis of the role of inhibition. (C) Responses of the model without inhibition. (D) Responses of the model with inhibition.
The differences between Figures 1(C) and 1(D) reflect the inhibition by $X_3$ in the latter case. We can see that the feedback control mechanism has several effects. First, there is a slight overshoot, especially in $X_1$: after its dip, $X_1$ actually exceeds its steady-state value before returning to it. Secondly, the responses are faster. For instance, we may ask how long it takes before $X_3$ is restored to within 5% of its normal state. Without feedback, this happens at $t = 697$ and $t = 1645$ for the two perturbations respectively. With feedback, $X_3$ is back within 5% at times $t = 459$ and $t = 1410$. Depending on the specific situation, a faster response might not be important, but it could also mean a critical advantage, improved fitness and possibly even survival. It is hard to see from the graphs, but the feedback control also softens the response in a sense that $X_2$ and $X_3$ do not deviate from the normal state as much as in the unregulated case. In a later section we will see a more complex example where regulation leads to entirely new behaviours. The important point of this first, simple example is to get an impression of the conversion of some biological phenomenon into mathematics and of some type of simple analysis. But we are getting ahead of ourselves.

The design process of a canonical model is clearly structured [9,10]. It begins with a diagram (as in Figure 1A) that shows the players and their interactions. Such a diagram has nothing to do with mathematics but requires sufficient biological insight. In a way, this diagram is the most important part of the model because, if it is wrong, the model will also be wrong.

**Step 1**

As the first model construction step, the information in the diagram is converted into several lists. The first list contains the main players, whose size, amount or concentration changes over time. In a biochemical systems model, typical main players are metabolites; in an ecological system, the main players may be population sizes of species. In canonical modelling, the main players are called dependent variables, because the quantity representing each such variable depends on what is happening in the system. The second list contains important components that do not change over time, at least not within the time period of any anticipated experiment. In a metabolic system, these components may be enzyme activities, constant influxes or factors and co-factors that are necessary and assumed to be available, even though one does not really care in the model where they come from and what happens to them if they are no longer needed. These components are classified as independent variables. Sometimes the categorization into dependent and independent variables is not obvious at the beginning and it happens quite frequently that a variable eventually moves from one list to the other. Nevertheless, to get started we need to define these lists, and each important contributor to the system has to appear on one and only one of them. The third list is initially not quite so important. It contains quantifiers for external or internal conditions. Typical are pH, temperature, characteristics of buffers, and other physical and chemical quantities that may be determinants for what happens in the system.
and that are necessary but not of primary interest. The last list to be collated is actually more like a matrix, that is, two-dimensional. Here we need to establish for each dependent variable \( X \), which of the (dependent or independent) variables \( Y_j \) have a direct effect on \( X \). More precisely, we distinguish whether any given \( Y_j \) affects the production or the degradation of \( X \). This information can be read directly off the diagram. An effect may be positive (augmenting, activating) or negative (inhibiting). For instance, if the dependent variable \( X \) is produced from material in pool \( Y \), and if variable \( I \) inhibits this production, then the list of variables affecting the production of \( X \) includes \( Y \) and \( I \).

At one point, either now or later (or both), the lists are to be extended to include further information that, for instance, quantifies the normal size, amount, concentration or activity of a variable, and the strength or importance of effects that are exerted by one variable onto another. In the end, the lists should generally include all types of data or observations that might help us put numbers on components or processes in the system.

**Step 2**

To construct model equations, we focus on one dependent variable at a time and formulate a very simple equation of the form

\[
\dot{X} = \text{production} - \text{degradation}
\]

This is a differential equation, but that should not worry us, because it will be solved by computer. Independent variables are constant and therefore not modelled as differential equations. To specify the production and degradation terms, we retrieve from the list of interactions all variables directly affecting the production (or degradation) of the variable under investigation. As an example, consider variable \( X_1 \) in Figure 1(A). Its production is affected only by \( X_3 \) and its degradation only by \( X_1 \) itself. The rule for each term is simple: every variable that directly affects this term is to be included as a factor and raised to some as yet unknown power, called a kinetic order. Furthermore, one multiplier, called a rate constant, is to be included in each term. Thus the production and degradation terms for \( X_1 \) are:

Production = \( \alpha_1 X_3^{g_13} \)

Degradation = \( \beta_1 X_1^{h_11} \)

In the process, we have introduced some conventional nomenclature: the rate constants of the production and degradation terms are called \( \alpha \) and \( \beta \) respectively, and the kinetic orders are called \( g \) and \( h \) respectively. Rate constants are always positive or zero. Kinetic orders are positive (negative, zero) if the represented effect is positive (negative, zero). Each rate constant carries one index (subscript) that indicates to which equation (dependent variable) it belongs; in our case 1. The kinetic orders carry two indices. The first refers again to the equation (1 in the production and degradation terms), and the second to the variable that exerts the effect (3 in the production term...
and 1 in the degradation term). That’s it. If a production term, say of $X_4$, contains two variables, $X_2$ and $X_3$, the power-law representation is:

\[
\text{Production} = \alpha_4 X_2^{g_{42}} X_3^{g_{43}}
\]

and we can see immediately that this is a generalization of the typical mass action representation, which would be:

\[
\text{Production} = k_4 X_2 X_3
\]

Similarly, if the degradation of variable $X_5$ in some model depended on $X_1$, $X_5$ and $X_6$, we would know immediately what the corresponding term in the canonical model would look like, namely:

\[
\text{Degradation of } X_5 = \beta_5 X_1^{b_{51}} X_5^{b_{55}} X_6^{b_{56}}
\]

Thus the process is straightforward and can even be executed by a computer [11]. Following through with this construction for all dependent variables results in a canonical S-system model that contains in each equation one or none production terms and one or none degradation terms. When pondering the process so far, recall how difficult we found the construction of equations for some biological phenomenon in the absence of guidelines.

One should note that a few variants of this type of model are possible; they have been discussed at length in the literature [9, 12, 13]. One should also point out how straightforward this model design process is in comparison with other methods. For instance, if we wanted to use a traditional enzyme kinetic model for the degradation of $X_5$, we would have to determine the specific mechanism of the process and its modulators, and every choice could lead to a distinctly different format, as it is succinctly illustrated in [7].

**Step 3**

The really hard part of any type of modelling is the determination of numerical values for all parameters, and canonical models are no exception. This parameter estimation task is difficult for several reasons, including the lack of the right types of data and computational issues. Three classes of estimation procedures are available for canonical S-system models. The first relies on experiments where one variable is altered and the changes in production or degradation terms are measured. If such experiments are available, the kinetic orders can be determined with linear regression, and the rate constant follows relatively easily [14]. The second method requires time series experiments. In this case, the biological system is slightly perturbed and one measures all metabolite concentrations at a sequence of time points afterwards. It is easy to imagine that these types of experiments are not trivial, but they are emerging in the literature in growing numbers. The estimation of parameter values from such data constitutes a difficult computational task, and many groups around the world are trying to solve this issue (for references see [15–21]). The third method is somewhat indirect, but owing to the types of
experiments that biochemists and biologists have executed in the past, the
majority of actual estimations of canonical models have used this method. It
consists of reformulating, as power-law functions, other types of production
and degradation functions that are found in the literature and for which
experimentalists have measured parameter values, such as Michaelis–Menten
rate laws (e.g. [22,23]).

Because parameter estimation is a complex issue, we will skip further
details here and assume that we had somehow managed to obtain appropriate
parameter values.

**Working with canonical systems models**

Suppose we have established our model and obtained all necessary parameter
values. The next tasks fall into two categories: (i) model diagnostics and
refinements; and (ii) model utilization.

**Diagnostics and refinements**

Even if a model (of any type) may have been constructed from the best
available knowledge and no mathematical errors had occurred, the result is
not always a satisfactory model. Possible reasons are unlimited and include
reliance on faulty data, mixing of data from different experimental conditions
or even organisms, missing components or processes, or misjudged importance
of some of the players. Indeed, it is generally not easy to find out how good
a model really is. Ultimately every model is based on simplifications and
omissions so that every model will eventually fail. Furthermore, the right data
for validation are often missing, making model diagnostics somewhat indirect.
The typical tool of the trade for weeding out models that are probably wrong
is sensitivity analysis, and to some degree stability analysis [9,12,14]. Sensitivity
analysis investigates the following question: if one changes one of the structural
or input components of the system by a small amount, how much will some
output feature of the system be affected? For instance, if a rate constant is
decreased by 5%, how much will the steady state of the system change? If the
answer is 20-fold, then there is probably something wrong with the model,
because variations of a few per cent happen regularly in nature and if they
would result in such dramatic changes in the functioning of the system, the
system would probably be prone to catastrophic failure. The ‘input’ features
considered in sensitivity analysis typically include all parameters as well as
the independent variables, whereas typical ‘output’ features are steady-state
concentrations, fluxes through variables or features of the time courses
between a perturbation and the return of the system to the steady state. In
dramatic cases, the system may become unstable and not even return to the
steady state but ‘die’ in a sense that one or several variables become zero. Some
of the diagnostics can be done with algebra and calculus, whereas the majority
of aspects require computational methods, including massive simulation
studies of very many possible situations [9].
If a model is deemed too sensitive, if it loses stability where it shouldn’t, or if it simply gives answers that are not consistent with observations or well-founded expectations, amendments and refinements are in order. Often sensitivity analysis provides clues for how the model is to be altered to remedy a particular problem. In general, however, model improvements are not always easy to conceive and require a lot of thought and effort (see [9] for details).

**Model utilization**

A good model can be used in a number of ways. One class of investigations asks what is within the realm of the model, what is probable, what is possible and what is not. Some parts of such an investigation can be accomplished with purely mathematical means, but much will usually rely on computational methods. As an example, consider the small gene regulatory system in Figure 2(A). Suppose gene $X_1$ activates itself as well as gene $X_2$, whereas gene $X_2$ suppresses $X_1$. It is difficult to predict with intuition alone how this little system responds to perturbations. Without regulation, the two genes would act independently, but here they obviously control (or at least affect) each other’s function. Does the system have a stable steady state? Does up-regulation of one gene lead to up- or down-regulation of the other gene? Will the system begin to oscillate? Following the design steps above, it is easy to construct an S-system model, and for the time being we just assume parameter values as shown in Figure 2(B). It is straightforward to compute the steady state (either mathematically or with the freeware PLAS [24]); it has the values

![Figure 2. Model of a simple gene circuit in which two genes affect the expression level of each other](image-url)

(A) Diagram of the gene circuit. (B) Model equations. Changing the free parameter $\alpha_1$ can lead to qualitatively different model responses (C and D).
Now suppose that $X_1$ is up-regulated to 1.2 at $t=10$. The simulation shows that the system rather quickly returns to its original steady state after some oscillations ensue (Figure 2C). However, the system is very sensitive. If we change the rate constant $\alpha_1$ from 0.9 to 1.02, the response is entirely different. Now the oscillations increase, and the system eventually enters a pattern of stable oscillations with the same period and amplitude (Figure 2D): the system never returns to its steady state! Mathematical analysis [25] reveals that there is a critical threshold for $\alpha_1=1$, where the dramatic change in system behaviour happens. The important take-home message is that the different responses and the critical threshold cannot be detected from the system diagram (Figure 2A) with hand waving and casual argumentation, but that the different responses depend on a complicated combination of the numerical values of some key parameters that can only be deciphered with a thorough analysis.

Related to determining possible system behaviours is the question of choke points. These are processes that most significantly limit the flux through the model. In other words, if a choked process could be enhanced, the flux through the entire system or through subsystems of particular interest could be substantially increased. Such choke points are of significance as potential drug targets or as system components (enzymes, genes) whose alteration could potentially lead to greater yield, for instance in a biotechnological system where some valuable organic compound is being produced.

As a continuation to the task of manipulating a system toward some desired goal, one may ask whether it is possible to optimize the system with respect to a stated objective. For example, citric acid is being produced worldwide at a rate of over one million tons per year, mostly with the black fungus *Aspergillus niger* [26]. Much effort has been invested over the past century to optimize the process. For future improvements or similar optimization tasks, it is at least theoretically possible to set up models describing citric acid production in the fungus and with mathematical means to optimize yield while ensuring that no metabolic processes needed for survival are unduly affected. If one uses canonical S-system models, the mathematics behind such an optimization task is actually not all that complicated [12,22]. The main difficulty is again parameter estimation.

Presently the most academic, but ultimately the most important and applied goal, is to discover the design and operating principles with which nature solves tasks. The generic questions are: why does nature solve a task using mechanism A instead of mechanism B? Why is the production of some component inhibited and not its degradation activated? Why are some genes regulated with repressors and others with inducers? Canonical models are particularly well suited for these types of analyses because they facilitate the construction and comparative analysis of models of observed designs as well as of different, hypothetically plausible designs [27–29]. A glimpse of the power of the careful analysis of alternative designs is termed Demand Theory [30], which accurately predicts the mechanisms with which microbial genes are regulated.
Summary

- The selection of a biochemical model is important and not trivial.
- BST (Biochemical Systems Theory) provides guidance for model design and analysis.
- System representations based on power-law models have great potential.
- Many biochemical phenomena have been analysed successfully.
- Even seemingly simple models can exhibit surprising responses.
- The long-term goal of modelling is to understand design and operating principles.
- Further information is available at www.bst.bme.gatech.edu

This work was supported in part by a Molecular and Cellular Biosciences Grant (MCB-0517135 to E.O.V.) from the National Science Foundation.

References

© The Authors Journal compilation © 2008 Biochemical Society


