Comparing systemic properties of ensembles of biological networks by graphical and statistical methods

Rui Alves\(^1,2,3\) and Michael A. Savageau\(^1,*\)

\(^1\)Department of Microbiology and Immunology, University of Michigan Medical School, 5641 Medical Science Building II Ann Arbor, MI 48109-0620, USA, \(^2\)Grupo de Bioquimica e Biologia Teoricas, Instituto Rocha Cabral, Calçada Bento da Rocha Cabral 14, 1250 Lisboa, Portugal and \(^3\)Programa Gulbenkian de Doutoramentos em Biologia e Medicina, Departamento de Ensino, Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, 1800 Oeiras, Portugal

Received on July 28, 1999; revised on November 11, 1999; accepted on December 3, 1999

Abstract

Motivation: When dealing with questions that concern a general class of models for biological networks, large numbers of distinct models within the class can be grouped into an ensemble that gives a statistical view of the properties for the general class. Comparing properties of different ensembles through the use of point measures (e.g. medians, standard deviations, correlation coefficients) can mask inhomogeneities in the correlations between properties. We are therefore motivated to develop strategies that allow these inhomogeneities to be more easily detected.

Results: Methods are described for constructing ensembles of models within the context of a Mathematically Controlled Comparison. A Density of Ratios Plot for a given systemic property is then defined as follows: the y axis represents the value of the systemic property in a reference model divided by the value in the alternative model, and the x axis represents the value of the systemic property in the reference model. Techniques involving moving quantiles are introduced to generate secondary plots in which correlations and inhomogeneities in correlations are more easily detected. Several examples that illustrate the advantages of these techniques are presented and discussed.

Contact: Savageau@umich.edu

Introduction

The only rigorous way to characterize and compare alternative biological designs for a particular class of systems is through the use of mathematical models and quantitative methods of analysis. In pursuing these goals we must address three critical issues. First, biologically meaningful behaviors must be identified (or, as is more commonly the case, hypothesized) and characterized by quantitative measures. Second, a representation of the alternatives must be capable of describing the phenomena of interest in quantitative terms. Third, comparisons will require analyses that explore a range of parameter values and use statistical methods to evaluate the results.

The first issue is obviously critical if the results are to be biologically significant; however, there is no prescription for discovering those biological behaviors that are based on natural selection or those that occur at random with high probability. The behaviors that are important characteristics of a given biological system can only be discovered by experimental means. Hypotheses must be generated and tested in each case, and this process will vary considerably according to the systems being studied. The behavioral repertoire of nonlinear systems can be quite diverse including saturation, thresholds, memory, time delays, synchrony, stable limit cycles and strange attractors.

The second issue is critical to any quantitative comparison of alternative systems. We require a mathematical language (or formalism) that is sufficiently flexible to represent the diverse behaviors that are likely to be encountered in the quantitative description of a nonlinear biological system. The power-law formalism (Savageau, 1996) is a most likely candidate for this language. It can be viewed as a canonical nonlinear representation from three different perspectives. From a fundamental perspective, it provides a generalization of mass-action kinetics, which is the most widely used representation of biological systems at the molecular level. From a recasting perspective, it provides a globally accurate representation that can be made mathematically equivalent to any sufficiently differentiable nonlinear system. From a local perspective, it provides a general representation that is guaranteed to be accurate over a range of variation about a nominal operating point.

*To whom correspondence should be addressed.
The third issue is critical because values for many of the parameters in any given complex system will not have been measured, and for those that have the estimates will often be poor. Moreover, even if we had a complete set of accurate parameter values with which to study the behavior of a system, the results would only apply to that particular system. In any case, we would have to vary the parameters over a range of values and statistically analyze the results to determine the properties of the general class of systems to which the particular system belongs.

Our purpose in this and the following paper is to present a methodology for dealing with this third issue and to illustrate its use in the simplest setting where the essentials of the methodology can be made most transparent. Hence, we shall focus on a class of systems for which the biologically relevant behavior is relatively simple and well defined (namely, unbranched amino-acid biosynthetic pathways with a single homeostatic steady state) and for which the local nonlinear representation, which is the simplest of the representations within the power-law formalism, is appropriate. At the end of the second paper we will return to these issues and indicate how the methods presented here might be applied to systems with more complex behaviors requiring more general representations within the power-law formalism. The methods themselves provide an extension of a previously developed approach for making well-controlled comparisons.

In the study of complex biological networks, models with alternative designs or structure are often compared to determine which of them provides the better representation for some observed phenomenon (e.g. Ni and Savageau, 1996). When comparing structurally different models for the same phenomenon, it is difficult to know whether the differences observed are accidental or inherent differences that can be attributed specifically to the alternative designs. The method of Mathematically Controlled Comparison (Savageau, 1972; for a review see Irvine, 1991) was proposed to address this issue.

In brief, the steps involved in this method are as follows. First, mathematical models are formulated for the alternative designs being compared. For example, a biosynthetic pathway with end-product inhibition and an identical one without it. One model, generally the more complex, is designated the reference; the other is designated the alternative. Second, the parameters of the alternative model are fixed relative to those of the reference model. Each process in the alternative model that is identical to one in the reference model is assigned a set of parameter values that is identical to the corresponding set in the reference model. This is referred to as internal equivalence. Each process in the alternative model that is different from the corresponding process in the reference model will have a set of parameter values that is unique to the alternative model, and these parameters represent degrees of freedom that must be constrained in an effort to reduce the accidental differences between the models. Each constraint is established by equating the expressions for a systemic property common to the two models. The set of constraint equations is then solved to determine values for the unique parameters of the alternative model in terms of values for the parameters of the reference model. This is referred to as external equivalence. Finally, having eliminated all the degrees of freedom, the two models are analyzed to determine the differences that remain.

The critical step in this method is the solution of the constraint equations. The models are described by nonlinear equations that in general have no analytical solution. However, the discovery of a canonical nonlinear representation that is locally valid and amenable to analytical solution (Savageau, 1969a, 1969b; for a review see Savageau, 1996) removes the difficulty associated with this critical step in many cases (Savageau, 1972, 1976). This canonical nonlinear representation within the power-law formalism is referred to as an S-system and it has the following systematic structure:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}, \quad i = 1, 2, \ldots, n.$$  

For each dependent concentration $X_i$ in a biochemical model there exists an aggregate production function and an aggregate consumption function. These aggregate functions are approximated by a first-order Taylor series in a logarithmic space, which in Cartesian space leads to the product of power-law functions. An exponent of zero for any $X_j$ means that that variable has no direct influence on the rate of the corresponding aggregate process, a positive exponent means that the variable and the rate of the aggregate process are positively correlated, and a negative exponent means that they are negatively correlated. In a steady state, (1) becomes a linear equation in logarithmic space and can be solved analytically. Likewise, various systemic properties can be calculated analytically and used to form constraints by equating the analytical expressions for corresponding systemic properties in the two models. These constraint equations can then be solved to determine values for the unique parameters of the alternative model in terms of values for the parameters of the reference model.

Once internal and external equivalence between the models is established in this manner, we can proceed to analyze the models and compare their systemic behaviors by taking ratios of their corresponding properties. The steady-state properties that are typically analyzed in Mathematically Controlled Comparisons include concentrations, fluxes, logarithmic gains, parameter sensitivities, and stability margins. For the purposes of this paper, these
systemic properties will be represented by $M$. The ratio of $M$ in the reference model to $M$ in the alternative model exhibits one of three possible properties.

1. The analytical ratio of $M$ values is always equal to 1, independent of parameter values. This means that the property being analyzed is always the same in the two models.

2. The analytical ratio of $M$ values is always larger (smaller) than 1, independent of parameter values. This means that the property being analyzed is always larger (smaller) in the reference model than in the alternative model. However, if the numerical values for the parameters are not known we can not say how much larger (smaller) the property is.

3. The analytical ratio of $M$ values is larger or smaller than 1, depending on the parameter values. In this case it is difficult to say anything about the property by simple examination of the analytical ratio.

The uncertainties associated with properties 2 and 3 will be addressed by the numerical methods being proposed in this paper. Moreover, these methods will allow us to draw statistical conclusions about the relative merits of various biological designs.

Methods
If we knew the numerical values for all the parameters of the reference model, then we could calculate the numerical values for all the parameters of the alternative model that is internally and externally equivalent. However, knowledge of all the parameter values is rarely available for any model. Furthermore, using just one set of parameter values restricts the interpretation to the specific pair of models being compared. These limitations can be overcome by creating a large ensemble of reference models with randomly generated sets of parameter values that adequately sample the parameter space. For each of these one can then construct the alternative model that is internally and externally equivalent.

There are two types of parameters that appear in the S-system representation (equation (1)): exponential parameters (kinetic orders) and multiplicative parameters (rate constants). The exponential parameters, which are weighted averages of more elementary kinetic orders, typically have values less than 4 in magnitude (Voit and Savageau, 1987). The multiplicative parameters, which reflect the different time scales present within the model, for most cases of interest are within 4 orders of magnitude of each other (i.e. within 4 log_{10} units). The results given in the following section are not critically dependent upon this particular choice of limits for the parameter space that needs to be sampled.

By using randomly generated numbers we can sample the relevant parameter space, apply selection and create a large ensemble of biologically relevant numerical models for both the reference and alternative designs, and make an ensemble of numerical comparisons. The amount of data generated by this approach can be overwhelming. The following subsections describe several ways to treat and interpret these data. In a following paper (Alves and Savageau, 2000) these methods are applied to a specific class of biochemical control mechanisms in a context different from that of mathematically controlled comparisons. Subsequent papers will provide examples of specific applications within the mathematically controlled comparison framework.

Basic treatment and analysis of the comparisons
The first problem in analyzing a large number of comparisons is deciding how to represent the data. Since we are comparing the value of a given property $M$ between the reference model and its alternative, one obvious way to represent the data is by taking the ratio of $M$ in the reference model to $M$ in the alternative model.

$$R = \frac{M_{\text{reference}}}{M_{\text{alternative}}}.$$  (2)

When dealing with an ensemble of comparisons we must calculate the ratio, $R$, of $M$ values for each reference model and its alternative model that is internally and externally equivalent. These data then can be treated by calculating some quantile of interest for the ensemble of ratios, thus determining whether $M$ is statistically larger in the reference models or their alternatives. This, however, will not give us much information, even if we included calculations for the dispersion of the results.

Density plots
More information can be obtained from density plots of $R$ versus $M$, where $M$ is a property measured in the reference model; e.g. the sensitivity, $S(X_t, \alpha_1)$, of a given intermediate, $X_t$, to fluctuations in the rate constant, $\alpha_1$, for the first reaction of the pathway. Some density plots where the ratio is typically smaller than 1 are presented in Figures 1–3. Note that in Figure 1A we have a situation in which the ratio of $S(X_t, \alpha_1)$ is uniformly scattered throughout the entire region bounded by $R = 1$ and $R = 0$. Figures 2A and 3A show different non-traditional distributions. Figure 3A shows a case in which $M$ can take only discrete values.

Density plots can be used to determine rank correlations between $M$ and $R$. Traditionally we calculate non-parametric rank correlations by using point measures such as the Spearman or Kendal rank correlation coefficients (e.g. Wherry, 1984; Krauth, 1988). These methods find linear and non-linear rank correlations between variables; however, it is not always easy to find such correlations in
Fig. 1. Uncorrelated Density of Ratios Plots. A: Density Plot of $R$ versus $M$ for two alternative models. There is a uniform distribution of values on both axes. B: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 50$. C: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 500$. See text for discussion.

Fig. 2. Correlated Density of Ratios Plots. A: Density Plot of $R$ versus $M$ for two alternative models. B: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 50$. C: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 500$. See text for discussion.
Fig. 3. Correlated Discrete Density of Ratios Plots. A: Density Plot of $R$ versus $M$ for two alternative models for which $M$ and $R$ assume discrete values. B: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 50$. C: Moving median plot of $\langle R \rangle$ versus $\langle M \rangle$ for the data in panel A and a window size of $W = 500$. See text for discussion.
Examples and discussion

Moving median plots of \( L_2 \) lists can be used to compare the relative effectiveness of two different classes of models on the basis of some criterion. For example, assume that \( M \) measures the sensitivity of a model to fluctuations in a given parameter and that this parameter sensitivity should be as low as possible according to the criterion of model robustness. The ratio \( R \) of \( M \) values in the reference model to \( M \) values in the alternative model, which is otherwise internally and externally equivalent to the reference model, is plotted and from this density plot one forms the moving median plot of \( \langle R \rangle \) versus \( \langle M \rangle \). Examples of such plots that exhibit various patterns are shown in Figures 2B and C. The value for \( \langle R \rangle \) is a function of \( \langle M \rangle \) with a minimum around \( \langle M \rangle \approx 1 \). With \( \langle M \rangle \approx 1 \), the value for \( M \) of the reference model is much less than that of the alternative model. With values for \( \langle M \rangle \) that are increasing or decreasing away from 1, the value for \( M \) of the reference model approaches that of the alternative model.

The selection of an appropriate window size is critical. If \( W \) is too small (e.g. 5), the \( Q_{0.5} \) plot will not differ significantly from the raw density plot. If \( W \) is too large, the correlation between \( \langle R \rangle \) and \( \langle M \rangle \) will be lost, or at least attenuated. This can be seen by comparing the curves for the two different window sizes in Figures 2B and C. As the window size increases from 50 to 500, the slope of the branch for \( \langle M \rangle \) less than 1 decreases (if the window size is increased further, the slope eventually becomes 0). This happens because the early samples of \( R \) are contaminated with latter samples and the correlation with the lower values of \( M \) is lost. With larger window sizes the slope of the branch for \( \langle M \rangle \) greater than 1 also decreases. As \( W \) approaches \( N \), the slope of the curve on either side of \( \langle M \rangle \approx 1 \) tends toward 0 and the \( Q_{0.5} \) plot provides no more information than calculating the median of the entire sample. Thus, the advantages of a \( Q_{0.5} \) plot only become apparent at intermediate window sizes. There is, to our knowledge, no good way of deciding the optimal size for the window \( W \); this depends on the sample size \( N \) and on the nature of the sample itself and must be determined by trial and error.

Figure 3A illustrates a case in which the values of \( M \) can only assume a finite number of discrete values. Figures 3B and C show the corresponding plots for \( \langle R \rangle \) versus \( \langle M \rangle \) of the reference model. A correlation between \( \langle R \rangle \) and \( \langle M \rangle \) is evident at low values of \( \langle M \rangle \) but disappears as \( \langle M \rangle \) increases. In addition, the \( Q_{0.5} \) plot in Figures 3B and C shows the dispersion in the moving median at each value of \( \langle M \rangle \), unlike the \( Q_{0.5} \) plots in Figures 1B and C and in Figures 2B and C. This dispersion occurs because there are several pairs in the list \( L_1 \) that have the same discrete value for \( M \) but different discrete values for \( R \). As the window \( W \) moves through a series of identical \( M \) values, the median value for \( M \) will remain unchanged whereas the median value for \( R \) will change. One can construct discrete density plots for any of the previous examples by designing classes for the values of \( M \) and by representing each class by the median of the class interval. Thus, the plot of \( \langle R \rangle \) versus \( \langle M \rangle \) in cases such as these can give us information not only about frequencies and correlations but also about dispersion of the results.

Acknowledgments

This work was supported in part by a joint PhD fellowship PRAXIS XXI/BD/9803/96 granted by PRAXIS XXI.
through Programa Gulbenkian de Doutoramentos em Biologia e Medicina (RA), the US Public Health Service Grant RO1-GM30054 from the National Institutes of Health (MAS), and the US Department of Defense Grant N00014-97-1-0564 from the Office of Naval Research (MAS). We thank Armindo Salvador, Rui Gardner Oliveira, and Charlotte Eriksson for critically reading early versions of this manuscript and making useful comments.

References