Toward Automated Cell Model Development through Information Theory†

A. Sayyed-Ahmad, K. Tuncay, and Peter J. Ortoleva*

Center for Cell and Virus Theory, Department of Chemistry, Indiana University, Bloomington, Indiana 47405

Received: February 28, 2003; In Final Form: June 30, 2003

The objective of this paper is to present a methodology for developing and calibrating models of complex reaction/transport systems. In particular, the complex network of biochemical reaction/transport processes and their spatial organization make the development of a predictive model of a living cell a grand challenge for the 21st century. However, advances in reaction/transport modeling and the exponentially growing databases of genomic, proteomic, metabolic, and bioelectric data make cell modeling feasible, if these two elements can be automatically integrated in an unbiased fashion. In this paper, we present a procedure to integrate data with a new cell model, Karyote, that accounts for many of the physical processes needed to attain the goal of predictive modeling. Our integration methodology is based on the use of information theory. The model is integrated with a variety of types and qualities of experimental data using an objective error assessment approach. Data that can be used in this approach include NMR, spectroscopy, microscopy, and electric potentiometry. The approach is demonstrated on the well-studied *Trypanosoma brucei* system. A major obstacle for the development of a predictive cell model is that the complexity of these systems makes it unlikely that any model presently available will soon be complete in terms of the set of processes accounted for. Thus, one is faced with the challenge of calibrating and running an incomplete model. We present a probability functional method that allows the integration of experimental data and soft information such as choice of error measure, a priori information, and physically motivated regularization to address the incompleteness challenge.

A. Introduction

There has been a long standing interest in obtaining a quantitative understanding of a living cell as a physicochemical system. Barriers to accomplishing this goal are the hierarchical complexity of a cell's spatial organization and the underlying network of reaction/transport processes, as well as the challenge of calibrating the many parameters that appear in such a model. The purpose of this paper is to present a methodology based on the integration of modeling and data through information theory that can be used in this or other complex reaction/transport systems modeling efforts. There are a number of reasons for developing a predictive cell model, including understanding the origin and nature of cellular life, drug design/treatment optimization, design of optimal microbes for environmental remediation and biochemical functions, stem cell research, and predicting the emergence of drug-resistant bacteria and identifying potential drug targets.

The complexity of the cellular reaction/transport network and hierarchical internal spatial organization have been put forth as arguments against the feasibility, or indeed the desirability, for developing a physicochemical model. However, considering the potential benefits of such a model, we contend that the opposite is the case. It is the immensity of this complex, hierarchical dynamical system that makes a model and computational simulator a necessity if the benefits of predictability are ever to be obtained. For any procedure to be successful, it must overcome the complexity of the system with the immensity of the data set used. However, to be practical, the procedure must be automatable. The approach outlined here has these features. Our overall approach is suggested Figure 1. In Figure 1a, we suggest that a variety of data types are to be integrated through information theory to develop and calibrate a cell model. In Figure 1b, we present a schematic flowchart for a computation that minimizes the error in model-predicted versus observed data to yield an optimal set of model parameters.

A comprehensive mathematical model has a large number of input parameters. For a cell, for example, there is the metabolic network reaction rate and equilibrium coefficients, membrane permeability, etc. It is difficult to get some of these parameters experimentally. To use the variety of data types for calibrating purposes as suggested in Figure 1a, a cell model must predict values of these observables. Clearly, a genomic, proteomic, metabolic, bioelectrical model is therefore required. Recently available models are Karyote,[1] GEPASI 3,[2−8] Mcell,[9,10] E-Cell,[11,12] and Virtual Cell.[13−18] Karyote is a multicompartmentalized, multiple time scale cell simulator that it is ideally suited for the implementation of our model/data integration strategy. A brief description of Karyote is provided in section F. To illustrate the concept, the use of multidimensional spectroscopy (MDS) data is suggested in Figure 1b wherein a cell model predicted protein population can be used with rules of tryptic digestion and known properties of the digest fragments to develop a synthetic spectrum. This synthetic spectrum information is compared with an observed spectrum and an error measure is calculated. Thus, to use protein spectra for calibration, a real model must predict the dynamic proteome. This involves accounting for transcription and translation biopolymerization, as well as amino acid and nucleotide synthesis. In addition, the kinetics of posttranslational modification must be accounted for in such a way that enzyme and ribosome creation/destruction are predicted by Karyote. The procedure is used to integrate the model with a variety of types and qualities of data as suggested in Figure 1a. The fundamental quantity on which the formulation is based is the probability that the model is correct and accurate once a set of assumptions are made. This probability will be, in a sense, subjective in that while the model

† Part of the special issue "Charles S. Parmenter Festschrift".

10.1021/jp0302921 CCC: $25.00
Published on Web 09/13/2003

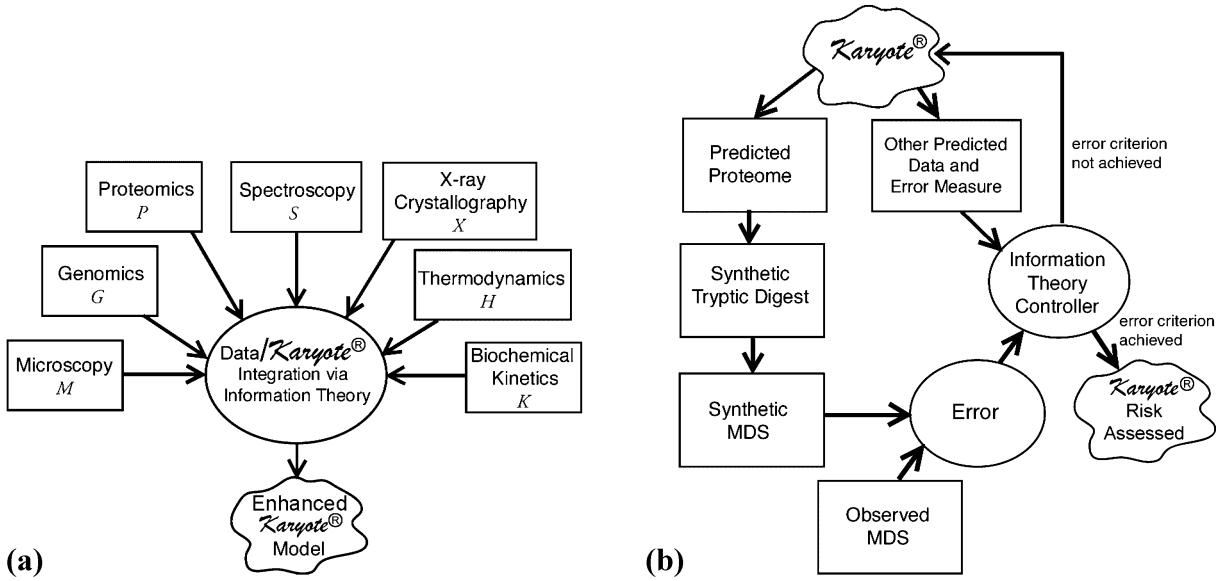

Figure 1. (a) Multiple types and qualities of data are integrated to automatically yield improvements in Karyote as new data becomes available. (b) Integration of Karyote with a variety of data is used to compute the most probable values of the least well-constrained model parameters via our information theory method. The method also yields the most probable time course of the concentrations of key chemical species of which the mechanism of production or degradation are not known. The computation involves execution of a number of Karyote simulations that increase linearly with the number of parameters to be calibrated and that can be run in parallel. The case for multidimensional spectroscopy is illustrated.

may have wider applicability; the uncertainty measure is established only for the class of phenomena of interest. This viewpoint allows us to select measures of correctness of the model that reflect our interest. For example, we might say that accuracy would imply that errors in predicting relative intracellular concentrations of a wide range of orders of magnitudes (e.g., nanomolar to molar) must be similarly minimized. Our probabilistic formulation allows us to estimate uncertainties in all model predictions, and, as we shall show, allows us to calibrate an incomplete model.

Historically, most calibration problems are formulated as, $\mathbf{Ax} = \mathbf{y}$, where $\mathbf{y}$ is a vector of observed quantities, and $\mathbf{x}$ is the vector of unknown model parameters. For a nonlinear model, the matrix $\mathbf{A}$ usually depends on $\mathbf{x}$. Because the problem is usually ill-posed, $\mathbf{A}$ is ill-conditioned.[19] The error $E = ||\mathbf{Ax} - \mathbf{y}||^2$ is a quadratic measure to be minimized with respect to $\mathbf{x}$. A number of techniques have been used to regularize such systems. Tikhonov's approach is a commonly used technique in which a small regularization parameter q is introduced to modify E to $||\mathbf{Ax} - \mathbf{y}||^2 + q||\mathbf{x}||^2$. However, the selection of q significantly affects the inversion. This technique is equivalent to the minimization of E subjected to the constraint $||\mathbf{x}||^2 = f$ through the use of the Lagrange multipliers. Minimization of the modified error damps the large oscillations in the least-squares solution. The Levenberg−Marquardt technique uses a full Newton approach and introduces another regularization parameter to the diagonal of the Jacobian matrix.[20,21] Once again, the choice of the regularization parameter is difficult, and the usual practice is to change it as the simulation progresses to minimize its effect.[22,23] In practice, multiple regularization techniques can be employed simultaneously.[19] Applications of the regularization techniques are presented in Player et al.,[24] Rao et al.,[23] Kytomaa and Weselake,[22] Torres et al.,[25] and Mendes and Kell.[26] Mendes and Kell compiled a review of optimization techniques applied to biological systems. They used the metabolic simulator GEPASI along with a large number of optimization techniques (such as steepest descent, truncated Newton, and genetic algorithm) to estimate five rate constants of the mechanism of irreversible inhibition of HIV proteinase. In this paper, we propose several improvements including use of information theory to construct a probability density function that can be used to assess the uncertainty in calibrated parameters and predicted cell behavior, use of different error measures to improve the optimization technique, development of physically motivated regularization techniques for problems in which the least-well-known parameters are functions of space or time (as this is equivalent to providing new information, in a novel approach, we impose regularization constraints on the probability density functions), and use of a consistent approach to weigh the importance of different error measures.

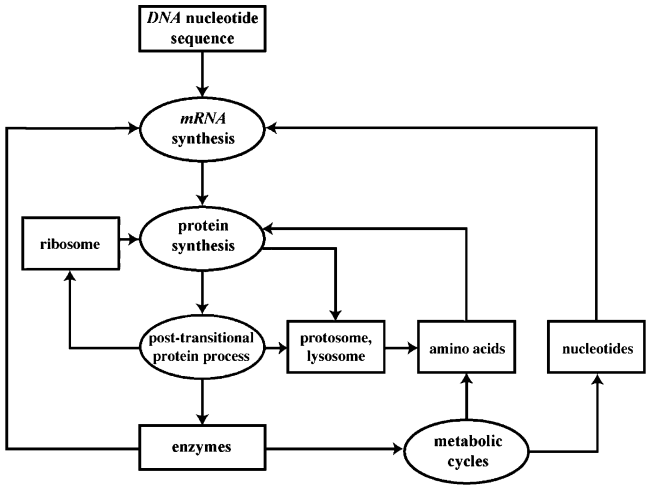

Figure 2. Partial schematic Karyote flowchart showing how DNA nucleotide sequence data is used in a self-consistent way to generate cell reaction/transport dynamics by feedback control and coupling of metabolic, proteomic, and genomic biochemistry.

In this paper, we present results on the implementation of our algorithm based on Karyote cell model.[1,27,28] A partial Karyote flowchart is seen in Figure 2. Because an extensive set of processes is accounted for in Karyote, it is the type of model that is ideally suited for the present approach. These technical challenges, presentation of our formulation, and application to the modeling of *Trypanosoma brucei* are presented herein.

B. Information Theory Model/Data Integration

Our information theory formalism is based on the construction of the probability density for poorly constrained factors. The three types of factors we account for are as follows: type *A*, discrete parameters (e.g., stoichiometric coefficients specifying the numbers of each molecular species participating in a reaction or parameters determining protein sequence/function rules); type *B*, continuous parameters (e.g., reaction rate coefficients, membrane transport parameters, and equilibrium constants that can reside in a continuous range); type *C*, functions (e.g., the time course of the concentration of chemical species of which the enzymatic role is known but the mechanism of creation/destruction is not known).

To estimate the most probable values of factors *A* and *B* and the time course of factor *C*, we introduce a method that surmounts the limitations of regularization techniques used in other approaches. First, we introduce the probability $\rho(\Gamma)$, ($\Gamma = A$, *B*, or *C*). The entropy *S* of information theory[29] is a measure of the overall uncertainty that we have about the value of Γ; it is defined in our formulations[28,30] via

$$S = -\int_{\Gamma} \rho \ln \rho \tag{B.1}$$

In this expression, $\int$ implies a sum over the discrete variables *A*, an integration over the continuous parameters *B*, and a functional integration over *C*. Normalization of the probability $\rho[\Gamma]$ implies

$$\int_{\Gamma} \rho = 1 \tag{B.2}$$

Experiments are divided into N_e groups labeled $k = 1, 2, ..., N_e$ for each of which there is a set of observed data values, $O^{(k)}$. For example, $O^{(1)}$ could be the time course of a set of intracellular constituent concentrations as they change in response to an injected chemical disturbance, $O^{(2)}$ can be the normal proteome, $O^{(3)}$ can be the proteome of the virally infected cell, and $O^{(4)}$ can be a set of membrane potentials in a rest state or as they change in response to an electrode-imposed disturbance. Through Karyote, we compute the model predictions, $\Omega^{(k)}(\Gamma)$, that correspond to $O^{(k)}$. Typically, these $\Omega^{(k)}$ are indirectly related to Γ. Because Karyote predictions depend on the choice of the parameters Γ, so does $\Omega^{(k)}$. The choice of the error measures is discussed in the next section. In general, the error measure $E^{(k)}$ should vanish as the difference between the predictions and the observation goes to zero.

$$E^{(k)} \to 0 \quad \text{as} \quad \Omega^{(k)} \to O^{(k)} \tag{B.3}$$

The entropy is proposed as a measure of our uncertainty in the state of the system. Thus, for discrete parameters (type *A*) if Γ is known to be a particular value Γ_0, then $\rho = 1$ for $\Gamma = \Gamma_0$ and 0 otherwise, implying that $S = 0$. If all values of Γ are equally likely, then *S* takes on its largest value. Hence, to be "objective", ρ should be determined as the probability that maximizes *S* constrained only with the available information. Thus, we maximize *S* subject to normalization (eq B.2) and the estimated error in the available data. Among the latter are the error conditions

$$\int_{\Gamma} \rho E^{(k)} = E^{(k)*} \tag{B.4}$$

Here $E^{(k)*}$ is the average value of $E^{(k)}$, and it is based on estimated experimental errors in the data and in the mathematical and numerical model.

From the physics of a system and from our general experience, we often know that time-dependent variables change smoothly on a time interval smaller than some characteristic time. Because data is often sparse, it is necessary to apply homogenization constraints on the time dependence of the continuous variables, $C(t)$. For example, assume that estimates based on known reactions suggest that $C(t)$ varies on a second time scale or longer not, for example, on a nanosecond scale. Then we impose a constraint on the expected rate of change of $C(t)$:

$$\int_{\Gamma} \rho \int_0^{t_f} dt \, \frac{1}{2}\left(\frac{\partial C_j}{\partial t}\right)^2 = t_f X_j \quad j = 1, 2, ..., N_t. \tag{B.5}$$

for the *j*th time-dependent parameter, C_j; the value of X_j represents the value of the square of the rate of change of C_j averaged over the ensemble and the total time t_f of the experiment. N_t is the number of time-dependent functions to be estimated. In general, one might need to apply regularization on the space dependence of some variables, $D(\vec{r})$. Such constraints eliminate unphysical high-frequency content of the solution. (e.g., spatial regularization of diffusion coefficient that is known only at few spatial points). The constraint can be expressed as

$$\int_{\Gamma} \rho \int_{\Omega} d^3r \, \frac{1}{2}|\nabla \vec{D}_i|^2 = \Psi_i \quad i = 1, 2, ..., N_s. \tag{B.6}$$

Here N_s is the number of space-dependent functions to be estimated. Introducing Lagrange multipliers β_k, Λ_j, and Π_i, we find that the ρ that maximizes *S* subject to (eqs B.2, B.4, B.5, and B.6) takes the form

$$\ln \rho = -\ln Q - \frac{1}{2}\sum_{i=1}^{N_s} \Pi_i \int_{\Omega} d^3r \, |\nabla \vec{D}_i|^2 - \frac{1}{2}\sum_{j=1}^{N_t} \Lambda_j \int_0^{t_f} dt (\partial C_j/\partial t)^2 - \sum_{k=1}^{N_e} \beta_k E^{(k)}[\Gamma] \tag{B.7}$$

The normalization coefficient *Q* is given by

$$Q = \int_{\Gamma} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_s} \Pi_i \int_{\Omega} d^3r \left|\nabla \vec{D}_i\right|^2 - \frac{1}{2}\sum_{j=1}^{N_t} \Lambda_j \int_0^{t_f} dt (\partial C_j/\partial t)^2 - \sum_{k=1}^{N_e} \beta_k E^{(k)}[\Gamma] \right) \tag{B.8}$$

By finding Lagrange multipliers (β_k, Λ_j, and Π_i), we construct the most unbiased probability distribution of the model input parameters Γ.

C. Data Types and Error Measures

The error measures $E^{(k)}$ of section B are a central element of our information theory approach. It is our opinion that the choice of error measure itself can be viewed as a type of information that can be justifiably folded into the approach. For example, suppose that from experience it is known that one error measure is more sensitive to the calibration of a given parameter than others. Alternatively, one error measure may emphasize one subset of data (e.g., large values) versus another (e.g., small

ones). Thus, a judiciously chosen error measure can reflect our knowledge of what is important in assessing the accuracy of the model.

Cells often can be induced (e.g., by manipulating the culture medium) to remain in a steady state of substrate uptake and product expulsion. Let the predicted steady-state concentrations be denoted by $\{c_i^p; i = 1, 2, ..., N\}$ for an N-component system, while the observed values, c_i^o, of these quantities are assumed to be known. Let $h(x,y)$ be a positive function of x and y. Then one class of error measure is

$$E = \sum_i h(c_i^p, c_i^o) \tag{C.1}$$

The challenge is to choose the form of $h(x,y)$ that fits certain criteria that one may have on model accuracy. In our formulation, these criteria are as follows: h is zero when $x = y$ and must be positive otherwise; for fixed y, h must be a monotonically increasing function of x as $|x - y|$ increases; h should reflect any valuation one may have (e.g., all values are equally important or the larger values are most important).

Specific examples of error measures for concentrations are as follows:

$$E = \sum_i |c_i^p - c_i^o|^{1/2} \tag{C.2}$$

$$E = \sum_i (\ln c_i^p - \ln c_i^o)^2 \tag{C.3}$$

$$E = \sum_i \left(\frac{c_i^p - c_i^o}{c_i^o} \right)^2 \tag{C.4}$$

$$E = \sum_i \left(\frac{c_i^p - c_i^o}{\sigma_i^o} \right)^2 \tag{C.5}$$

where σ_i^o is a weighting factor (e.g., the uncertainty in the observed data).

The above error measures have different characteristics. In our studies steady-state concentrations range over several orders of magnitude that the choice of $h(x,y) = [\ln x - \ln y]^2$ is a good measure to make use of all measured data. Thus while x and y can vary over several orders of magnitudes, this measure treats all species on a relatively equal footing. Our study indicates that the simple measure $(x - y)^2$ has rather poor behavior (see Figure 3a). The steady-state concentrations were predicted by Karyote for *T. brucei* cell.[27] When the simple quadratic error measure was used, error was weakly dependent on the rate coefficient, while the use of eq C.3 resulted in a well-defined minimum. In this sense, knowledge of the optimal error measure is a type of information.

D. A Priori Information

Considering the number of model parameters that are poorly constrained, the above procedure may not be sufficient to determine them when only a small amount of data is available. Thus, we suggest approaches that will guarantee solubility based on qualitative knowledge. First, consider a set of expected errors $F^{(k)}$, $k = 1, 2, ..., N_x$. Associated with one of the N_x expected errors is a group of $N_a^{(k)}$ model parameters, $\gamma^{(k)}$ $\{\gamma_1^{(k)}, \gamma_2^{(k)}, ..., \gamma_{N_a^{(k)}}^{(k)}\}$, for which we have experience. For example, if they are rate coefficients that we have estimated from experience

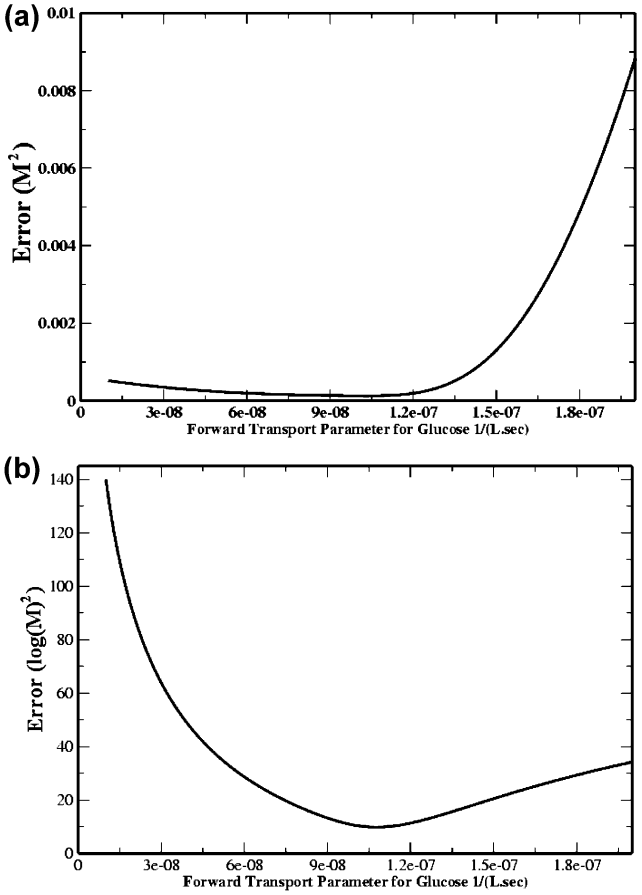

Figure 3. Different error measures show different response as a function of forward transport coefficient for glucose between a glycosome and the cytosol in *T. brucei* model using Karyote: (a) the simple error measure (eq C.5) with $\sigma_i^o = 1$ is biased to large concentrations; (b) the log-difference error (eq C.3) more equally weighs the full range of concentration values. It shows a distinct global minimum.

with similar reactions to be $\gamma^{(k)}$, then let $F^{(k)}$ be given by

$$F^{(k)} = \sum_{i=1}^{N_a^{(k)}} h(\gamma_i^{(k)}, \bar{\gamma}_i^{(k)}) \tag{D.1}$$

From a database of similar reactions, one may gather statistical information, denoting the expected value of $F^{(k)}$ by $F^{(k)}*$. We impose the conditions

$$\int_\Gamma \rho F^{(k)} = F^{(k)}* \tag{D.2}$$

This allows one to determine the N_x Lagrange multipliers that are introduced in the entropy maximization. To illustrate the essence of this approach, consider the following problem with one type of error. The maximization of entropy subject to normalization (eq B.2) and error constraint (eqs B.4 and D.2) yields

$$\ln \rho = -\ln Q - \beta E - \sum_{i=1}^{N_x} \lambda_i F^{(i)} \tag{D.3}$$

where β and λ_i are Lagrange multipliers. If F is taken as a quadratic function, it implies a Gaussian envelope that helps to stabilize the numerical solution.

Next, we introduce an "irrelevance" constraint. In this case, we consider certain parameter values to be irrelevant once they reach an asymptotic range. For example, consider the reaction

$$X + Y \rightarrow Z \tag{D.4}$$

of rate coefficient q and rate law $q(KXY - Z)$. As $q \rightarrow \infty$, then to lowest order $KXY = Z$ and the rate becomes independent of q. To show this, expand X, Y, and Z in a Taylor series in q^{-1}. After a short transition period, the system evolves to the equilibrium manifold $QX_0Y_0 = Z_0$, where $X = X_0 + q^{-1}X_1, + ...$ With this, an error measure will become independent of q beyond a crossover value, and hence, ρ becomes independent of q in that asymptotic range. Let W_l^f and W_l^r be the forward and reverse rates of the lth reaction. Then as the associated rate coefficient q_l exceeds the crossover, $W_l^f \approx W_l^r$. Hence, the quantity

$$\xi_l = \frac{W_l^f - W_l^r}{W_l^f + W_l^r} \tag{D.5}$$

provides a measure of proximity to the asymptotic limit. If there are N_{cross} of these q_l ($l = 1, 2, ..., N_{cross}$) then consider the measure

$$G = \sum_{l=1}^{N_{cross}} \left\{ \frac{1}{\xi_l^2} + \left(\frac{\bar{q}_l}{q_l} \right)^2 \right\} \tag{D.6}$$

The explicit $\bar{q}_l/q_l$ term accounts for the fact that the rate coefficients are not expected to fall too far below typical values, $\bar{q}_l$, and certainly not below zero. Then we impose the constant

$$\int \rho G = N_{cross} g \tag{D.7}$$

where g is a value of a typical ξ_l^{-2} beyond which crossover is expected (say 10^{-3}).

E. Numerical Approach

The key to the implementation of our information theory method is an efficient numerical algorithm for calibrating the continuous parameters B and the discretized functions C, which together constitute a set of N_p parameters denoted ($\mathbf{x} = x_1, x_2, ..., x_{N_p}$).

1. Single Data Set. For one error type (and associated data set), the results of section B imply

$$\ln \rho = -\ln Q - \beta E(\mathbf{x}) \tag{E.1}$$

The most probable parameter values are at ρ_m, the global maximum of ρ. This occurs at E_m, the global minimum of E, that is, at the solution of

$$\frac{\partial E}{\partial x_i} = 0, \qquad i = 1, ..., N_p \tag{E.2}$$

For metabolic kinetic networks, as an illustrative example, steady-state concentration measurements are used to construct an error measure. Steady-state probing is crucial to determine the error response of the model with respect to steady-state concentrations. The simulation time is essential to determine the most probable values of the parameters (see Figure 4).

By expanding E around the most probable value $\mathbf{x}_m$ of $\mathbf{x}$ and dropping cubic and higher order terms in the deviation from $\mathbf{x}_m$, we get (see Appendix I)

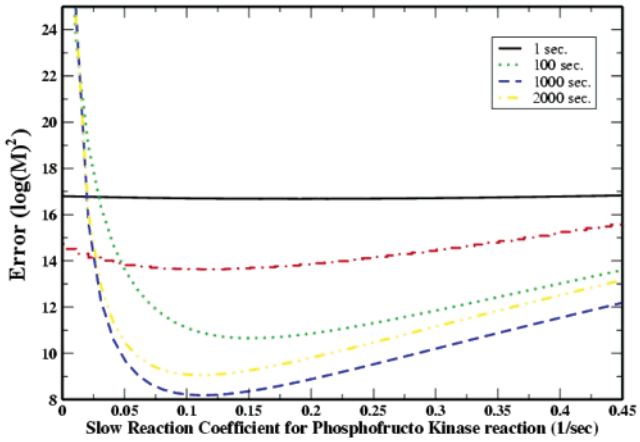

Figure 4. The error topography of the error measure depends on simulation time. The longer the simulation time is, the deeper are the minima that we get for that specific parameter, as expected because experiments were done at steady-state concentrations.

$$\frac{1}{Q} \approx \rho_m = \prod_{i=1}^{N_p} \sqrt{\frac{N_p \lambda_i}{4\pi(E^* - E_m)}} \tag{E.3}$$

and

$$\beta = \frac{N_p}{2(E^* - E_m)} \tag{E.4}$$

hence

$$\rho(\mathbf{x}) \approx \rho_m \exp\{-(^1/_2)\beta \Delta \mathbf{x}^T \mathbf{H}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x}\} \tag{E.5}$$

The λ_i are the eigenvalues of $H_{ij}(\partial^2 E/(\partial x_i \partial x_j))$, the Hessian matrix of E calculated at the minimum E_m of E. E^* is the expected value of the error evaluated from experimental and model uncertainties. Caution must be used for adopting quadratic approximation (see Figure 5). Expected values can be found using the probability distribution and the Hessian matrix of the output predictions $P(\mathbf{x})$ calculated at the most probable value of the parameters.

$$\bar{P} = P(\mathbf{x}_m) + \frac{(E^* - E_m)}{N_p} \sum_{i=1}^{N_p} \frac{\gamma_{ii}}{\lambda_i} \tag{E.6}$$

Also one can calculate the uncertainty in the predictions using the output response vector and the Hessian matrix of the model prediction $P(\mathbf{x})$ calculated at the most probable value of the parameters

$$\sigma_p^2 = \frac{2(E^* - E_m)}{N_p} \sum_{i=1}^{N_p} \frac{\kappa_{ii}}{\lambda_i} + \frac{3(E^* - E_m)^2}{N_p^2} \sum_{i,j=1}^{N_p} \frac{\gamma_{ij}^2}{\lambda_i \lambda_j} - \frac{(E^* - E_m)^2}{N_p^2} \left(\sum_{i=1}^{N_p} \frac{\gamma_{ii}}{\lambda_i} \right)^2 \tag{E.7}$$

Here γ_{ii} are the diagonal elements of the transformed Hessian matrix of P calculated at the minimum of E. $\kappa_{ii} = (\partial P/\partial x_i)^2$ are the diagonal elements of the tensor product of the prediction output response vector, $\partial P/\partial x_i$, and its transpose calculated at the minimum of E.

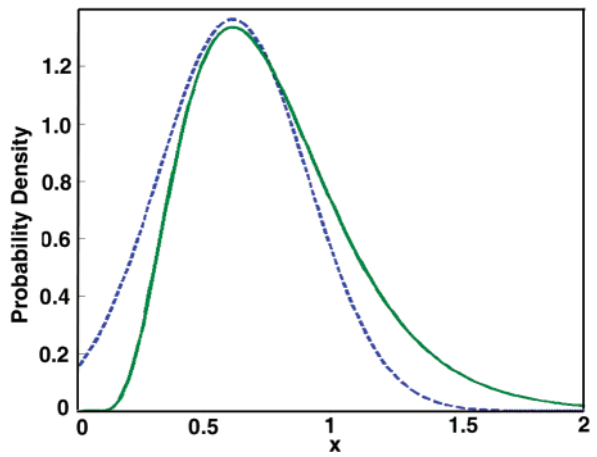

Figure 5. The dashed curve is a Gaussian approximation to the probability distribution constructed using a quadratic expansion of ln ρ around the most probable value. While agreement is good in this case, the Gaussian approximation is not always appropriate.

2. Multiple Data Sets. Here, we maximize the entropy to find the probability of the model parameters $S[\rho]$ with respect to $\rho(\mathbf{x})$ and subject to normalization and eq B.4. Introducing Lagrange multipliers, $-\ln Q$, and β_k, we maximize the auxiliary function to obtain

$$\ln \rho(\mathbf{x}) = -\ln Q - \sum_{k=1}^{N_e} \beta_{(k)} E^{(k)}(\mathbf{x}) \tag{E.8}$$

As in the single error measure case, we find the most probable value of $\mathbf{x}$ by minimizing the total error of E^T

$$E^T = \sum_{k=1}^{N_e} \beta_{(k)} E^{(k)}(\mathbf{x}) \tag{E.9}$$

subject to normalization and error constraints (eq B.4). We thus solve $N_p + N_e + 1$ equations

$$\frac{\partial E^T}{\partial x_i} = \sum_{k=1}^{M} \beta_{(k)} \frac{\partial E^{(k)}}{\partial x_i} = 0, \quad i = 1, ..., N_p \tag{E.10}$$

$$\int E^{(k)}(\mathbf{x}) \rho(\mathbf{x},\boldsymbol{\beta})\, d\mathbf{x} - E^{(k)*} = 0, \quad k = 1, 2, ..., N_e \tag{E.11}$$

and

$$\int \rho(\mathbf{x},\boldsymbol{\beta})\, dx - 1 = 0 \tag{E.12}$$

Using a Metropolis−Monte Carlo algorithm to evaluate eqs E.11 and E.12, one needs thousands of model runs to evaluate each integral. However, with the use of a quadratic truncation of E, the computational time is reduced dramatically. This allows us to compute these integrals analytically as a function of β. By doing so, we get (see Appendix II)

$$\frac{1}{Q} \approx \rho_r(\mathbf{x}_r,\boldsymbol{\beta}) = \prod_{i=1}^{N} \sqrt{\frac{\lambda_i(\mathbf{x}_r,\boldsymbol{\beta})}{2\pi}} \exp\left(-\frac{\mathbf{X}_i^2(\mathbf{x}_r,\boldsymbol{\beta})}{2\lambda_i(\mathbf{x}_r,\boldsymbol{\beta})}\right) \tag{E.13}$$

and

$$E^{(k)}(\mathbf{x}_r,\boldsymbol{\beta}) + \frac{1}{2}\sum_{j=1}^{N_p}\left(\sum_{k=1}^{N_p}\theta_{jk}\right)\frac{(\lambda_j(\mathbf{x}_r,\boldsymbol{\beta}) + \mathbf{X}_j^2(\mathbf{x}_r,\boldsymbol{\beta}))}{\lambda_j^2(\mathbf{x}_r,\boldsymbol{\beta})} = E^{(k)*} \tag{E.14}$$

where $\mathbf{X}_r = \mathbf{G}_r \nabla E|_{\mathbf{x}=\mathbf{x}_r}$, $\boldsymbol{\theta}_r = \mathbf{G}_r^T J_r^{(k)} \mathbf{G}_r$, $\mathbf{G}_r$ are the eigenvectors of the total error (eq E.9) Hessian matrix evaluated at $\mathbf{x} = \mathbf{x}_r$, and $J_r^{(k)}$ is the Hessian matrix of the kth error type.

Denote eqs E.10 and E.11 by $\{f_i, i = 1, 2, ..., N_p, N_p + 1, ..., N_p + N_e\}$ and β by $x_{N_p+1}, ..., x_{N_p+N_e}$. The first N_p equations follow from error minimization, while the remainder follow from the error constraints. Note that we do not need to solve for Q because we impose normalization on eq E.12 to get an approximate normalization constant. The above system of nonlinear equations can be written as

$$\underline{\mathbf{f}}(\mathbf{x}) = 0 \tag{E.15}$$

We solve eq E.15 using the Newton−Raphson method starting with an initial guess and constructing the Jacobian matrix, which represents the sensitivity of the equations to changes in the variables. The evaluation of the Jacobian matrix and the error minimization equations cannot be obtained analytically for the complex reaction/transport systems of interest here. While automated differentiated methods (ADIFOR) can be used to develop accurate expressions, such an approach is memory intensive and not easily parallelized. However, a forward difference scheme is found to be easily coded and parallelized. We have implemented a finite-parameter perturbation method to calculate the error response and the Jacobian (for one error type) as follows:

$$\frac{\partial E}{\partial x_i} \approx \frac{E(\mathbf{x} + h_i\hat{\mathbf{e}}_i) - E(\mathbf{x})}{h_i} \tag{E.16}$$

and

$$\frac{\partial^2 E}{\partial x_i \partial x_j} \approx \frac{E(\mathbf{x} + h_i\hat{\mathbf{e}}_i + h_j\hat{\mathbf{e}}_j) - E(\mathbf{x} + h_i\hat{\mathbf{e}}_i) - E(\mathbf{x} + h_j\hat{\mathbf{e}}_j) + E(\mathbf{x})}{h_i h_j}. \tag{E.17}$$

in which $\hat{e}_i$ is a unit vector in the ith direction and h_i is a small perturbation. For multiple error types, the above will be the upper right part of the Jacobian; the rest can be calculated in the same way by perturbing the Lagrange multipliers. The number of model runs needed to calculate the Jacobian is $(N_p + 1) + N_p(N_p + 1)/2$. This can be reduced to $(N_p + 1)$ if one uses Newton−Gauss or a steepest descent gradient scheme. One can use a hybrid method (e.g., starting with a steepest descent scheme and after a few iterations applying Newton−Raphson or Newton−Gauss (to get a quadratic convergence). Once the Jacobian is constructed, we solve

$$\underline{\underline{J}}\Delta\underline{\mathbf{x}}_n = -\underline{\mathbf{f}} \tag{E.18}$$

and

$$\underline{\mathbf{x}}_{n+1} = \underline{\mathbf{x}}_n + \omega\Delta\underline{\mathbf{x}}_n \tag{E.19}$$

here $J_{ij} = \partial f_i/\partial x_j$ for Newton−Raphson method; J for Newton−Gauss is obtained by dropping of predictions' second derivatives terms in the Newton−Raphson Jacobian. ω is obtained using a line search along the direction of Newton−Raphson or Newton−Gauss direction. We update the solution by iteratively applying this procedure until the elements of f become smaller than a prescribed tolerance. For insufficient or poor quality data or a poor choice of error measure, the scheme will not converge as expected. In summary, this iterative procedure is much more

efficient than Monte Carlo methods because one solves for the most probable value of the model parameters directly. The algorithm is parallelizable because the many simulations of the model that are required can be run simultaneously; the number of simulations required scales linearly with the number of parameters to be determined for a Newton−Gauss optimization technique.

3. Time Regularization Constraints. We now give more details on our strategy for probability functional calculations. We return to problems involving the three types of factors to be calibrated $\Gamma = (A,B,C)$ as in section B. Introducing Lagrange multipliers β_k and Λ_j, we find that the ρ that maximizes S subject to constraints (eqs B.2, B.4, and B.5) takes the form

$$\ln \rho = -\ln Q - \frac{1}{2}\sum_{j=1}^{N_t}\Lambda_j\int_0^{t_f} dt\,(\partial C_j/\partial t)^2 - \sum_{k=1}^{N_e}\beta_k E^{(k)}(\Gamma) \quad \text{(E.20)}$$

The factor Q is a constant to be determined by imposing normalization (eq B.2). The most probable value of Γ is that which maximizes ρ. For type A parameters, this follows from a discrete search; for $B = (B_1, B_2, ..., B_{N_b})$ and $C = (C_1, C_2, ..., C_{N_t})$, one must solve

$$\sum_{k=1}^{N_e}\beta_k\frac{\partial E^{(k)}}{\partial B_j} = 0, \quad j = 1, 2, ..., N_b \quad \text{(E.21)}$$

and

$$\Lambda_j\frac{\partial^2 C_j}{\partial t^2} + \sum_{k=1}^{N_e}\beta_k\frac{\delta E^{(k)}}{\delta C_j} = 0, \quad j = 1, 2, ..., N_t \quad \text{(E.22)}$$

This is a functional differential equation that has similarities in its behavior to a steady-state diffusion equation in the time dimension t. The functional derivatives $\delta E^{(k)}/\delta C_j$ measure the degree to which $E^{(k)}$ changes when the form of the function $C_j(t)$ changes by an infinitesimal amount. As the Λ-parameters get larger, C becomes a smoother function of time. The values of the β and Λ parameters are determined in our procedure via imposition of the conditions eqs B.4 and B.5. We solve the above equations for the most probable values of A, B, and C numerically. We use a finite difference scheme so that the homogenization constraints take the form

$$\frac{1}{2}\int_0^{t_f} dt\,(\partial C_j/\partial t)^2 = \frac{1}{2}\sum_i\left(\frac{C_{j,i+1} - C_{j,i}}{\Delta t}\right)^2\Delta t \quad \text{(E.23)}$$

Equation E.22 can be written as

$$\Delta t\Lambda_j\frac{C_{j,i+1} - 2C_{j,i} + C_{j,i-1}}{\Delta t^2} + \sum_{k=1}^{N_e}\beta_k\frac{\partial E^{(k)}}{\partial C_{j,i}} = 0,$$
$$j = 1, 2, ..., N_t \qquad i = 1, 2, ..., N_j \quad \text{(E.24)}$$

where Δt is the spacing between the discretized values of C_j and N_j is the number of discretization intervals for species j. A Newton−Raphson method is used for solving the coupled equations (E.21, E.24).

A simple reaction model illustrates this approach. The model involves three species, X, Y, and C, which are taken to participate in the reactions

$$\begin{aligned} &X + Y \rightarrow 2X \\ &2X \rightarrow \text{products} \\ &2Y \rightarrow \text{products} \\ &C + X \rightarrow \text{products} \\ &C + Y \rightarrow 2Y \end{aligned} \quad \text{(E.25)}$$

While all the reactions affecting X and Y are assumed to be known, those affecting the catalyst C are not. Consider now the challenge of determining the time course of the catalyst concentration, $C(t)$, given limited or noisy data on $X(t)$ at a set of discrete times, given that C is known at $t = 0$ and the final time t_f (5 min). We assumed

$$C(t) = e^{-|\sin(\omega t)|} \quad \text{(E.26)}$$

and then generated $X(t)$ via the numerical solution of the mass action rate laws for the mechanism (eq E.25); this was taken as the observed data, and various levels of noise were added to evaluate the effect of uncertainty in the data. Figure 6 shows a comparison of results for various levels of noise in the experimental data. Even when there is a large amount of data, solution without regularization is vulnerable to noise in the experimental data. The physically motivated homogenization

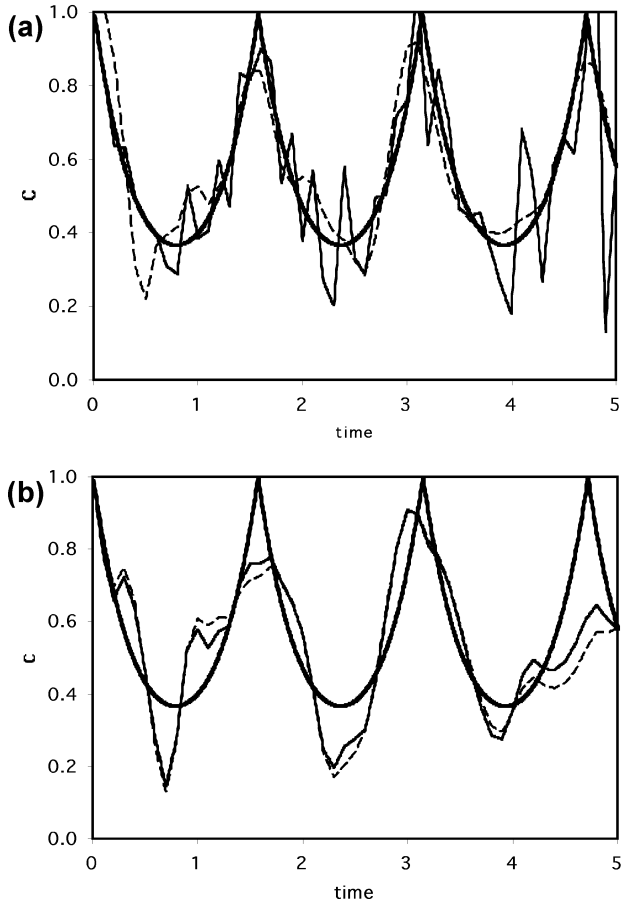

Figure 6. (a) Comparison of the time course of $C(t)$ as known (bold) and predicted with and without regularization (dashed and solid line). To the 41 data points used, a random error of 0.3% was added to determine the effect of experimental data uncertainty on the evaluation of $C(t)$. In the absence of regularization, the high-frequency oscillations are unacceptably large. (b) Even when the level of noise is increased significantly (2% and 3% for thin solid and dashed lines, respectively), we obtain satisfactory results.

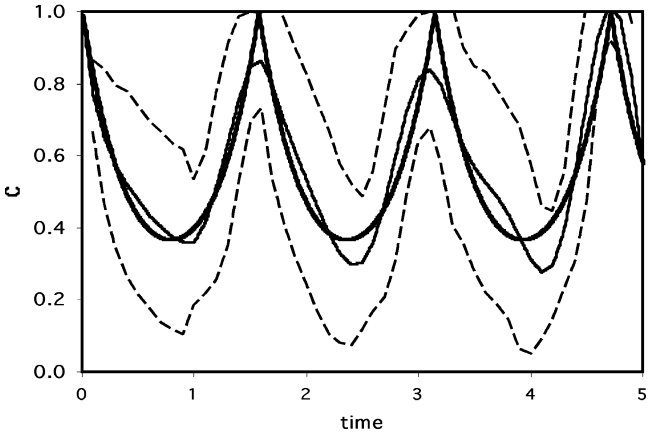

Figure 7. The rms deviation of $C(t)$ (dashed lines) showing the uncertainty of the results when the expected error is 0.01.

(eq B.5) increases the allowable noise in the experimental data by an order of magnitude. Because this method is based on an objective probability analysis, it provides an estimate of uncertainty in the predictions (e.g., see Figure 7). The above approach yields accurate results even with limited and noisy data, a situation typical for data on cellular and other complex systems. Imposing small values of (variable name of eq B.5) decreases the oscillatory behavior of C and narrows the probability distribution as well. Therefore, if a species is known to have a smooth time course, use of this information via eq B.5 leads to a narrow probability distribution of C.

F. Application to *T. brucei*

An extensive study of *T. brucei* was done by Visser and Opperdoes.[31] Steady-state concentrations of different metabolites were measured using a variety of analytical techniques such as high-pressure liquid and ion-exchange chromatography. Electrophoresis is used for protein and enzyme concentrations. Bakker et al.[32] and Navid and Ortoleva[27] studied *T. brucei* glycolysis. Navid and Ortoleva simulated glycolysis using a metabolic network that consists of 28 fast (equilibrium) and 11 slow reactions. The system consists of 59 chemical species in three compartments (mitochondrion, glycosome, and cytoplasm). In Karyote, the dynamics of the metabolite concentrations are obtained by solving[1]

$$V^{\alpha}\frac{\mathrm{d}c_i^{\alpha}}{\mathrm{d}t} = \sum_{\alpha' \neq \alpha}^{N_c} A^{\alpha\alpha'} J_i^{\alpha\alpha'} + V^{\alpha} R_i^{\mathrm{p,slow}} + V^{\alpha} \sum_{k=1}^{N_f} \nu_{ik}^{\mathrm{fast}} \frac{W_k^{\alpha,\mathrm{fast}}}{\epsilon}, \quad i = 1, 2, ..., N_s \quad \text{(F.1)}$$

where $A^{\alpha\alpha'}$ is the shared boundary surface area separating compartments α and α', $J_i^{\alpha\alpha'}$ is the net flux of species i from α to α', c_i^{α} is the concentration of species i in compartment α, N_c, N_s, and N_f are the number of compartments, chemical species, and fast reactions, respectively, $R_i^{\alpha,\mathrm{slow}}$ is the net reaction rate for slow reactions involved with chemical species j in compartment α, V^{α} is the volume of compartment α, $W_k^{\alpha,\mathrm{fast}}$ is the rate of reaction k in compartment α, ν_{ik}^{fast} is the stoichiometric coefficient for species i in reaction k in compartment α, and ϵ is the ratio of the short to long characteristic time.

We calibrate the parameters listed in Table 1 with available experimental steady-state concentrations. As seen in Tables 1 and 2, the values of the parameters developed by our procedure give lower error than those obtained by Navid and Ortoleva, which themselves give a lower error than those of Bakker et al.

G. Conclusions

The overall goal of this work is to develop a natural integration of modeling and laboratory approaches. Because there are uncertainties in both, the natural framework for this integration is information theory (i.e., probability theory). We have established the relationship between the completeness of the model and of the experimental data set. For example, an extensive data set can be used to establish relationships among variables that are not included in the physics and chemistry of the model; thus while reaction mechanism for the creation/destruction of one chemical species may not be known, our formulation shows how experimental data can be used to construct the most probable time course of concentration of this species.

The information theory framework allows for the inclusion of qualitative data/physical intuition (i.e., soft data) in a variety of ways. Experience gained in determining the most sensitive error measure for given types of model parameters and of data can be considered as information that is naturally integrated in our formulation. Qualitative information such as upper and lower limits on the time scale of processes can be naturally introduced via our regularization approach. The numerical computations can be stabilized by incorporating our knowledge of asymptotic behavior of the reaction system (e.g., when the rate coefficient for a reversible reaction is beyond a certain value, then that reaction is at equilibrium, and therefore, predictions of the model become insensitive to the value of parameter used). Finally, experience gained on reaction mechanisms analogous to those in the model of interest can be used to guide the structure of the probability distribution via an a priori information approach.

TABLE 1: List of Parameters Calibrated for the *T. brucei* Glycolysis Model[a]

parameter type	reaction	initial estimate	calibrated value
slow reaction rate coefficient	hexokinase	0.0658 (s^{-1})	0.0872 (s^{-1})
slow reaction rate coefficient	glycerol-3-phosphate dehydrogenase	0.0395 (s^{-1})	0.9040 (s^{-1})
slow reaction rate coefficient	GAP dehydrogenase	57.90 (s^{-1})	81.470 (s^{-1})
slow reaction rate coefficient	GAP dehydrogenase	57.90 (s^{-1})	25.150 (s^{-1})
slow reaction rate coefficient	phosphofructokinase	0.321 (s^{-1})	0.0912 (s^{-1})
slow reaction rate coefficient	phosphoglycerate kinase	0.125 (s^{-1})	0.0263 (s^{-1})
slow reaction rate coefficient	pyruvate kinase	0.0743 (s^{-1})	2.205 (s^{-1})
slow reaction rate coefficient	glycerol kinase	1.00 (s^{-1})	0.109 (s^{-1})
fast reaction equilibrium constant	phosphoglycerate mutase	0.187 (s^{-1})	0.128 (s^{-1})
fast reaction equilibrium constant	enolase	6.70 (s^{-1})	1.511 (s^{-1})
forward transport coefficient	glycerol-3-phosphate	4×10^{-5} ($L^{-1}\cdot s^{-1}$)	4.001×10^{-5} ($L^{-1}\cdot s^{-1}$)
forward transport coefficient	pyruvate	7×10^{-9} ($L^{-1}\cdot s^{-1}$)	4.8×10^{-9} ($L^{-1}\cdot s^{-1}$)
forward transport coefficient	dihydroxy-acetone-phosphate	1×10^{-5} ($L^{-1}\cdot s^{-1}$)	9.76×10^{-6} ($L^{-1}\cdot s^{-1}$)

[a] Fast reactions are considered to be at equilibrium; thus, only the equilibrium constants were calibrated. For slow (finite rate) reactions, rate coefficients were calibrated.

TABLE 2: Comparison of Calculated and Measured Steady-State Metabolite Concentrations for Glycolysis under Aerobic Conditions in *T. brucei*[a]

		Karyote				Bakker et al.	
species	exptl concn (aerobic) (mM)	concn (aerobic) (mM)[b]	% error[b]	calibrated concn (aerobic) (mM)	% error	concn (aerobic) (mM)	% error
G6P	4.4	1.0	77	4.40	0	0.44	90
FBP	2.4	0.55	77	2.41	1	0.13	95
F6P	1.9	1.4	26	1.93	2	26	1268
GAP	0.47	0.25	46	0.28	40	0.074	84
DHAP(g/c)	2.6	3.8	46	4.26	64	1.6	38
1−3-BPG	0.77	0.2	74	0.74	4	0.028	96
3PG(g/c)	4.8	1.7	65	4.98	4	0.68	86
2PG	0.59	0.3	49	0.60	2	0.13	78
PEP	0.85	2.0	135	0.91	7	0.85	0
pyruvate	21	21.6	3	20.7	1	21	0
nGly-3-P(g/c)	2	0.4	80	1.68	16	1.1	45

[a] In column two are the measured concentrations by Visser and Opperdoes.[31] In column three are the results for Karyote simulation of the same system. In column seven are the results of a similar simulation by Bakker et al.[32] Karyote's results have smaller average margins of error in comparison to Bakker's results. Improvement due to the use of Karyote is seen by comparing column three and seven. Improvement due to a better calibration is seen by comparing column three and five. The designations g and c denote glycosomal and cytosolic concentrations, respectively. [b] Navid and Ortoleva, 2002.

Our methodology allows for the construction of the full probability distribution. However, these computations can be carried out most efficiently when a Gaussian approximation is used to construct the probability density for the least known factors. This probability in the Gaussian approximation can be used to calculate the probability distribution for model predictions, the latter not necessarily Gaussian even though the former was. As a model, prediction is in general a complex nonlinear function of the unknown parameters.

The methodology allows for the objective integration of multiple data sets of various types and quality (e.g., NMR, mass spectroscopy, microelectrode). To take full advantage of such a spectrum of data, a model of a complex system, like a cell, must be sufficiently comprehensive to utilize a broad range of data types. For example, the model should be based on a large network of metabolic reactions to use data on small molecules and should have proteomic and genomic components to use mass spectroscopy data on tryptic digest of proteins. A key link in the utilization of a variety of experimental data types is the development of modules that transform the output of the model (concentrations of chemical species, populations of various proteins) into the experimentally measured quantities (e.g., NMR and mass spectroscopy). Thus the development of the physical models and numerical algorithms needed for the translation modules is an important next step in the development of our approach.

Acknowledgment. This work was supported in part by grants from the U.S. Air Force Research Laboratory/DARPA (Grant No. USAF/F30602-02-0001) and the U.S. Department of Energy (Grant No. DE-FG-02-02ER-25498083104). We greatly appreciate the cooperation of A. Navid (who developed the reaction network for the *T. brucei*) and E. Weitzke (who developed the original Karyote cell model); without their cooperation, this work would not have been possible.

Appendix I. One Error Type Approximation

Entropy maximization for one error type yields

$$\ln \rho = -\ln Q - \beta E(\mathbf{x}) \tag{I.1}$$

where Q is a normalization factor and β is a Lagrange multiplier. We expand ln ρ around the most probable values of the parameters set $\mathbf{x}_m$. Clearly $\mathbf{x}_m$ is maximizing the probability distribution because ln ρ is a monotonic function of ρ.

Define

$$\rho_m = \int \exp\left(\frac{-\beta \Delta \mathbf{x}^{\mathrm{T}} \mathbf{H}_{\mathrm{E}}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x}}{2}\right) \mathrm{d}\mathbf{x} \tag{I.2}$$

hence,

$$\ln \rho \approx \ln \rho_m + \Delta \mathbf{x}^{\mathrm{T}} \nabla \ln \rho|_{\mathbf{x}=\mathbf{x}_m} + \frac{1}{2} \Delta \mathbf{x}^{\mathrm{T}} \mathbf{H}_{\ln \rho}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x}$$

$$\approx \ln \rho_m + \frac{1}{2} \Delta \mathbf{x}^{\mathrm{T}} \mathbf{H}_{\ln \rho}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x} \tag{I.3}$$

This approximation is always valid when we have a narrow probability distribution where the quadratic term is the dominant factor. For multimodal probability distribution, when one of the maxima is much larger than the others, it is legitimate to ignore the latter. This allows us to give a complete description of the probability using few parameters (i.e., averages and variances). However, the idea of best estimate and confidence intervals would be irrelevant when the multimodal probability density has comparable maxima.

Now, we have

$$\rho(\mathbf{x}) \approx \rho_m \exp \frac{-\beta \Delta \mathbf{x}^{\mathrm{T}} \mathbf{H}_{\mathrm{E}}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x}}{2} \tag{I.4}$$

Normalization implies

$$\rho_m \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \frac{-\beta \Delta \mathbf{x}^{\mathrm{T}} \mathbf{H}_{\mathrm{E}}|_{\mathbf{x}=\mathbf{x}_m} \Delta \mathbf{x}}{2} \mathrm{d}x_1 \cdots \mathrm{d}x_{N_p} = 1 \tag{I.5}$$

Spectral decomposition of the error Hessian matrix ($\mathbf{H}_{\mathrm{E}}$) implies

$$\mathbf{H}_{\mathrm{E}} = \mathbf{G}^{\mathrm{T}} \lambda \mathbf{G} \tag{I.6}$$

Because the Hessian matrix of the error is positive definite around $\mathbf{x}_m$, one can evaluate the quadratic integration in eq I.4 analytically. By doing so, we find

$$\rho_m = \prod_{i=1}^{N_p} \sqrt{\frac{\beta \lambda_i}{2\pi}} \tag{I.7}$$

where N_p is the number of model parameters. β is evaluated using the error constraint (eq B.4) and the quadratic expansion of E. With the use of the transformation $\Delta\mathbf{x} = \mathbf{G}^T\mathbf{s}$, the constraint integral

$$\int\left\{E_m + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}_E\Delta\mathbf{x}\right\}\rho(\mathbf{x})\,d\mathbf{x} \approx E^* \tag{I.8}$$

can be transformed to

$$E_m + \frac{\rho_m}{2}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} ds_1... ds_{N_p}\sum_{i=1}^{N_p}\lambda_i {s_i}^2 \exp\left(-\frac{\beta}{2}\sum_{i=1}^{N_p}\lambda_i {s_i}^2\right) = E^* \tag{I.9}$$

Evaluating the above integral yields

$$E_m + \frac{\rho_m}{2}\sum_{i=1}^{N_p}\lambda_i\sqrt{\frac{2\pi}{(\lambda_i\beta)^3}}\prod_{\substack{j=1\\ i\neq j}}^{N_p}\sqrt{\frac{2\pi}{\lambda_j\beta}} = E^* \tag{I.10}$$

And hence, we get

$$\beta = \frac{N_p}{2(E^* - E_m)} \tag{I.11}$$

Substituting β from eq I.11 into eq I.7 yields

$$\rho_m = \prod_{i=1}^{N_p}\sqrt{\frac{N_p\lambda_i}{4\pi(E^* - E_m)}} \tag{I.12}$$

Similarly we can calculate the expected value of a model predictable output P to be

$$\bar{P} = P(\mathbf{x}_m) + \frac{(E^* - E_m)}{N_p}\sum_{i=1}^{N_p}\frac{\gamma_{ii}}{\lambda_i} \tag{I.13}$$

where γ_{ii} are the diagonal elements of the transformed P Hessian matrix $\boldsymbol{\gamma} = \mathbf{G}^T\mathbf{H}_P\mathbf{G}$. E_m is the error evaluated at $\mathbf{x}_m$. The uncertainty in the predictions can be deduced from the probability distribution of the input parameters. We expand a predictable quantity P around the most probable values of model parameters. Taylor expansion of P gives

$$P(\mathbf{x}) \approx P(\mathbf{x}_m) + \Delta\mathbf{x}^T\nabla P + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}_P\Delta\mathbf{x} \tag{I.14}$$

and

$$P(\mathbf{x})^2 \approx P(\mathbf{x}_m)^2 + \Delta\mathbf{x}^T\nabla P\nabla P^T\Delta\mathbf{x} + P(\mathbf{x}_m)\Delta\mathbf{x}^T\mathbf{H}_P\Delta\mathbf{x} + \frac{1}{4}\Delta\mathbf{x}^T\mathbf{H}_P\Delta\mathbf{x}\Delta\mathbf{x}^T\mathbf{H}_P\Delta\mathbf{x} + ... \tag{I.15}$$

Now,

$$\langle P(\mathbf{x})^2\rangle = \rho_m\int P(\mathbf{x})^2 e^{-\beta E(\mathbf{x})}\,d\mathbf{x} \tag{I.16}$$

can be approximated by

$$P(\mathbf{x}_m)^2 + \frac{2P(\mathbf{x}_m)(E^* - E_m)}{N_p}\sum_{i=1}^{N_p}\left(\frac{\gamma_{ii}}{\lambda_i} + \frac{\kappa_{ii}}{\lambda_i P(\mathbf{x}_m)}\right) + \frac{3(E^* - E_m)^2}{{N_p}^2}\sum_{i,j=1}^{N_p}\frac{{\gamma_{ij}}^2}{\lambda_i\lambda_j} \tag{I.17}$$

where $\kappa = \mathbf{G}\nabla P\nabla P^T\mathbf{G}^T$.

With use of eqs I.13 and I.17, the uncertainty $\sigma_P^2 = \langle P(\mathbf{x})^2\rangle - \langle P(\mathbf{x})\rangle^2$ in a model prediction P is found to be

$$\sigma_P^2 = \frac{2(E^* - E_m)}{N_p}\sum_{i=1}^{N_p}\frac{\kappa_{ii}}{\lambda_i} + \frac{2(E^* - E_m)^2}{{N_p}^2}\sum_{i,j=1}^{N_p}\frac{{\gamma_{ij}}^2}{\lambda_i\lambda_j} \tag{I.18}$$

Appendix II. Multiple Error Types Approximation

For multiple error types, we need to solve the highly nonlinear constraints equations

$$\frac{\partial \ln Q}{\partial\beta_k} + E^{(k)*} = 0, \quad k = 1, 2, ..., N_e \tag{II.1}$$

where N_e is the number of error types and Q is the partition function for multiple error types, which can be evaluated by

$$Q = \int \exp(-\sum_{k=1}^{N_e}\beta_k E^{(k)})\,d\mathbf{x} \tag{II.2}$$

Denote the integration over parameters space by $\langle\ \rangle$; then the Jacobian of the above nonlinear system (eq II.1) is found to be

$$\frac{\partial^2 \ln Q}{\partial\beta_i\partial\beta_j} = \langle E^{(i)}E^{(j)}\rangle - \langle E^{(i)}\rangle\langle E^{(j)}\rangle, \quad i,j = 1, 2, ..., N_e \tag{II.3}$$

which reduces to

$$\frac{\partial^2 \ln Q}{\partial\beta_i\partial\beta_j} = \langle E^{(i)}E^{(j)}\rangle - E^{(i)*}E^{(j)*}, \quad i,j = 1, 2, ..., N_e \tag{II.4}$$

or

$$\frac{\partial^2 \ln Q}{\partial\beta_i\partial\beta_j} = \langle (E^{(i)} - E^{(i)*})(E^{(j)} - E^{(j)*})\rangle, \quad i,j = 1, 2, ..., N_e \tag{II.5}$$

Now, consider an arbitrary $\boldsymbol{\beta} \neq 0$,

$$\boldsymbol{\beta}^T\langle(\mathbf{E} - \mathbf{E}^{(*)})^T(\mathbf{E} - \mathbf{E}^{(*)})\rangle\boldsymbol{\beta} = \langle\boldsymbol{\beta}^T(\mathbf{E} - \mathbf{E}^{(*)})^T(\mathbf{E} - \mathbf{E}^{(*)})\boldsymbol{\beta}\rangle \tag{II.6}$$

which implies

$$\boldsymbol{\beta}^T\langle|(\mathbf{E} - \mathbf{E}^{(*)})\boldsymbol{\beta}|^2\rangle > 0 \tag{II.7}$$

that is, the Jacobian of the nonlinear system eq II.1 is positive definite. This is a necessary condition for having a unique solution of eq II.1. Another necessary condition for having a solution is

$$(E^{(k)*} - E^{(k)}_{min}) > 0, \quad k = 1, 2, ..., N_e \tag{II.8}$$

However, the above nonlinear system and the Jacobian evaluations are computationally expensive. Using Monte Carlo methods to calculate the partition function Q needs thousands of model runs. If quadratic terms are dominant in the error functions, then a Gaussian approach for the probability distribution can be taken as follows.

Denote $E = \sum_{k=1}^{N_e}\beta_{(k)}E^{(k)}$; then the quadratic expansion of ln ρ around a reference point $\mathbf{x}_r$ yields

$$\ln\rho \approx \ln\rho_r - \Delta\mathbf{x}^T\cdot\nabla E|_{\mathbf{x}=\mathbf{x}_r} - \frac{1}{2}\Delta\mathbf{x}^T\mathbf{J}|_{x=x_r}\Delta\mathbf{x} \tag{II.9}$$

where

$$\nabla E = \sum_{k=1}^{N_e} \beta_{(k)} \nabla E^{(k)} \tag{II.10}$$

and

$$J_{ij} = \sum_{k=1}^{N_e} \beta_{(k)} \frac{\partial^2 E^{(k)}}{\partial x_i \partial x_j} \tag{II.11}$$

Substituting eq II.8 into normalization, we get

$$\rho_r \int \exp\left(-\Delta \mathbf{x}^T \cdot \nabla E\Big|_{\mathbf{x}=\mathbf{x}_r} - \frac{1}{2}\Delta \mathbf{x}^T \mathbf{J}\Big|_{\mathbf{x}=\mathbf{x}_r} \Delta \mathbf{x}\right) d\mathbf{x} = 1 \tag{II.12}$$

The spectral decomposition of E Hessian matrix $\mathbf{J}$ implies

$$\mathbf{J} = \mathbf{G}^T \lambda \mathbf{G} \tag{II.13}$$

and

$$|\mathbf{G}| = 1 \tag{II.14}$$

We transform the integration variables using

$$\Delta \mathbf{x} = \mathbf{G}^T \mathbf{s}, \tag{II.15}$$

and

$$\mathrm{X} = \mathbf{G} \nabla E|_{\mathbf{x}=\mathbf{x}_r} \tag{II.16}$$

Equation II.12 then becomes

$$\rho_r \prod_{i=1}^{N_p} \int_{-\infty}^{\infty} \exp\left(-\mathrm{X}_i s_i - \frac{1}{2}\lambda_i {s_i}^2\right) ds_i = 1 \tag{II.17}$$

which can be simplified to

$$\rho_r \prod_{i=1}^{N_p} \sqrt{\frac{2\pi}{\lambda_i}} \exp\left(\frac{{\mathrm{X}_i}^2}{2\lambda_i}\right) = 1 \tag{II.18}$$

Therefore

$$\frac{1}{Q} \approx \rho_r = \prod_{i=1}^{N} \sqrt{\frac{\lambda_i(\boldsymbol{\beta})}{2\pi}} \exp\left(-\frac{{\mathrm{X}_i}^2(\boldsymbol{\beta})}{2\lambda_i(\boldsymbol{\beta})}\right) \tag{II.19}$$

One can use the partition function Q to approximate the constraints (eq II.1) and the Jacobian (eq II.4) using a forward difference scheme. Similarly, we can expand $E^{(k)}(\mathbf{x})$ around $\mathbf{x}_r$,

$$E^{(k)}(\mathbf{x}) = E^{(k)}(\mathbf{x}_r) + \Delta \mathbf{x}^T \nabla E^{(k)}|_{\mathbf{x}=\mathbf{x}_r} + \frac{1}{2}\Delta \mathbf{x}^T \mathbf{J}^{(k)}|_{\mathbf{x}=\mathbf{x}_r} \Delta \mathbf{x} \tag{II.20}$$

Substituting in eq B.4, we get

$$\rho_r \int E^{(k)} \exp\left(-\Delta \mathbf{x}^T \cdot \nabla E\Big|_{\mathbf{x}=\mathbf{x}_r} - \frac{1}{2}\Delta \mathbf{x}^T \mathbf{J}\Big|_{\mathbf{x}=\mathbf{x}_r} \Delta \mathbf{x}\right) d\mathbf{x} \approx E^{(k)*} \tag{II.21}$$

This reduces to

$$E^{(k)}(\mathbf{x}_r) + \frac{1}{2}\left[\prod_{i=1}^{N_p} \sqrt{\frac{\lambda_i}{2\pi}} \exp\left(-\frac{{\mathrm{X}_i}^2}{2\lambda_i}\right)\right] \int \mathbf{s}^T \mathbf{G}_r^T \mathbf{J}_r^{(k)} \mathbf{G}_r \mathbf{s} \exp\left(-\mathrm{X}_r^T \mathbf{s} - \frac{1}{2}\mathbf{s}^T \lambda \mathbf{s}\right) d\mathbf{s} \approx E^{(k)*} \tag{II.22}$$

Now let $\boldsymbol{\theta}_r = \mathbf{G}_r^T \mathbf{J}_r^{(k)} \mathbf{G}_r$; then the preceding becomes

$$E^{(k)}(\mathbf{x}_r) + \frac{1}{2}\left[\prod_{i=1}^{N_p} \sqrt{\frac{\lambda_i}{2\pi}} \exp\left(-\frac{{\mathrm{X}_i}^2}{2\lambda_i}\right)\right] \int \mathbf{s}^T \boldsymbol{\theta}_r \mathbf{s} \exp\left(-\mathrm{X}_r^T \mathbf{s} - \frac{1}{2}\mathbf{s}^T \lambda \mathbf{s}\right) d\mathbf{s} = E^{(k)*} \tag{II.23}$$

which reduces to

$$E^{(k)}(\mathbf{x}_r) + \frac{1}{2}\left[\prod_{i=1}^{N_p} \sqrt{\frac{\lambda_i}{2\pi}} \exp\left(-\frac{{\mathrm{X}_i}^2}{2\lambda_i}\right)\right] \sum_{j=1}^{N_p} \sum_{k=1}^{N_p} \theta_{jk} \frac{(\lambda_j + {\mathrm{X}_j}^2)}{{\lambda_j}^2} \prod_{l=1}^{N_p} \sqrt{\frac{2\pi}{\lambda_l}} \exp\left(\frac{{\mathrm{X}_l}^2}{2\lambda_l}\right) = E^{(k)*} \tag{II.24}$$

that is,

$$\frac{1}{2}\sum_{j=1}^{N_p} \left(\sum_{k=1}^{N_p} \theta_{jk}\right) \frac{(\lambda_j(\mathbf{x}_r,\boldsymbol{\beta}) + {\mathrm{X}_j}^2(\mathbf{x}_r,\boldsymbol{\beta}))}{{\lambda_j}^2(\mathbf{x}_r,\boldsymbol{\beta})} = E^{(k)*} - E^{(k)}(\mathbf{x}_r,\boldsymbol{\beta}) \tag{II.25}$$

which is an equivalent approximation to eq II.1.

References and Notes

(1) Weitzke, E.; Ortoleva, P. Simulating Cellular Dynamics Through a Coupled Transcription, Translation, Metabolic Model. *Computat. Biol. Chem.*, in press.

(2) Mendes, P. *Comput. Appl. Biosci.* **1993**, *9*, 563.

(3) Mendes, P. *Trends Biochem. Sci.* **1997**, *22*, 361.

(4) Mendes, P. Presented at NATO ARW on Technological and Medical Implications of Metabolic Control Analysis, Visegard, Hungary, 1999.

(5) Mendes, P. *NATO Sci. Ser., 3; High Technol.* **2000**, *74*, 149.

(6) Mendes, P.; Kell, D. B. *NATO Sci. Ser., 3; High Technol.* **2000**, *74*, 3.

(7) Mendes, P.; Martins, A. M.; Cordeiro, C.; Freire, A. P. *Eur. J. Biochem.* **2001**, *268*, 3930.

(8) Mendes, P.; Kell, D. B. *Bioinformatics* **2001**, *17*, 288.

(9) Bartol, T.; Stiles J. R.; Sejnowski, T.; Salpeter, M. Salpeter E. *MCell is: A General Monte Carlo Simulator of Cellular Microphysiology*; 1997. Found on website http://www.mcell.cnl.salk.edu.

(10) Bartol, T.; Stiles J. R.; Sejnowski, T.; Salpeter, M.; Salpeter E. *Synapses* **2001**, 681.

(11) Tomita, M.; Hashimoto, K.; Shimizu, T. S.; Matsuzaki, Y.; Miyoshi, F.; Saito, K.; Tanida, S.; Yugi, K.; Venter, J. C.; Hutchison, C.A. *Bioinformatics* **1999**, *15*, 72.

(12) Tomita, M.; Hashimoto, K.; Takahashi, K.; Matsuzaki, Y.; Matsushima, R.; Saito, K.; Yugi, K.; Miyoshi, F.; Nakano, H.; Tanida, S.; Saito, Y.; Kawase, A.; Watanabe, N.; Shimizu, T. S.; Nakayama, Y. *New Gener. Comput.* **2000**, *18*, 1.

(13) Schaff, J.; Fink, C.; Slepchenko, B.; Carson, J.; Loew, L. *Biophys. J.* **1997**, *73*, 1135.

(14) Schaff, J.; Loew, M. *Biocomputing Proceedings of the 1999 Pacific Symposium*; World Scientific Publishing Co.: River Edge, NJ, 1999.

(15) Schaff, J.; Loew, M.; Fink, C.; Slepchenko, B.; Moraru, I.; Watras, J. *Biophys. J.* **2000**, *79*, 163.

(16) Schaff, J.; Loew, M.; Slepchenko, B.; Choi, Y.; Wagner, J.; Resasco, D. *Chaos* **2001**, *11*, 115.

(17) Schaff, J.; Loew, M. *Trends Biotechnol.* **2001**, *19*, 401.

(18) Schaff, J.; Loew, M.; Fink, C.; Slepchenko, B.; Carson, J. H. *Annu. Rev. Biophys. Biomol. Struct.* **2002**, *31*, 423.

(19) Kirsch, A. *An Introduction to the Mathematical Theory of Inverse Problems*; Springer-Verlag: New York, 1996.

(20) Levenberg, K. *Q. Appl. Math.* **1944**, *2*, 164.
(21) Marquardt, D. W. *SIAM J.* **1963**, *189*, 421.
(22) Kytomaa, H.; Weselake, K. *Comput. Mech.* **1994**, *15*, 161.
(23) Rao, L.; He, R.; Wang, Y.; Yan, W.; Bai, J.; Ye D. *IEEE Trans. Magn.* **1999**, *35*, 1562.
(24) Player, M. A.; Weereld, J.; Allen, A. R.; Collie, D. A. L. *Electron. Lett.* **1999**, *35*, 2189.
(25) Torres, N. V.; Voit, E. O.; Glez-Alcon, C.; Rodriguez, F. *Biotech. Bioeng.* **1997**, *55*, 758.
(26) Mendes, P.; Kell, D. B. *Bioinformatics* **1998**, *14*, 869.
(27) Navid, A.; Ortoleva, P. *J. Theor. Biol.*, submitted for publication, 2003.
(28) Ortoleva, P., manuscript in preparation.
(29) Jaynes, E. T. *Phys. Rev.* **1957**, *106*, 620.
(30) Tuncay, K.; Ortoleva, P. *Institute of Mathematics and its Applications (IMA)*; Springer-Verlag: New York, 2002; pp 131, 161.
(31) Visser, N.; Opperdoes, F. R. *Eur. J. Biochem.* **1980**, *103*, 623.
(32) Bakker, B. M.; Michels, P. A. M.; Opperdoes, F. R.; Westerhoff, H. V. *J. Biol. Chem.* **1997**, *272*, 3207.