CONTRIBUTED PAPER

Addressing the Need for a Model Selection Framework in Systems Biology Using Information Theory

This paper develops the argument that information-theoretic model selection metrics should be extended to nonnested model comparison applications in systems biology.

By Frank DeVilbiss and Doraiswami Ramkrishna

ABSTRACT | The field of systems biology thrives upon the use of models to organize biological knowledge and make predictions of complex processes that are hard to measure. When attempting to generate model descriptions for metabolic systems, one arrives at a crossroads. A variety of mathematical explanations are available for metabolic data with varying degrees of resolution from simple to complex. Biological modelers often rely upon subjective arguments to choose one framework over another. While there is no universal rule to determine the absolute utility of a model, certain metrics founded on information theoretical principles, demonstrate promise in providing a coherent, rational, and objective basis for addressing this model selection problem in systems biology. A model seeks to capture the regularity in biological data. Models that best capture regularity in data without excessive complexity are the most useful for applications in optimization and control. To demonstrate the efficacy of such an approach, several metabolic model selection scenarios are investigated. This work develops the argument that information theoretic model selection metrics should be extended to nonnested model comparison applications in systems biology. It also makes a novel comparison of kinetic, constraint-based, and cybernetic models of metabolism based not only on model accuracy, but also model complexity. The results show the strengths of lumped hybrid cybernetic model (L-HCM) and flux balance analysis (FBA) for applications in steady state flux prediction. Also, the hybrid cybernetic model's (HCM) merit in the modeling of dynamic changes in fluxes is also established.

KEYWORDS | Biological systems modeling; cybernetics; information theory

I. INTRODUCTION

When given an arbitrary set of data, one can generate a host of different mathematical descriptions for it. Metabolic systems are no exception and embody an important branch of systems biological study. In order to predict the effects of perturbations to metabolic networks such as deleting genes or inhibiting enzymes, it is useful to first use a model to understand, without additional experimentation, the effects of such modifications. To model the changes in metabolic systems, one can select kinetic, constraint-based, or cybernetic formulations. Each of these metabolic models is unique in formulation and widely used for similar goals.

In very general terms, the utility of a model is derived from its ability to describe regularity in data. Regularity, or coherence in a set of data, means that the data are generated as the result of some intelligible process [1]. For metabolic flux data, each type of model is able to capture the coherence of metabolic processes to a certain degree. These models also have varying degrees of complexity which are used to explain the behavior of data. To establish which model is best for the purposes of

Manuscript received September 14, 2015; accepted February 20, 2016. Date of publication June 7, 2016; date of current version January 18, 2017. This work was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370.

The authors are with the School of Chemical Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: ramkrish@ecn.purdue.edu).

Digital Object Identifier: 10.1109/JPROC.2016.2560121

optimization and control, it is proposed that models be selected on the basis of how well they are able to capture the regularity of data without being excessively complex.

In this work, a number of widely used mathematical approximations of metabolic systems are compared according to the ability of each to capture regularity in data. While there is much discourse on the merits of each modeling framework [2]–[6], no systematic method has been implemented to quantitatively and simultaneously measure the relative accuracies and complexities of metabolic models. This work applies information theoretic metrics, well known in other fields, to address this problem. Treating each model as an entity that compresses data for communication through a channel, one can quantitatively evaluate how well a model balances accuracy and complexity. Models that accurately reproduce data with low complexity require less information to communicate and embody a more compressed description of a process's data. This method of evaluation is especially useful in situations where model formulations are vastly different (i.e., nonnested).

The establishment of the model that best minimizes penalties for both error and parameterization is deemed to be the best model for an application related to the optimization and control of biological systems. Restated, there is a point in diminishing returns for model complexity. Additional parameters may enhance accuracy, but each additional parameter has an intrinsic cost associated with it.

A set of four distinct models of metabolic fluxes are judged in their ability to describe metabolic reaction rates at a given steady state. Following this, dynamic metabolic models are compared in their ability to predict changing metabolic fluxes. These dynamic models of metabolic fluxes have never been compared in this objective fashion. Neither have such a wide range of steady state descriptions of metabolic fluxes been compared. The application of these metrics in these scenarios helps to establish a new way of thinking about metabolic model selection. Moreover, a quantitative framework for comparing nonnested biological models is necessary to introduce to the field of systems biology.

II. METHODS

A. Theory

To develop the model selection framework, it is first useful to review some basic tenets of information theory for those who might be unfamiliar. In the field of communication, signals that are being passed through a channel are analyzed and compressed depending on how much regularity is present in a given message. Compression is a useful tool in that shorter messages lead to faster communication. For example, consider the two messages below:

1) 01101011101001000 1011;
2) 00001000000000000000.

Sequence 1) is generated from some arbitrary process where either 0 or 1 share equal probabilities of occurrence. On the other hand, sequence 2) is generated from a process where the probability of 1 is 1/20. To communicate either sequence, one could send each 1 or 0 value individually or one could compress the information down into a shorter sequence. Given the fact that either ones or zeros are equally probable in 1), it is virtually incompressible. On the other hand, sequence 2) can be shortened in a number of ways.

For example, sequence 2) can be simply described by specifying the position of the single one rather than the whole sequence assuming the decoder understands the compression scheme. In case 2), one could rewrite the sequence as "5" specifying the location of the single 1. For this 20-b sequence, specifying the position of a single one in any of 20 possible locations requires up to 5 b ($\lceil(\log_2(20)\rceil$) instead of 20 for compressing any position in the sequence of 20 b. This represents a compression factor of 0.25 compared to communicating the entirety of the original sequence. Other, more efficient coding schemes are possible and the fundamental limit of compression of these data sequences is quantified using Shannon's entropy [7]

$$H(x) = \sum_i p_i \log(p_i) \tag{1}$$

in which p_i represents the probability of a 0 or 1 occurring in the sequence. The motivation for compression is increasing the overall rate of communication. The more compressed a message is, the less time is spent communicating it. In terms of entropy, the highest entropy sequence will consist of bits generated by the method of 2). In the same way that a message can be compressed by a proper coding scheme, we can say that biological data can be compressed by a model. It is here where the minimum description length principle (MDL) becomes useful in that one can reinterpret the model selection problem as one of data compression [1]. The aim is to shrink the data D into some D' from which D can be perfectly reconstructed after compression. For some model M, there is a length of the data $L(D')$ that is determined as

$$L(D') = L(D|M) + L(M) \tag{2}$$

where $L(D|M)$ expresses the data in terms of the model and $L(M)$ is a description of the models complexity [1], [8]. The term $L(D|M)$ accounts for the extra information that needs to be transmitted in order to describe the model prediction's distance from the real data. For example, if the model prediction comes close to the data, less information is needed to communicate the model error

than if the prediction is far from the data in the same way that smaller integers can be communicated by fewer bits than larger ones (e.g., in the case of integers, the number 2 can be encoded into binary as "10" at 2 b versus 20 as "10100" at 5 b). The complexity term $L(M)$ defines the amount of information that must be communicated in order to describe the model and is typically defined by the number of parameters in the model. Together, the model and a specification of its error can be used to perfectly reconstruct the data from D' to D.

To represent $L(D')$ for some model, one can apply metrics such as Akaike's information criterion (AIC) [9] or Schwarz's Bayesian information criterion (BIC) [10]. These information theoretic metrics take the form of either

$$\mathrm{AIC} = n\log(\hat{\sigma}^2) + 2k \tag{3}$$

or

$$\mathrm{BIC} = n\log(\hat{\sigma}^2) + k\log(n). \tag{4}$$

In the above, $\hat{\sigma}^2$ is the error of the model, k is the number of parameters within the model, and n is the number of data points that the model approximates. Note that these metrics are valid asymptotically and correction factors are applied in the case of limited data. These correction factors increase the penalty on extra parameters when there are fewer data points.

AIC is formulated using Kullback–Liebler divergence and seeks to select a candidate model that best describes reality. This is due to the fact that KL divergence is a measure of the extra information needed to transmit some information using a model distribution as compared to some real generating distribution [11]. BIC, built from a Bayesian arguments, seeks to select a "true" model from a possible set of models. More specifically, it applies a likelihood function to gauge the probability that a model is true given some observed data. AIC penalizes parameters less severely than BIC which means that BIC tends to favor simpler models than AIC. For derivations of these metrics, one should consult [9] and [10].

Both metrics above are made from different arguments, but they embody a similar principle. A model that optimally describes the data will strike a balance between accuracy and complexity.

To further explain these metrics in the context of a communication problem, consider the transmission of the model in the place of the raw data through a communication channel. In order to communicate the model itself, the parameters need to be transmitted with a certain accuracy. Encoding parameters to a precision of $\delta_m = 1/\sqrt{n}$ is the most reasonable way to do this as

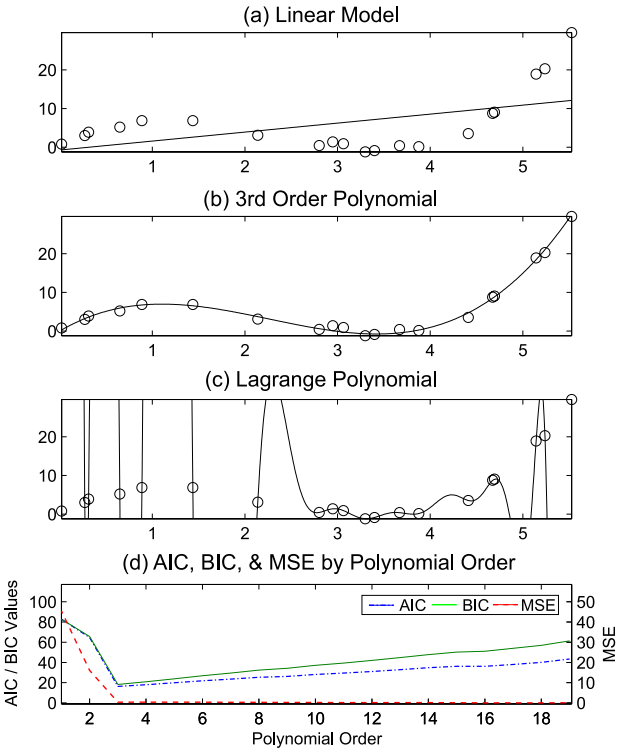

Fig. 1. ***Example of model selection problem with polynomials. (a) Fit of a linear polynomial for the data. (b) Fit of the third order polynomial. (c) Fit of a Lagrange polynomial of order $n-1$. (d) Behavior of MSE in red while AIC and BIC are shown in blue and green, respectively, for each order of polynomial. Note that both metrics are minimized for the third-order polynomial model.***

$1/\sqrt{n}$ represents the magnitude of estimation error on the parameters themselves [12]. To transmit a model's k parameters to this precision through the channel, one will need to use $-k\log_2 1/\sqrt{n}$ bits which makes up the latter portion of BIC.

The error of the model will also require communication which is approximated by the mean squared error for the data set. Mean squared error represents the average magnitude of error in the description of each data point. For a particular data point, magnitude of error is relevant because, as stated previously, larger numbers require more bits to communicate. When considering these model metrics in the context of data transfer, they are formally referred to as two-stage description length or two-stage MDL. Note that two-stage MDL has the same form as BIC.

To further illustrate the use of these information criteria, let us consider some arbitrary data set as shown in Fig. 1. There are a range of polynomial models that one could use to fit the n data points from an overly simple linear model of order 1 in Fig. 1(a) to an overfitting Lagrange polynomial of order $n-1$ in Fig. 1(c). While

the Lagrange polynomial captures the data with no error, the third-order polynomial has a better qualitative fit to the data.

When using a model for the purposes of optimization and control, one desires a model that will be accurate without being overly complex. By applying AIC and BIC to the analysis of biological models, one can gain a relative sense of a model's balance between accuracy and complexity relative to other descriptions of the data. An important advantage offered by these metrics is that they can be used to compare nonnested models. In other words, these information criteria are valid for comparing models of vastly different formulations such as constraint-based and differential equation models.

The model selection approach highlighted contrasts with other methods of model selection that focus merely on accuracy. Approaches such as cross validation are useful for comparing models but do not offer any insight into the relative complexity of different models. Other methods that seek to prevent overfitting such as regularization are not necessarily useful for the applications discussed here.

B. Metabolic Models

To demonstrate the value of these information criteria for biological modeling applications, a set of models will be developed to describe the same set of biological data. This set of models includes kinetic, constraint-based, and cybernetic models. These models can all be used to predict metabolic fluxes, or the rates of different intracellular chemical reactions, in cultures of cells growing on different carbon sources. These models are all developed using experimental data representing quantities such as the dynamic changes in concentration of carbon sources, and the growth rates of cells to arrive at these predictions. Each model has varying complexity and predicts metabolic fluxes with different amounts of error. It is the goal of this work to show how well each model balances between these.

To understand the general features of each modeling framework, the structure of all three classes of models will be highlighted. To start, one must consider the nature of metabolic systems. They are composed of connected chemical reactions that form networks. This network is articulated in a stoichiometric matrix $\mathbf{S}$ of m metabolites by n reactions. To model the changes in extracellular species, one can use

$$\mathbf{x} = \begin{bmatrix} \mathbf{s} \\ \mathbf{p} \\ c \end{bmatrix} \tag{5}$$

where $\mathbf{s}$ and $\mathbf{p}$ are vectors of n_s substrates and n_p products, respectively, and c is the concentration of cells often referred to as biomass. Combining the extracellular variable $\mathbf{x}$ with a vector for the cell density normalized intracellular components $\mathbf{m}$ yields an expression that describes the time rate of changes of extracellular and intracellular variables

$$\begin{bmatrix} \frac{1}{c}\frac{d\mathbf{x}}{dt} \\ \frac{d\mathbf{m}}{dt} \end{bmatrix} = \mathbf{S}\mathbf{r}. \tag{6}$$

Above, $\mathbf{r}$ represents the rates of metabolic reactions or fluxes. In the kinetic model, this differential expression is solved using expressions for $\mathbf{r}$ that approximate the fluxes of the chemical reactions as a function of $\mathbf{x}$ and $\mathbf{m}$. These flux expressions typically use Michaelis–Menten kinetics such as

$$r_i = \frac{V_i^{\max} m_i}{K_i + m_i} \tag{7}$$

where $V_i^{\max}$ and K_i are the maximum reaction rate and saturation constants, respectively. These parameters rely on experimental data and can change significantly for different reactions. Given that metabolic networks can be composed of thousands of reactions, kinetic models can be quite complex. Also, the kinetics used can also include enzyme influences and reaction inhibition. Kinetic models are typically very high-resolution pictures of cellular processes.

Constraint-based models such as flux balance analysis (FBA) embody a much simpler approach to predicting metabolic fluxes [13]. FBA makes two major assumptions to do so. One is that intracellular metabolites are at some pseudosteady state or

$$\frac{d\mathbf{m}}{dt} = 0. \tag{8}$$

The other is that the cells organize their metabolic fluxes to optimize some objective function. This objective function typically takes the form of maximizing the yield of biomass. This objective function will be used in all of the proceeding scenarios. FBA is written as an optimization problem as

$$\begin{aligned} \max \quad & J = \mathbf{c}^{\mathbf{T}}\mathbf{r} \\ \text{subject to} \quad & \mathbf{S}\mathbf{r} = 0 \\ & \mathbf{a} < \mathbf{r} < \mathbf{b} \end{aligned} \tag{9}$$

and can be solved using LP. Above, the product $\mathbf{c}^{\mathbf{T}}\mathbf{r}$ represents the combination of fluxes that are maximized.

Fluxes are constrained to be in the null space of **S** and must satisfy a specified set of upper and lower bounds referred to as **b** and **a**. These lower and upper bounds are determined by experimental evidence as well as thermodynamic constraints on the reactions (i.e., some reactions only work in the forward direction). The experimental evidence is typically used to constrain the uptake rates of substrates consumption and product formation in the model. Other intracellular constraints can be used, but these quantities can be difficult to measure. To model the dynamic changes of fluxes in this work, the static optimization approach will be used from dFBA where a model will be used to approximate the changes in constraints over time [14], [15].

Cybernetic models use dynamic objective functions that optimize the system to achieve goals at each time through the inclusion of control variables that regulate enzyme synthesis and activity. Instead of exhaustively describing the kinetics of each reaction as the kinetic model does, hybrid cybernetic models (HCMs) decompose the reaction network into a set of pathways or macroscopic reactions termed elementary modes (EMs) [16] that are expressed at varying levels over time. To do so, the pseudosteady state assumption must be made like in FBA. The flux vector can be decomposed into a set of rates through the EMs as

$$\mathbf{r} = \mathbf{Z}^T \mathbf{r}_M \tag{10}$$

where $\mathbf{Z}$ represents the network's n_Z EMs and $\mathbf{r}_M$ represents the regulated uptake rate of each EM. Given the use of the pseudosteady state hypothesis, the changes in extracellular concentrations are tracked in the following way

$$\frac{1}{c}\frac{d\mathbf{x}}{dt} = \mathbf{S_x Z^T r_M}. \tag{11}$$

As stated previously, in kinetic models, the reaction rates of each chemical transformation are tracked. In HCM, the regulated rate expressions for $\mathbf{r}_M$ take similar Michaelis–Menten forms, however, they include regulation v_i and enzyme e_i terms

$$r_{M,i} = e_i v_i \frac{r_{M,i}^{\text{kin,max}} s_i}{K_{M,i} + s_i}. \tag{12}$$

Above, parameters $r_{M,i}^{\text{kin,max}}$ and $K_{M,i}$ are similar to the parameters from the kinetic model with one main distinction. They describe the rate of uptake into an EM or set of reactions instead of the rate of a single reaction. In HCM, the change of enzymes which regulates $\mathbf{r}_M$ is

$$\frac{de_i}{dt} = \alpha + u_i \frac{k_{E,i} s_i}{K_i' + s_i} - (\mu + \beta) e_i. \tag{13}$$

In HCM, enzymes are generated by a constitutive formation and induced formation which make up the first two terms above. The last is an expression for the depletion of enzymes due to growth dilution and degradation. Cybernetic control variables u_i and v_i guide the induced synthesis of enzymes and the allosteric regulation of enzyme activity, respectively. Induced enzyme formation is expressed as some function of each pathways return on investment (ROI) p_i. ROIs typically are defined as each pathway's rate of substrate uptake or growth rate. To calculate the control of enzyme formation, ROIs are compared for each pathway in

$$u_i = \frac{p_i}{\sum_j p_j} \tag{14}$$

where the denominator represents the sum of ROIs for all pathways. This means that the fraction of a finite resource pool devoted to the production of enzymes for one pathway is proportional to the ROI for that pathway. Similarly, the activity of the different metabolic pathways is controlled by the v_i variable which takes the form

$$v_i = \frac{p_i}{\max_j p_j}. \tag{15}$$

The pathway with the highest ROI will be fully expressed. All other pathways with lower ROIs will be down-regulated proportionally.

For a given metabolic network, the number of EMs can be quite high. Therefore, yield analysis is used to reduce the EMs in HCM down to a minimal set that spans a given yield space [17]. This makes the generation of an HCM model for the subsequent results facile and reduces the number of parameters for the model's specification.

Another version of cybernetic models that will be analyzed is the L-HCM. The formulation of this model is quite similar to HCM with one main distinction. Instead of enumerating uptake rate constants for all pathways, the EMs are lumped together into families based on their structural returns on investment. These family modes are then expressed as a function of some dynamic metabolic objective function. The procedure for lumping EMs is somewhat complex and is best explained in [18].

Both HCM and L-HCM employ objective functions to dynamically maximize the rate of carbon uptake in the models used in the subsequent scenarios.

C. Comparison Method

To compare the ability of each model with one another, both AIC and BIC are computed for the set of models to gauge which one minimizes their values. Consistency among the modeling frameworks for minimizing both criteria is considered. The metabolic model that best minimizes these information criteria for different scenarios is identified as the best model for the purposes of optimization and control applications. In other modeling scenarios, such as those in which models are developed for the purposes of biological discovery, AIC and BIC should not be used for model selection as quantifying the tradeoff between accuracy and simplicity is not necessarily relevant.

III. RESULTS

A. Modeling Dynamic Fluxes of the Aerobic Growth of E. coli

To compare how well various biological models compress metabolic fluxes, four different models describing the aerobic growth of *E. coli* were compared. In this growth scenario, *E. coli* first consumes extracellular glucose while generating acetate as a byproduct. Once the glucose has been exhausted, the culture then shifts its metabolic state to consume the acetate product. The models chosen to describe various aspects of this system are 1) a detailed kinetic model; 2) dynamic FBA; 3) HCM; and 4) L-HCM.

The reaction network used in this model selection exercise was taken from [19] where the metabolic network was summarized in Fig. 2 which was also used as the kinetic model for this comparison. The network includes a variety of metabolic pathways including glycolysis, gluconeogenesis, and the TCA cycle. Note that some metabolic reactions are truncated with others in order to simplify the network. For example, the network's reaction for G6P's conversion to FBP lumps together two reactions and ignores the intermediate product of F6P.

The kinetic model [19] is taken as the basis for this exercise as this model type is generally the most labor intensive to develop in terms of parameters and structure. Then, dFBA, HCM, and L-HCM formulations were developed for the kinetic model's growth scenarios. The structured biomass equation used to develop these models was extrapolated from a prior *E. coli* model [20] and takes the form of

$$\begin{aligned} &3.7478\ \mathrm{AcCoA} + 5.1971\ \mathrm{alKG} + 1.32915\ \mathrm{G6P} \\ &\quad + 18.271\ \mathrm{NADPH} + 1.787\ \mathrm{OxA} + 0.5191\ \mathrm{PEP} \\ &\quad + 1.496\ \mathrm{PG3} + 2.8328\ \mathrm{Pyr} + 59.89551887\ \mathrm{ATP} \\ &\qquad \rightarrow 1\ \mathrm{Biomass} + 3.547\ \mathrm{NADH}. \end{aligned} \tag{16}$$

1) Steady State Flux Predictions: Each of the models tested generates a description of this network's metabolic fluxes for both dynamic scenarios and steady state scenarios. Each one of these model descriptions of fluxes can then be verified using steady state flux data for growth on glucose and acetate from [21]. Fig. 3 shows the correlation plots describing the accuracy of each model's steady state flux description with the appropriate Pearson's

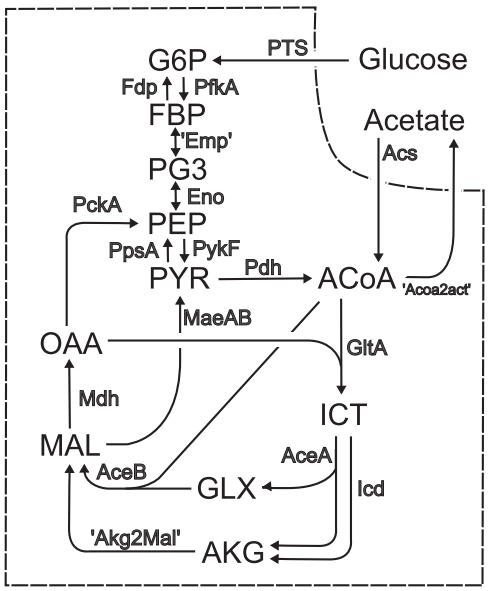

Fig. 2. *Schematic of the simplified network used to construct the various models. Reaction names used in subsequent plots are listed next to their respective arrows.*

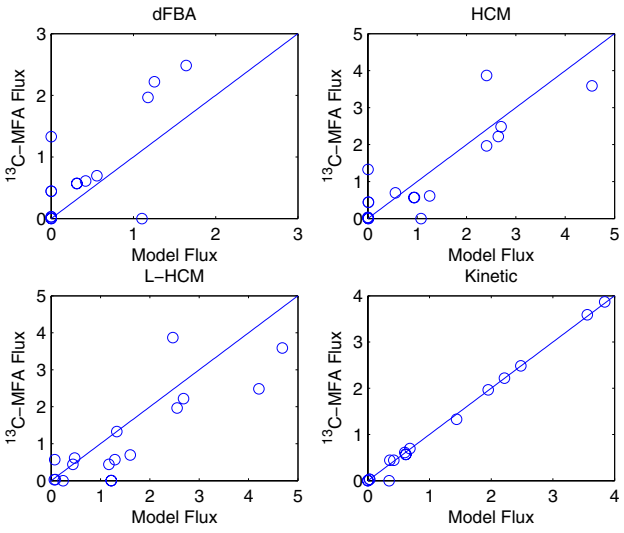

Fig. 3. *Correlation plots for each model's description of steady state flux data for E. coli growing on glucose. The horizontal axis shows the flux values predicted by each model and the vertical axis corresponds to the experimental values for steady state fluxes taken using carbon-13 labeling experiments.*

Table 1 Correlations, Parameterization and Information Criterion for the Model Set for the Steady State Fluxes

	Kinetic	FBA	HCM	L-HCM
Parameter No.	193	3	12	11
$\rho_{Glucose}$	0.9972	0.8172	0.8269	0.8548
$AIC_{Glucose}$	113.38	**1.1445**	4.0746	3.6997
$BIC_{Glucose}$	193.78	**2.3943**	9.0739	8.2823
$\rho_{Acetate}$	0.9980	0.3290	0.3307	0.7858
$AIC_{Acetate}$	186.63	14.411	18.635	**0.3933**
$BIC_{Acetate}$	267.03	15.661	23.634	**4.9760**

product-moment correlation coefficient. Note that FBA is used to calculate the steady state fluxes instead of dFBA and that the kinetic model was parameterized, in part upon the flux data.

Fig. 3 clearly shows the kinetic model's superior ability to describe the rates of metabolic reactions. Nonetheless, it is evident that all models provide reasonable descriptions of metabolic fluxes. The Pearson metric is listed in Table 1 for each condition. The model with the most parameters, being the kinetic model with 193 experimental constants, produced the most accurate approximation of the experimental flux values for both growth on glucose and acetate. The next most accurate experimental flux description is produced by L-HCM. Despite ranking third and fourth, HCM and FBA also provide good approximations of the experimental fluxes of growth on glucose. Both HCM and dFBA, however, show low accuracy in generating steady state approximations of the acetate fluxes.

Model comparison metrics were calculated for both the glucose and acetate conditions to gauge how well each model performed. Parameters for the models were tabulated. L-HCM was the second most parameterized model with 14 parameters. These parameters include the kinetic parameters as well as those used to lump the EMs together. HCM had 12 parameters. FBA had only three parameters which were the uptake and excretion rates of substrates and products in this system. Note that the objective functions in these different models were not parameters, but structures for the various models which did not imply additional penalty for model complexity.

The FBA model for the glucose steady state fluxes showed the best minimization of both information criteria. FBA demonstrated poor performance in predicting the fluxes for the steady state growth on acetate and was outperformed by L-HCM for both AIC and BIC. Despite the kinetic model's very high correlation with experimental flux data for both conditions, it was severely penalized by its large number of parameters.

2) Dynamic Flux Predictions: Metabolic fluxes are difficult to measure experimentally and require carefully controlled experiments. Because of this, time-series experimental fluxes are unavailable for this system. To compare the dFBA, HCM, and L-HCM models on their ability to model dynamic fluxes, the kinetic model's prediction of dynamic fluxes will be treated as an experimental approximation of the true dynamic metabolic fluxes for this system. The use of this artificial data is justified in part by the fact that the model was parameterized upon a wide variety of data for multiple levels of cellular processes. It incorporates a great span of regulatory phenomena including transcription factors, transcription-factor metabolite interactions, gene expression, enzyme production, kinase reactions, phosphatase reactions, and protein degradation. Given the high degree of complexity of the kinetic model, it will be reasonable to assume that it provides a close to true approximation of the real dynamic fluxes for the *E. coli* system that is being modeled.

Fig. 4 shows the comparison of the information theoretic metrics with the artificial data. Generally speaking, HCM and L-HCM overpredict the flux rates through the glyoxylate pathway. They also overpredict the flux of malate to pyruvate during growth on both glucose and acetate. On the other hand, dFBA underpredicts the fluxes through glycolysis and the TCA cycle while L-HCM and HCM provide good qualitative descriptions of the simulated data. L-HCM overpredicts the futile cycling from G6P to FBP during the consumption of glucose while HCM and dFBA underpredict this. HCM best predicts the flux through pyruvate dehydrogenase while dFBA is lower and L-HCM is higher.

As in the steady-state flux model comparison, model selection metrics were calculated for the set of models. The complexities of the HCM and L-HCM models did not change. However, to incorporate the dynamics of the system, five additional parameters were added to generate the dFBA description of the data. The artificial data were generated at 15-min intervals for a 10-h period of growth. The model that best minimizes the information criterion for the dynamic flux data is HCM which is shown in Table 2. This is followed by L-HCM. dFBA places last for both information criteria.

IV. DISCUSSION

The results provided by the model selection framework presented in this work varied by application. In the comparison of the different models' abilities to predict steady state flux data, there was a mixed outcome. While FBA was highly capable of describing the steady state fluxes for the growth on glucose, it was less able to provide an accurate description of the fluxes for the growth on acetate. L-HCM, conversely, placed third in its minimization of the information criterion for glucose. However, it best minimized them for the acetate fluxes. L-HCM also

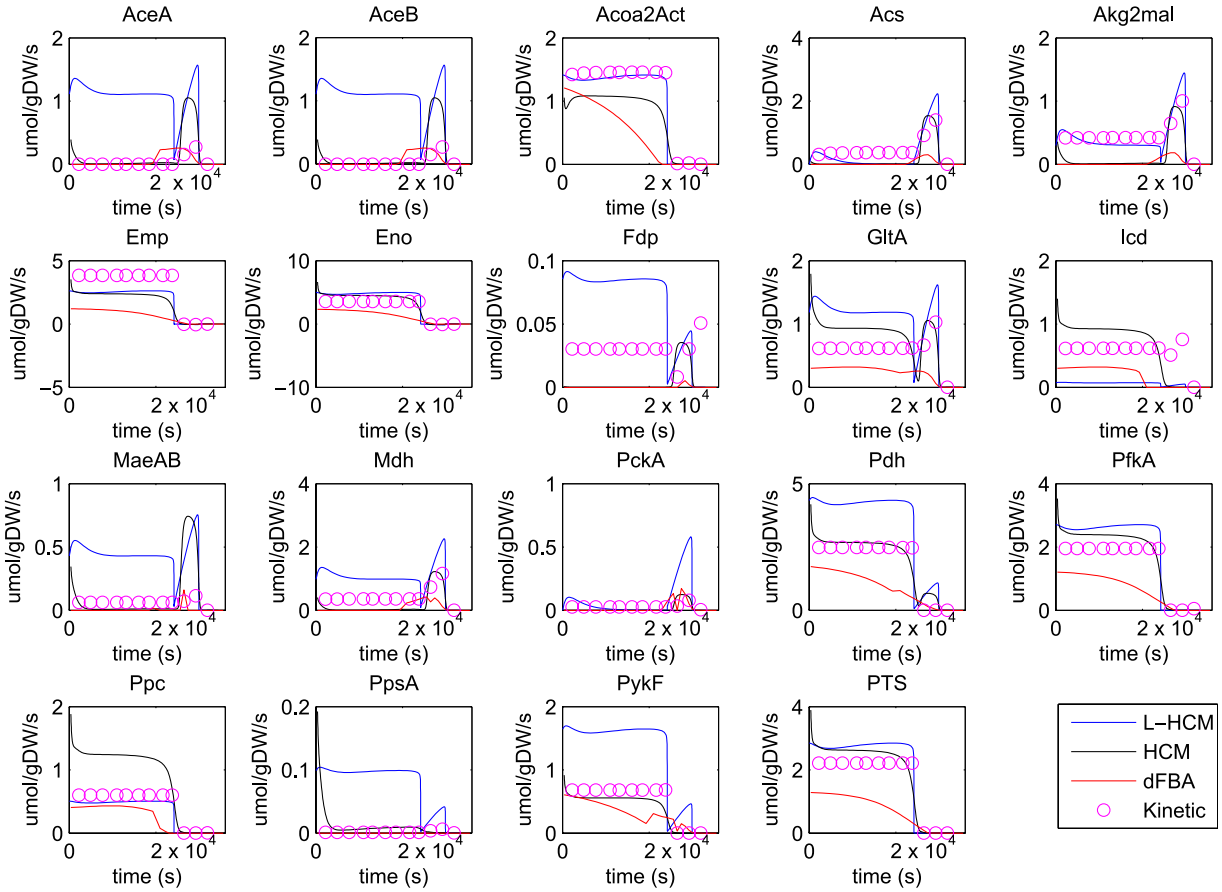

Fig. 4. *Dynamic flux profiles for all three models. The flux titles correspond to the reaction names given in Fig. 2. The purple circles represent the values for the artificial flux data. The red line represents flux predictions made by dFBA. The black and blue lines represent the flux predictions made by HCM and L-HCM, respectively.*

provided the best steady state flux description for the glucose fluxes, after the kinetic model, but its complexity penalized it into the third place. This additional complexity, allowed it to capture the acetate steady state fluxes best. Because of the information criterion's ability to penalize models for their level of complexity, FBA has a clear advantage when it is able to provide an accurate estimate.

The ability of L-HCM to capture steady state fluxes has also been demonstrated in prior work [22]. The strength of L-HCM is that it combines multiple EMs into a smaller subset of lumped elementary modes. The lumping procedure takes into account experimental data related to the yield of products and biomass for various substrates. The strength of this lumping procedure is clearly seen in the steady state flux results.

Table 2 Information Criteria for Dynamic Flux Descriptions for the Artificial Data

	FBA	HCM	L-HCM
Parameter No.	7	12	11
AIC	3580.8	**-1607.4**	2926.5
BIC	3612.7	**-1552.7**	2976.6

When treating these models as data compressing entities, it is noteworthy how much FBA, HCM, and L-HCM reduce the values of the information criteria relative to the complex, kinetic description of the steady state fluxes. All three models reduce the information criterion values by 2 orders of magnitude representing a significantly reduced description length for the data. It is intuitive that the complex kinetic description of the system would provide the most excessively complex description of the equilibriated flux state for this system. The information criteria are able to take this intuitive statement and establish it in quantitative sense. Moreover, they show the degree to which the kinetic model overexplains steady state data.

The scenarios tested above showed that HCM provides the best description of the artificial dynamic flux data despite the fact that it did not perform well in estimating the steady state fluxes. This is most likely due to the fact that HCM employs a combination of EMs over

time to describe the changes in fluxes. In steady state scenarios, HCM will only express one of these numerous EMs that embodies some extreme edge of the yield space as determined using yield analysis. This comes from the cybernetic policy which will selectively produce only one pathway's bulk enzyme pathway once a steady state is reached because there is only one pathway that will have a maximum unregulated rate for the substrate concentration present at a steady state. Notwithstanding HCM's difficulty in the capture of steady states for both of the substrate conditions modeled here, the inclusion of multiple EMs for a single substrate becomes an advantage for HCM as it can describe the span of an organism's metabolic yield space.

In contrast to HCM, the L-HCM model shows a markedly less accurate prediction of the dynamic fluxes for this system. Contrasting accurate steady state predictions, it is less capable of reproducing the artificial data for dynamic fluxes in this scenario. This might in part be due to L-HCM overpredicting the overall flux through different pathways which stems from how it lumps together the different modes. Also, it reduces the complexity of the model into merely two EMs which may not be sufficient in spanning the total space of allowable fluxes. This could mean that despite its dynamic objective function, it is limited to a more limited space in its description of dynamically altering flux profiles. Regardless, both HCM and L-HCM outperform dFBA in the compression of dynamic flux data. This makes the argument that rate-based objective functions are more descriptive of the data compared to yield-based ones.

On the whole, dFBA provides the least accurate prediction of dynamic metabolic fluxes. This could stem from the inadequacy of a yield-based objective function in describing the dynamic shifts in metabolic states. dFBA consistently underpredicted the values of fluxes which may be due to the biomass maximization objective function. This is due to the formulation of the LP problem where maintenance and other functions related to metabolism are ignored and therefore the total sum of fluxes may be contracted from a realistic state. In other words, the goal of the FBA model is to convert the most substrate into biomass with a given set of constraints and will ignore phenomena such as futile cycling. Other objective functions such as total flux minimization could provide better results but were not tested in this work. It has been shown that the accuracy of objective functions varies by scenario [5]. This would also help to explain the biomass maximization objective function differences in accuracy when describing the steady state scenarios. Also not tested were additional constraints on flux values for the metabolic system.

The artificial data for the fluxes were a reasonable substitution given the lack of actual dynamic data for this model system. It is possible that dynamic flux data may be different from what is indicated by the kinetic model. However, given the fact that the kinetic model incorporates a great number of regulatory phenomena and captures steady state fluxes with near-perfect accuracy, the best possible approximation of the dynamic fluxes is likely.

The fact that the model selection criteria were minimized for each model for a different application makes clear that certain models are more relevant to compress specific types of data. It is natural that FBA would be able to minimize this information criterion for the steady state scenarios given its low complexity. In the acetate steady state case, however, the fact that L-HCM outperforms FBA brings to light the fact that FBA's objective functions are not universally descriptive nor applicable in all substrate consumption scenarios. HCM, with its incorporation of multiple EMs, finds a balance between these two which captures the dynamic states predicted by the kinetic model best.

Also, the varied results by scenario point out the fact that these information criteria are not biased toward one approach. Their ability to compare vastly different formulations for the same metabolic system highlights their utility in systems biological applications. Information criteria treat each model as an alternate description length for the data. A key assumption of these metrics is that the model structure itself is not communicated, only the parameters. This assumption allows these metrics to compare nonnested models.

Finally, both AIC and BIC were minimized for the same model for all three model selection scenarios. This consistency is a good indication of their utility in reaching objective conclusions for model selection.

V. CONCLUSION

This work has shown, for the first time, how information criterion can be used for the comparison of nonnested systems biological models. While these metrics have been well established for many applications, their use has not been brought to the attention of metabolic modelers who could benefit from a deeper understanding of how different models balance between accuracy and complexity. This work has shown that L-HCM provides the most succinct description of steady state fluxes for *E. coli* growing on acetate. It has also demonstrated that FBA optimizes the information criterion for *E. coli* growing on glucose at steady state. Finally, it has shown that HCM minimizes these metrics for the description of artificial dynamic flux data.

These conclusions, however, are contingent upon the objectives of the modeling effort. Application of models for the purposes of optimization and/or control requires compromising model complexity in favor of rapid assessment of predictions. On the other hand, a thorough understanding of the changes in metabolic performance due to engineered perturbations can only come about by

using detailed models without serious compromise of complexity. Our analysis demonstrates quantitatively how different modeling frameworks perform when the model objectives are defined. Thus, conclusions made for one objective will not necessarily carry over to other circumstances. ■

Acknowledgment

The authors would like to thank Prof. M. Raginsky who provided crucial direction in this work. They would also like to thank P. Robles Granda for his help in treating this model selection problem.

REFERENCES

[1] P. D. Grnwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT Press, 2007.

[2] H.-S. Song, F. DeVilbiss, and D. Ramkrishna, "Modeling metabolic systems: The need for dynamics," *Current Opinion Chem. Eng.*, vol. 2, no. 4, pp. 373–382, Nov. 2013.

[3] J. D. Orth, I. Thiele, and B. Palsson, "What is flux balance analysis?" *Nature Biotechnol.*, vol. 28, no. 3, pp. 245–248, Mar. 2010.

[4] M. Malik and A. Abdullah, "A comparative study between flux balance analysis and kinetic model for C. acetobutylicum," in *Proc. 8th Malaysian Softw. Eng. Conf.*, Sep. 2014, pp. 264–267.

[5] R. Schuetz, L. Kuepfer, and U. Sauer, "Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli," *Molecular Syst. Biol.*, vol. 3, no. 1, p. 119, Jan. 2007.

[6] H.-S. Song and D. Ramkrishna, "When is the quasi-steady-state approximation admissible in metabolic modeling? When admissible, what models are desirable?" *Ind. Eng. Chem. Res.*, vol. 48, no. 17, pp. 7976–7985, Sep. 2009.

[7] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[8] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, 1998.

[9] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, 1974.

[10] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[11] K. P. Burnham and D. Anderson, "Model selection and multi-model inference," in *A Practical Information-Theoretic Approach*. New York, NY, USA: Springer-Verlag, 2003.

[12] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Stat. Assoc.*, vol. 96, no. 454, pp. 746–774, Jun. 2001.

[13] A. Varma and B. O. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110," *Appl. Environ. Microbiol.*, vol. 60, no. 10, pp. 3724–3731, Oct. 1994.

[14] R. Mahadevan, J. S. Edwards, and F. J. Doyle, III, "Dynamic flux balance analysis of diauxic growth in escherichia coli," *Biophys. J.*, vol. 83, no. 3, pp. 1331–1340, Sep. 2002.

[15] X. Feng, Y. Xu, Y. Chen, and Y. J. Tang, "Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of Shewanella Oneidensis MR-1," *PLoS Comput. Biol.*, vol. 8, 2012, Art. no. e1002376.

[16] J. I. Kim, J. D. Varner, and D. Ramkrishna, "A hybrid model of anaerobic E. coli GJT001: Combination of elementary flux modes and cybernetic variables," *Biotechnol. Progr.*, vol. 24, no. 5, pp. 993–1006, Sep. 2008.

[17] H.-S. Song and D. Ramkrishna, "Reduction of a set of elementary modes using yield analysis," *Biotechnol. Bioeng.*, vol. 102, no. 2, pp. 554–568, Feb. 2009.

[18] H.-S. Song and D. Ramkrishna, "Cybernetic models based on lumped elementary modes accurately predict strain-specific metabolic function," *Biotechnol. Bioeng.*, vol. 108, no. 1, pp. 127–140, 2011.

[19] O. Kotte, J. B. Zaugg, and M. Heinemann, "Bacterial adaptation through distributed sensing of metabolic fluxes," *Molecular Syst. Biol.*, vol. 6, no. 355, 2010, DOI: 10.1038/msb.2010.10.

[20] C. T. Trinh, P. Unrean, and F. Srienc, "Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses," *Appl. Environ. Microbiol.*, vol. 74, pp. 3634–3643, 2008.

[21] M.-K. Oh, L. Rohlin, K. C. Kao, and J. C. Liao, "Global expression profiling of acetate-grown Escherichia coli," *J. Biol. Chem.*, vol. 277, no. 15, pp. 13 175–13 183, Apr. 2002.

[22] H. Song *et al.*, "Dynamic modeling of aerobic growth of Shewanella oneidensis. Predicting triauxic growth, flux distributions, energy requirement for growth," *Metabol. Eng.*, vol. 15, 2012, DOI: 10.1016/j.ymben.2012.08.004.

ABOUT THE AUTHORS

Frank DeVilbiss received the B.S. degree in chemical engineering from the School of Chemical Engineering, Purdue University, West Lafayette, IN, USA, in 2011 and the Ph.D. degree also from Purdue University.

His current research focus is on modeling the dynamic control of metabolism using the concept of metabolic goals. He has developed such models to discover how drugs affect the onset of inflammation and has also devised methods to analyze large sets of biological data to extract these metabolic goals.

Doraiswami Ramkrishna received the B.S. degree in chemical engineering from the Bombay University Department of Chemical Technology, now known as ICT, Bombay, India, in 1960 and the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 1965.

After serving for two years on the faculty at Minnesota, he returned to India to join the IIT Kanpur as an Assistant Professor. In 1974, he went back to the United States as a Visiting Associate Professor at the University of Wisconsin and as a Visiting Professor the following year at the University of Minnesota before joining the Purdue University, West Lafayette, IN, USA, faculty in 1976 as a full Professor. In 1994, he was appointed H.C. Peffer Distinguished Professor. He is noted for his research on the application of mathematics to chemical and biochemical reaction engineering, biotechnology, particulate systems, and more recently personalized medicine. He is well known for the book *Linear Operator Methods in Chemical Engineering* (Englewood Cliffs, NJ, USA: Prentice-Hall, 1985) coauthored with Neal Amundson, and his book *Population Balances. Theory and Application to Particulate Systems in Engineering* (New York, NY, USA: Academic, 2000).

Dr. Ramkrishna is a recipient of several AIChE Awards, the Alpha Chi Sigma (1987), the Richard Wilhelm Award (1998), and the Thomas Baron Award (2004). He is a Fellow of professional societies, the American Institute of Medical and Biological Engineering (1996), and the American Institute of Chemical Engineers (2008). From Germany, he won the Senior Humboldt Award (2001) to visit the Max Planck Institute in Magdeburg. Bombay University honored him with the UDCT Diamond Award (1994), the Platinum Award (2009), and Ruia College with the Jewel of Ruia Award. He has held several Distinguished Professorships and delivered numerous Distinguished Lectures. In 2009, he was elected to the U.S. National Academy of Engineering, and in 2011 as a foreign member to the Indian National Academy of Engineering.