MODELING OF METABOLIC SYSTEMS USING GLOBAL OPTIMIZATION METHODS

[†]Eberhard O. Voit and [‡]Edward P. Gatzke

[†]*Department of Biometry and Eppidemiology*
Medical University of South Carolina
Charelston, SC 29425

[‡]*Department of Chemical Engineering*
University of South Carlona
Columbia, SC 29208
Corresponding author: gatzke@sc.edu

Abstract: This paper considers the ***metabolic engineering*** problem of dynamic modeling in complex biological systems. New areas under consideration include ***dynamic system modeling*** of metabolic systems using a Generalized Mass Action (GMA) representation. The modeling problem will be presented as a nonconvex global optimization problem to be solved using deterministic optimization techniques. Advanced control and estimation methods can be devised based on the input-output model of the nonlinear dynamic system. A five-state fermentation pathway is considered using global optimization techniques for modeling and a discrete-time GMA formulation.

Keywords: Biomedical systems, nonlinear dynamic modeling, model identification, global optimization, sampled-data systems

1. INTRODUCTION

The hallmark of biological systems is their organizational complexity, which is manifested in large numbers of components and multitudes of intricate nonlinear interactions. For instance, in a biochemical system, various metabolites are consumed or created through enzyme-catalyzed reactions. These reactions are often modulated by regulatory components that are produced and consumed by these same reactions in the same pathway or may be constituents of entirely different pathways. When such modulations are present, intuitive analyses by cause-and-effect reasoning are no longer sufficient for system analysis, and systematic mathematical approaches are needed to gain useful insight. These numerical approaches are commonly based on systems of ordinary differential equations.

In metabolic engineering, the analysis of biochemical systems is often directed toward manipulation and optimization. For instance, one goal may be the improvement of metabolic yield in a microorganism. Two structured approaches currently dominate the field. One is a linear analysis of the flux distribution within the system, where the key concept is the well-known stoichiometric matrix. The other approach is a convenient nonlinear representation of the individual reactions. Over the past three decades, several groups around the world have developed and furthered a

mathematical framework specifically dealing with this latter approach.

The basis of this framework, known in the field as Biochemical Systems Theory (BST) is the representation of reaction rates with products of power-law functions that include those and only those metabolites and modulators that directly affect a given rate. See Savageau (1969) and Voit (2000). As an example, if enzyme E catalyzes a bimolecular reaction between A and B, and if this reaction is inhibited by end product P, the power-law term for the rate v in BST may be written as:

$$v = \alpha A^{\gamma_A} B^{\gamma_B} E^{\gamma_E} P^{\gamma_P} \tag{1}$$

where α is the rate constant of the reaction, the concentrations of the biochemical species are A, B, E, and P, and γ_A, γ_B, γ_E, γ_P are apparent kinetic orders.

Under some assumptions which have been discussed extensively in the literature, the nonlinear BST models (in the so-called S-system form) can be effectively optimized. See Voit (1992). However, in the alternative representation of a Generalized Mass Action (GMA) system, which is more intuitive to most biochemists, such optimization falls into the realm of NonLinear Programming (NLP) problems, which are notoriously difficult to handle. Preliminary work by Torres and Voit (2002) indicates that the special power-law structure of GMA systems might be amenable to streamlined, efficient methods of optimization. The development and refinement of such methods is a long-term goal. Achieving this goal would have great reward because GMA systems are the simplest systems that contain both the stoichiometric approach and the S-system approach as immediate special cases. Furthermore, GMA systems contain mixtures of linear and S-systems and have been shown by Savageau and Voit (1987) to provide mathematically equivalent representations for essentially all smooth, nonlinear phenomena. If all GMA rates that determine the dynamics of variable X_i $(i = 1..n)$ are symbolically coded as $\phi_i(X_1, X_2, ...X_n, ...X_m)$, the dynamic response of a GMA system can be modeled as follows:

$$\begin{aligned} \frac{dX_1}{dt} &= \phi_1(X_1, X_2, ...X_n, ...X_m) \\ &\vdots \\ \frac{dX_n}{dt} &= \phi_n(X_1, X_2, ...X_n, ...X_m) \end{aligned} \tag{2}$$

Note that variables X_1 through X_n are time dependent, while variables X_{n+1} to X_m may be independent of time for a given experiment.

One method to modify the rate of change of dependent variables is to over-express a gene. This changes the activity of an enzyme, which is usually modeled as an independent variable in a GMA model. Additionally, other independent variables, such as the substrate concentration or some inhibitor or cofactor, could be manipulated to different degrees, thereby evoking a dynamic response in the system.

In the presented case study, which is adapted from the work of Galazzo and Bailey (1990) and Curto et al. (1995), the external glucose concentration will be manipulated for the system, forcing changes in the dependent variables as glucose is absorbed into the cell at different rates. The metabolic pathway under consideration for this work is shown in Figure 1. Solid arrows represent reactions and dotted arrows show modulations. State variables in the model are: X_1 = cytosolic glucose; X_2 = glucose-6-phosphate; X_3 = fructose-1,6-diphosphate; X_4 phosphoenol pyruvate; X_5 = ATP. Independent variables with constant values are: X_6 = effective hexose transport; X_7 = hexokinase/glucokinase; X_8 = phosphofructokinase; X_9 = glyceraldehyde dehydrogenase; X_{10} = pyruvate kinase; X_{11} = glycogen and trehalose production; X_{12} = glycerol production; X_{13} = ATPase; X_{14} = NADH/NAD+ ratio.

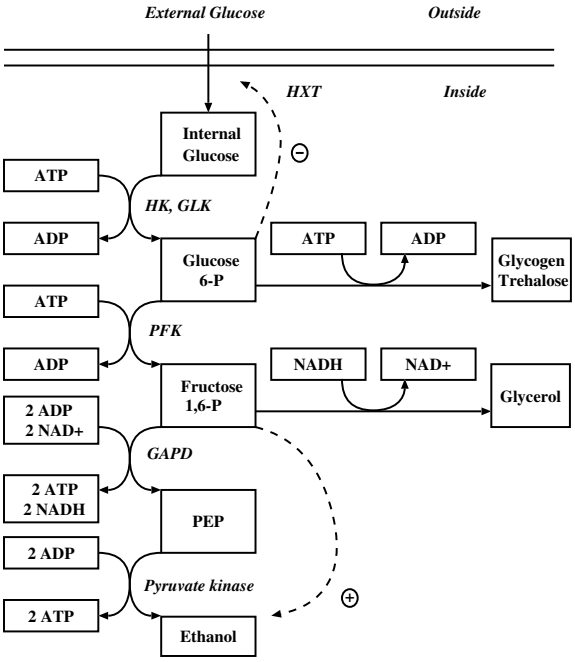

Fig. 1. Simplified model of anaerobic fermentation of glucose to ethanol, glycerol, and polysaccharides in Saccharomyces cerevisiae.

In general, a metabolic network may become quite complex for system involving many species. There may be many uses of such a model. Typically, a fully parameterized model is used for simulation, prediction, or optimizations. However, one might also have available measured concentrations at different points in time and attempt to deduce the structure of the pathway from these "metabolic profiles". See Voit and Almeida (2003). In general, this is a

formidable task, but in the case of a well-structured model such as a S-system or GMA model within BST, the task reduces to the simpler, yet challenging determination of parameter values that best describe the system and the measured profiles. The parameter estimation problem may be formulated as a nonconvex optimization problem to be solved using global optimization techniques. Due to the complexity of metabolic systems, a single set of model parameters not be readily apparent. Deterministic numerical methods may be used to approach this type of problem and determine the best model given the data.

2. MODELING FORMULATION

Given a system at steady state, a small perturbation in a metabolite concentration or in external conditions may cause significant transient response in metabolite concentrations, which provide insight into the structure of the metabolic network and the existence and magnitudes of the fluxes. In the GMA formulation, each flux representation requires determination of values for the rate constant α and the kinetic orders γ_i. The following global optimization scheme, based on system discretization and dynamic programming, can be used to determine the optimal values for these parameters. Here, $\hat{X}_i(k)$ is the metabolic concentration of species i in the model at time k. The rate of change for species i at time k is denoted by $\frac{d\hat{X}_i}{dt}(k)$. P is the number of measurement time points.

$$\begin{aligned}
\min_{\alpha,\gamma} \quad & \sum_{k=1}^{P} |e_i(k)| \ \forall i = 1..n \qquad (3)\\
s.t. \quad & \frac{d\hat{X}_1}{dt}(k) = e_1(k) + \phi_1(\hat{X}_1(k), \hat{X}_2(k), ...\hat{X}_m(k))\\
& \forall k = 1..P\\
& \vdots\\
& \frac{d\hat{X}_1}{dt}(k) = e_n(k) + \phi_n(\hat{X}_1(k), \hat{X}_2(k), ...\hat{X}_m(k))\\
& \forall k = 1..P\\
& \hat{X}_i(k) > 0\\
& \forall i = 1..n, k = 1..P
\end{aligned}$$

This formulation minimizes the total sum of absolute errors for rate of change in the dynamic model representation for the nonlinear system. Note that the concentrations are constrained to take only positive values. Also note that the nonlinear functions p_i in the formulation take the form of a sum of GMA reaction terms, as represented in the general GMA model reaction rate in Equation 1. Assuming measurements of all species are available at all times and that the rate of change of each species can be estimated, the problem dimensionality is greatly reduced. The parametric space becomes only a function of the reaction rate parameters, α and γ_i for each reaction.

The optimization problem can be seen as a nonconvex optimization problem in the general form:

$$\begin{aligned}
\min_{x} \quad & f(x) \qquad (4)\\
s.t. \quad & g_1(x) = 0\\
& \vdots\\
& g_m(x) = 0\\
& LB \le x \le UB
\end{aligned}$$

Here, the functions $f(x)$ and $g_i(x)$ may be nonconvex. In the original formulation of Equation 3 the objective function is a convex function but the model constraint equations are nonconvex nonlinear equality constraints. A nonconvex optimization problem in this form can be solved using standard branch-and-bound techniques. Deterministic branch-and-bound methods similar to those used in Mixed-Integer Linear Programming algorithms can be used to solve nonconvex NLP problems as in Soland (1971), Adjiman et al. (1998), and Tawarmalani and Sahinidis (2000). These methods rely on derivation of a convex relaxed lower bounding problem as described in McCormick (1976), Adjiman et al. (1996), Tawarmalani and Sahinidis (2000), and Gatzke et al. (2002). Recent range-reduction techniques have been shown to play a vital role in rendering more problems tractable as seen in Ryoo and Sahinidis (1995). As in all combinatorial optimization problems, reducing the problem dimension and solution space can lead to large improvements in solution efficiency. In this problem, the actual formulation may contain many variables (variables for α values , γ_i values, and model concentration values $\hat{X}_i(k)$). However, for estimation purposes, the actual solution space is significantly reduces, because branching only applies to α and γ_i.

Given a general nonconvex problem with continuous variables, $x \in \mathbb{R}^n$, any local solution using existing NLP methods will possibly provide an upper bound. The upper bounding problem can be expressed as described in Problem 4, where f and / or g may be nonconvex. After reformulation and introduction of new variables $w \in \mathbb{R}^o$, $z \in \mathbb{R}^{n+o}$. See McCormick (1976), Smith (1996), Tawarmalani and Sahinidis (2000), and Gatzke et al. (2002). An equivalent nonconvex problem can be expressed as:

$$\begin{aligned}
\min_{z} \quad & c_1^T z \qquad (5)\\
& A_1 z \le b_1\\
& h(z) = 0\\
& z^L \le \ z \ \le z^U
\end{aligned}$$

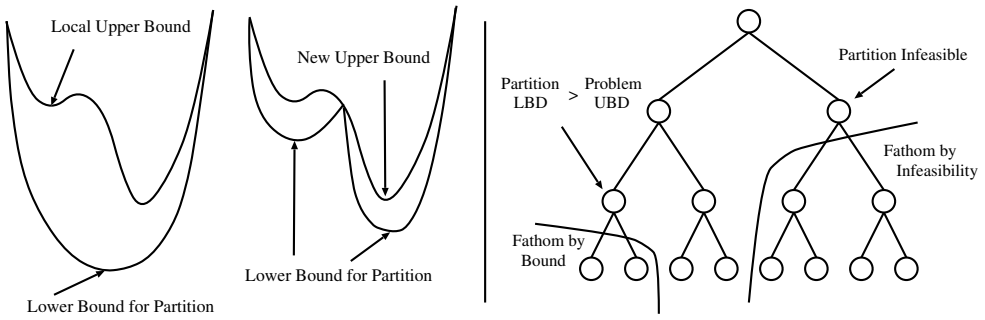

Fig. 2. Left: A single branch-and-bound step for a nonconvex function of a continuous variable. Right: Demonstration of implicit enumeration search for a branch-and-bound tree.

Here, the nonconvex constraints $h(z)$ are simple nonlinear expressions involving two or three variables where one variable is explicitly defined using a single nonlinear operation, *e.g.* $z_1 = z_2 z_3$ or $z_4 = e^{z_5}$. This reformulation is required so that the simple nonconvex expressions can be relaxed using known convex envelopes, and outer approximation of the nonlinear convex expressions described in Tawarmalani and Sahinidis (2000) and Gatzke et al. (2002) leads to convex lower bounding problem for a partition expressed as a Linear Programming (LP) problem:

$$\begin{aligned} \min_z \; & c_2^T z \\ & A_2 z \le b_2 \\ z^L \le \; & z \; \le z^U \end{aligned} \tag{6}$$

In this problem, c_2, A_2, b_2, z^L, and z^U depend on the current variable bounds for the variables in the original problem: x^L and x^U. This lower bounding LP problem can be solved using any valid LP technique.

The deterministic NLP solution proceeds according to the branch-and-bound algorithm illustrated in Figure 2. The original region is partitioned, and lower bounds are determined for each new partition. A partition is discarded if its lower bound exceeds the current upper bound for the problem, or if the lower bounding problem is infeasible. Once a feasible solution to Problem 4 is found, it serves as an upper bound for the global solution. The algorithm attempts to verify that the solution is the true global solution by systematically fathoming the remaining solution space. Range reduction methods can also be used to reduce portions of the solution space, possibly speeding convergence, Ryoo and Sahinidis (1995).

Range reduction techniques play a pivotal role in efficient solutions of nonconvex NLP problems. Reduction methods attempt to shrink the variable space without removing a region that may possibly contain the global solution of the problem. Interval analysis methods of Moore (1979) can be used to analyze the constraints in Problem 5 in order to modify the upper and lower bounds on z, reducing the possible solution space. Solution of Problem 6 provides a lower bound on the solution for a given partition. The Lagrange multipliers at the solution of the convex lower bounding problem can also be used to reduce the solution space, Ryoo and Sahinidis (1995). These bounds-tightening procedures may be repeated for a single partition, producing new variable bounds and a new lower bound for the partition while avoiding branching a partition. This may in some cases avoid the combinatorial growth of active subproblems.

In the formulation described by Problem 3, the only nonconvexity arises from the power-law rate terms of each GMA reaction. Each of these terms can be reformulated by introducing a logarithmic transformation as follows:

$$\begin{aligned} v &= \alpha A^{\gamma_A} B^{\gamma_B} E^{\gamma_E} P^{\gamma_P} \\ \frac{1}{\alpha}\ln(v) &= \gamma_A \ln(A) + \gamma_B \ln(B) + \gamma_E \ln(E) + \gamma_P \ln(P) \end{aligned} \tag{7}$$

Each logarithmic term is then replaced by a new variable, w_i as described by Torres and Voit (2002). After introducing these new variables, almost all constraints in the original formulation are linear. The only nonconvex relationships are the simple constraints of the form $w_i = \gamma_i \ln(X_i)$. As illustrated in Figure 3, X_i^L and X_i^U are known for a given partition and a secant can be used as a convex lower bound for this nonlinear function, while multiple linear first-order approximations may serve as linear upper bounding constrains for the nonlinear function.

3. EXAMPLE SYSTEM

For this system, we consider the fermentation pathway in *Saccharomyces cerevisiae* described in Curto et al. (1995). This is a relatively simple metabolic pathway system with five time-dependent states and, thus, five

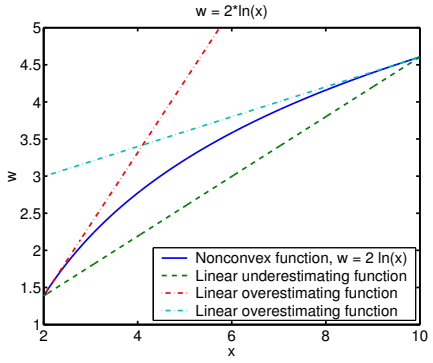

Fig. 3. Convex relaxation using linear constraints.

differential equations. The metabolic pathway map is given in Figure 1. Each reaction is modeled separately in the GMA formulation. For illustration, this GMA model is used as the allegedly "true" model for the generation of data and testing the optimization algorithm. The model equations are given in Figure 4. While this is a nonlinear continuous time dynamic model, it can also be represented with a discrete-time nonlinear model. For this illustration, we use a discrete time sampling rate of 0.001. Slower sampling rates for the discrete time model may result in unstable or inaccurate dynamic systems. This rapid sampling time is an obvious limitation, which is to be considered in future work.

Data from a one hour simulation were used to develop initial parameter estimates for the system. The model system was parameterized, resulting in 22 total parameters to be considered. Using a multistart unconstrained nonlinear optimization algorithm, these parameters were found using Matlab / Simulink, MathWorks (2000), to evaluate the objective function for the system for a given set of parameter values. The error terms for each species were scaled by the expected maximum deviation from the normal steady state operation. The scaling values used were:

$$\begin{bmatrix} 0.0025 & 0.05 & 0.3 & 0.0005 & 0.1 \end{bmatrix}$$

The resulting parameter values serve as an upper bound on the global solution. A comparison of the dynamic response of the continuous time process and the resulting discrete time model is given in Figure 5. The objective function (sum absolute error) at the resulting solution was 9.8479. It is the goal of the global optimization procedure to guarantee that the upper bound value is the global solution.

A branch-and-reduce algorithm was developed to determine optimal parameter values for this system. The lower bounding problems are posed as linear programming relaxations of the convex lower bounding problem for each partition. Each LP problem is solved using OSL from IBM, I. B. M. (1997). The branch-and-reduce procedure includes range reduction techniques that can be used to reduce the total number of nodes visited. The problem formulation was developed using a general purpose Maple script, Maplesoft (2000), that automatically generates code that can be automatically translated using DAEPACK. See Tolsma and Barton (2000) and Gatzke et al. (2002). DAEPACK generates convexification subroutines and gradient information for a given problem.

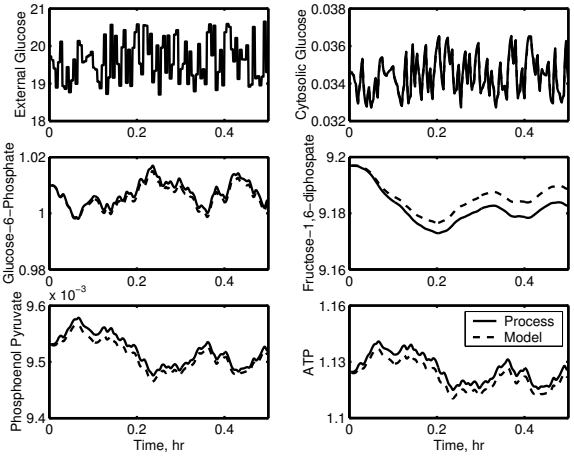

Fig. 5. Comparison of process and dynamic model response for changes in external glucose concentration.

For a given modeling problem with horizon length P with n dependent variables, nP equations can be written. These equations serve as the nonlinear equality constraints for the prediction of the model concentrations over the data set of interest. The model equations include the concentration values of the n species at the P points in the horizon of interest, as well as their instantaneous rate of change. The model equations also include variables representing the GMA reaction rate parameters for each mode flux (α and γ values). The number of new variables and constraints in the resulting convex reformulation will depend on the complexity of each term in the GMA formulation and the total number of terms.

Table 1. Computational Results for different bounds.

Allowable Parameter Perturbation	Solution Time (s)	Global Objective	Local Objective
10%	0.66	0.000	0.000
100%	1.45	0.000	0.000
200%	12.01	0.000	0.917
300%	14.41	0.000	0.884

Given initial bounds on the parameter values, the branch-and-reduce algorithm was able to reduce the

$$
\begin{aligned}
\dot{X}_1 &= 0.8122X_2^{-0.2344}X_6 - 2.8632X_1^{0.7464}X_5^{0.0243}X_7 \\
\dot{X}_2 &= 2.8632X_1^{0.7464}X_5^{0.0243}X_7 - 0.5232X_2^{0.7318}X_5^{-0.3941}X_8 - 0.0009X_2^{8.6107}X_{11} \\
\dot{X}_3 &= 0.5232X_2^{0.7318}X_5^{-0.3941}X_8 - 0.011X_3^{0.6159}X_5^{0.1308}X_9X_{14}^{-0.6088} - 0.04725X_3^{0.05}X_4^{0.533}X_5^{-0.0822}X_{12} \\
\dot{X}_4 &= 0.022X_3^{0.6159}X_5^{0.1308}X_9X_{14}^{-0.6088} - 0.0945X_3^{0.05}X_4^{0.533}X_5^{-0.0822}X_{10} \\
\dot{X}_5 &= 0.022X_3^{0.6159}X_5^{0.1308}X_9X_{14}^{-0.6088} + 0.0945X_3^{0.05}X_4^{0.533}X_5^{-0.0822}X_{10} - 2.8632X_1^{0.7464}X_5^{0.0243}X_7 \\
&\quad -0.0009X_2^{8.6107}X_{11} - 0.5232X_2^{0.7318}X_5^{-0.3941}X_8 - X_5X_{13}
\end{aligned}
$$

Fig. 4. GMA model equations for continuous time system.

initial partition size, tightening the lower bound on the partition. The algorithm only considers parameter value (α and γ) variables for branching. These values were constrained to $\pm 10\%$ of their original values to $\pm 300\%$. The algorithm was able to guarantee within $\varepsilon = 0.1$ that the initially found upper bound was the global solution, while the local solution method was not able to find the global solution in cases with parametric bounds $> \pm 200\%$.

4. CONCLUSIONS

A nonconvex optimization formulation has been presented for determination of metabolic pathway parameters using a GMA representation of the metabolic system. This formulation can then be solved using branch-and-bound methods to global optimality. The branch-and-bound method proceeds in a deterministic manner, providing lower and upper bounds on the quality of the solution as the solution proceeds. The proposed problem can be reduced significantly by only considering a subset of variables for branching. This reduction in problem dimensionality can significantly improve the convergence aspects of the algorithm.

Acknowledgements

The Authors would like to acknowledge financial support from the South Carolina Collaborative Grants Program.

References

C. S. Adjiman, I. P. Androulakis, C. D. Maranas, and C. A. Floudas. A Global Optimization Method, α BB, for Process Design. *Comput. Chem. Eng.*, 20:S419–S424, 1996.

C. S. Adjiman, S. Dalliwig, C. A. Floudas, and A. Neumaier. A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs - I Theoretical Advances. *Comput. Chem. Eng.*, 22(9):1137–1158, 1998.

R. Curto, A. Sorribas, and M. Cascante. Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis. Model definition and nomenclature. *Math. Biosc.*, 130:25–50, 1995.

J. L. Galazzo and J. E. Bailey. Fermentation pathway kinetics and metabolic flux control in suspended and immobilized Saccharomyces cerevisiae. *Enzyme Microb. Technol*, 12:162–172, 1990.

E. P. Gatzke, J. E. Tolsma, and P. I. Barton. Construction of Convex Function Relaxations Using Automated Code Generation Techniques. *Optimization and Engineering*, 3(3): 305–326, 2002.

I. B. M. . IBM Optimization Solutions and Library Linear Programming Solutions. Technical report, I. B. M., 1997.

Maplesoft. *Maple Reference Guide*. Springer Verlag, 2000.

The MathWorks. *Matlab 6.1*. Prentice Hall, 2000.

G. P. McCormick. Computability of Global Solutions to Factorable Nonconvex Programs: Part I - Convex Underestimating Problems. *Mathematical Programming*, 10:147–175, 1976.

R. E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.

H. S. Ryoo and N. V. Sahinidis. Global Optimization of Nonconvex NLPS and MINLPs with Application to Process Design. *Comput. Chem. Eng.*, 19(5):551–566, 1995.

M. A. Savageau. Biochemical Systems Analysis, I. Some Mathematical Properties of the Rate Law for the Component Enzymatic Reactions. *J. Theor. Biol.*, 25(365-369), 1969.

M. A. Savageau and E. O. Voit. Recasting Nonlinear Differential Equations as S-Systems: A Canonical nonlinear Form. *Math. Biosci.*, 87:83–115, 1987.

E. M. B. Smith. *On the Optimal Design of Continuous Processes*. PhD thesis, Imperial College, London, 1996.

R. M. Soland. An Algorithm for Separable Nonconvex Programming Problems II. *Management Science*, 17:759–773, 1971.

M. Tawarmalani and N. V. Sahinidis. Global Optimization of Mixed Integer Nonlinear Programs: A Theoretical and Computational Study. Technical report, University of Illinois, 2000.

J. Tolsma and P. I. Barton. DAEPACK: An Open Modeling Environment for Legacy Models. *Ind. Eng. Chem. Res.*, 39(6): 1826–1839, 2000.

N. V. Torres and E. O. Voit. *Pathway Analysis and Optimization in Metabolic Engineering*. Cambridge University Press, Cambridge, UK, 2002.

E. O. Voit. Optimization in integrated biochemical systems. *Biotechn. Bioengin.*, 40:572–582, 1992.

E. O. Voit. *Computational Analysis of Biochemical Systems*. Cambridge University Press, New York, 2000.

E. O. Voit and J. Almeida. Dynamic Profiling and Cononical Modeling: Powerful Partners in Metabolic Pathway Identification. In R. Goodacre and G. G. Harrigan, editors, *Metabolite Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Kluwer Academic Publishing, Dordrech, The Netherlands, 2003.