Model Identification: A Key Challenge in Computational Systems Biology

Eberhard O. Voit

Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, Georgia

The 2nd International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, 31 October – 3 November 2008
Overview

Systems Biology and Optimization

Choice of a Suitable Model

Bottom-up and Top-down Model Estimation

Technical Issues

Dynamic Flux Estimation

Open Problems
Systems Biology

Biological System → Data → Model

measure → explain → match
Biological System

Data

Model

- measure
- match
- optimize
- extrapolate
- manipulate

explain
Systems Biology

- Biological System
- Data
- Model

- measure
- match
- extrapolate
- manipulate
- optimize

- explain

Optimization
Systems Biology

Biological System

Data

Model

Optimization

Focus today
**Application: Pathway Modeling**

"Local" Data
- Literature, Brenda,
- *de novo* Experiments
  (Enzyme Kinetics)

"Global" Data
- Internet,
- *de novo* Experiments
  (Microarrays,
  Proteomics,
  Mass Spec, NMR,
  Time Series)

Model
- Structure
- "inverse problem"
- "Local Processes"

- Literature, KEGG, *de novo* Experiments

Understanding
- Extrapolation
- Manipulation
- Optimization
Overview of Modeling Process
Formulation of a Dynamical Systems Model

\[ \dot{X}_i = \frac{dX_i}{dt} = V_i^+ - V_i^- \]

\[ V_i^+ = V_i^+ (X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}) \]

complicated

[inside outside]

Big Problem: Where do we get functions from?
<table>
<thead>
<tr>
<th>Sources of Functions for Complex Systems Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physics:</strong> Functions come from theory</td>
</tr>
<tr>
<td><strong>Biology:</strong> No theory available</td>
</tr>
<tr>
<td><strong>Solution 1:</strong> Educated guesses: growth functions</td>
</tr>
<tr>
<td><strong>Solution 2:</strong> “Partial” theory: Enzyme kinetics</td>
</tr>
<tr>
<td><strong>Solution 3:</strong> Generic approximation</td>
</tr>
</tbody>
</table>
Why not Use “True” Functions?

\[ A + B \iff P + Q \]

![Diagram](image)

\[ v = \left( \frac{\text{num.1}}{\text{coef. AB}} \right) (A)(B) - \left( \frac{\text{num.1}}{\text{coef. AB}} \times \frac{\text{num.2}}{\text{num.1}} \right) (P)(Q) \]

\[
= \left( \frac{\text{constant}}{\text{coef. A}} \times \frac{\text{coef. A}}{\text{coef. AB}} \right) + \left( \frac{\text{coef. A}}{\text{coef. AB}} \right) (A)(B) + \left( \frac{\text{coef. B}}{\text{coef. AB}} \right) (B)
\]

\[
+ \left( \frac{\text{coef. AB}}{\text{coef. A}} \times \frac{\text{coef. A}}{\text{coef. AB}} \right) (A)(B) + \left( \frac{\text{coef. AP}}{\text{coef. A}} \times \frac{\text{coef. A}}{\text{coef. AB}} \right) (P)
\]

\[
+ \left( \frac{\text{coef. PQ}}{\text{coef. Q}} \times \frac{\text{constant}}{\text{coef. A}} \times \frac{\text{coef. A}}{\text{coef. AB}} \right) (P)(Q)
\]

\[
+ \left( \frac{\text{coef. ABP}}{\text{coef. AB}} \right) (A)(B)(P)
\]

\[
+ \left( \frac{\text{coef. BPQ}}{\text{coef. BQ}} \times \frac{\text{coef. B}}{\text{coef. AB}} \right) (B)(P)(Q)
\]

from Schultz (1994)
Why not Use Linear Functions?

Example: Heartbeat modeled as stable limit cycle

System of linear differential equations  
System of non-linear differential equations
Formulation of a Nonlinear Model for Complex Systems

**Challenge:**

Linear approximation unsuited

Infinitely many nonlinear functions

**Solution with Potential:**

\[
\dot{X}_i = \frac{dX_i}{dt} = V_i^+ - V_i^-
\]

Savageau (1969): Approximate \( V_i^+ \) and \( V_i^- \) in a logarithmic coordinate system, using Taylor theory.

Result: *Canonical Modeling; Biochemical Systems Theory.*
Example

Adenine Excretion as a Function of Plasma Adenine Concentration

Concentration and Log of Concentration of Plasma Adenine
\[ \dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_{n+m}^{h_{i,n+m}} \]

Each term is represented as a product of power-functions.

Each term contains and only those variables that have a direct effect; others have exponents of 0 and drop out.

\( \alpha \)'s and \( \beta \)'s are rate constants, \( g \)'s and \( h \)'s kinetic orders.

**Important:**
Each term contains exactly those variables that have a direct effect; others have exponents of 0 and drop out.
Mapping Structure  Parameters

\[ \alpha_2 X_1^{g_{21}} \]

\[ g_{41} = 0 \]

\[ g_{41} \leq 0 \]
Alternative Formulations Within BST

S-system Form:

\[ \dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \cdots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \cdots X_{n+m}^{h_{i,n+m}} \]

\[ \dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^+ - \sum V_{ij}^- \]
Alternative Formulations

S-system Form:

\[ \dot{X}_i = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \cdots X_{n+m}^{g_{i,n+m}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \cdots X_{n+m}^{h_{i,n+m}} \]

Generalized Mass Action Form:

\[ \dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^+ - \sum V_{ij}^- \]

\[ \dot{X}_i = \sum \pm \gamma_{ik} \prod X_j^{f_{ijk}} \]
Example of Canonical Model Design

\[
\begin{align*}
\dot{X}_2 &= 8X_1^{0.75} - 5X_2^{0.3} \\
\dot{X}_3 &= 5X_2^{0.3} - 5X_3^{0.5}X_4^{0.2} \\
\dot{X}_4 &= 12X_1^{0.5}X_4^{-1} - 4X_4^{0.8} \\
X_0 &= 1.1 \text{ (constant)}
\end{align*}
\]

GMA: \[
\dot{X}_1 = 20X_0X_3^{-0.9} - 8X_1^{0.75} - 12X_1^{0.5}X_4^{-1}
\]
S-system: \[
\dot{X}_1 = 20X_0X_3^{-0.9} - 19X_1^{0.64}X_4^{-0.45}
\]

\[X_2(t_0) = 1, \quad X_3(t_0) = 0.5, \quad X_4(t_0) = 6, \quad X_1(t_0) = 0.8\]
Example of Canonical Model Design

GMA: \[
\dot{X}_1 = 20X_0 X_3^{-0.9} - 8X_1^{0.75} - 12X_1^{0.5} X_4^{-1} \quad X_1(t_0) = 0.8
\]

S-system: \[
\dot{X}_1 = 20X_0 X_3^{-0.9} - 19X_1^{0.64} X_4^{-0.45} \quad X_1(t_0) = 0.8
\]
**Sphingolipid pathway (purely metabolic)**

1. Many metabolites
2. Many reactions
3. Many stimuli and agents regulate several enzymes of lipid metabolism
4. Some *in vivo* experiments

*Alvarez, Sims, Hannun, Voit*  
*JTB, 2004; Nature, 2005*
Applications

Pathways: purines, glycolysis, citric acid, TCA, red blood cell, trehalose, sphingolipids, ...

Genes: circuitry, regulation,...

Genome: explain expression patterns upon stimulus

Growth, immunology, pharmaceutical science, forestry, ...

Metabolic engineering: optimize yield in microbial pathways

Dynamic labeling analyses possible

Math: recasting, function classification, bifurcation analysis,...

Statistics: S-system representation, S-distribution, trends; applied to seafood safety, marine mammals, health economics
Advantages of Canonical Models

Prescribed model design: Rules for translating diagrams into equations; rules can be automated

Direct interpretability of parameters and other features

One-to-one relationship between parameters and model structure simplifies parameter estimation and model identification

Simplified steady-state computations (for S-systems), including steady-state equations, stability, sensitivities, gains

Simplified optimization under steady-state conditions

Efficient numerical solutions and time-dependent sensitivities

In some sense minimal bias of model choice and minimal model size; easy scalability
$V_i = R_i(S_p, M_i)$

$\frac{dX_i}{dt} = f_k(X_j, V_i)$

$V_i = R_i(S_p, M_i)$

$X_j$
• Lots of time-consuming work and effort!
• Very many a priori assumptions
  • What’s important, what isn’t?
  • Topology
  • Regulation
  • Functional forms
• Seldom consistent experiments
• Mixing and matching of organisms, strains, conditions
• Paucity of data for comparisons with documented responses
• Iterative nature of process time consuming
Alternative to Traditional Modeling: Top-Down Modeling

- Use information at the “global” level (*in vivo* time series data) to deduce (per model) structure and regulation at the “local” level (connectivity, signals, …)
Inverse Problems: Sandbox Example

Voit’s Box of Magic Tricks

VBoMT
Top-Down “Inverse” Modeling

\[ \dot{X} = \alpha \prod X^g - \beta \prod X^h \]
\[ \dot{Y} = \alpha' \prod Y^{g'} - \beta' \prod Y^{h'} \]
\[ \dot{Z} = \alpha'' \prod Z^{g''} - \beta'' \prod Z^{h''} \]

BST
Key Step: Parameter Estimation from Time Series Data

- According to computer scientists: trivial, solved.
- Many methods
- Most work sometimes
- None works always
- Estimation remains to be a frustrating topic!
- Example: Kikuchi et al. 2003
Recent Methods for Parameter Estimation in BST:
~ 100 papers; no method really good
Challenges of Inverse Modeling
Challenges of Inverse Modeling
Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix
Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

Model selection criteria:
- Data dynamics capture ability,
- Mathematical simplicity,
- Tractability,
- Results interpretability

- Infinite variety of formulations
Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

- Model selection criteria: Data dynamics capture ability, mathematical simplicity, tractability, results interpretability
- Infinite variety of formulations

- Computational capacity
- Slow convergence
- Lacking convergence or convergence to local minima
- Time consuming for integration of differential equations
Challenges of Inverse Modeling

- Overly noisy data
- Missing data points
- Uncertainties about the measurements
- Non-informative
- Ill-posed data matrix

- Computational capacity
- Slow convergence
- Lacking convergence or convergence to local minima
- Time consuming for integration of differential equations

- Model selection criteria:
  - Data dynamics capture ability,
  - mathematical simplicity,
  - tractability,
  - results interpretability

- Infinite variety of formulations

- Distinctly different yet equivalent solutions
- Non-equivalent solutions with similar error
- Error compensation
Old Trick: Slope Estimation
(at least as old as Voit & Savageau, 1982)

\[ S(t_k) \approx \dot{X} \mid t_k = f(X(t_k)) \]

\[ S_i(t_j) \approx f_i(X_1(t_j), X_2(t_j), \ldots, X_n(t_j); p_{i1}, \ldots, p_{iM_i}) \]

S-System:

\[ f_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \]

\[ S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \text{ at } t_k \]
Toward a New Trick

\[ S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \text{ at } t_k \]

- Terms become Numbers
- \( \beta_i \) and \( h_{ij} \) are guessed
- \( \alpha_i \) and \( g_{ij} \) are estimated from data
- \( h_{ij} \) are measured
New Trick: Alternating Regression

\[ S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} \quad \text{at} \quad t_k \]

\[ S_i - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}} = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} \quad \text{at} \quad t_k \]

Number \( = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} \quad \text{at} \quad t_k \)

\[ \log(\text{Number}) = \log(\alpha_i) + \sum g_{ij} \log(X_i) \quad \text{for all} \quad t_k \]

Linear regression yields \( \hat{\alpha}_i \) and \( \hat{g}_{ij} \)
Alternating Regression (cont’d)

\[ S_i \approx \alpha_i X_{1}^{g_{i1}} X_{2}^{g_{i2}} \ldots X_{n}^{g_{in}} - \beta_i X_{1}^{h_{i1}} X_{2}^{h_{i2}} \ldots X_{n}^{h_{in}} \text{ at } t_k \]

Use \( \hat{\alpha}_i \) and \( \hat{g}_{ij} \) and compute “\( \alpha \)-term”

Merge the numerical value of the \( \alpha \)-term with \( S_i \) and compute \( \hat{\beta}_i \) and \( \hat{h}_{ij} \) per linear regression for all time points.

Iterate between \( \alpha \) - and \( \beta \) - terms until convergence.
Alternating Regression (cont’d)

Results:
Extremely fast, if it converges.
Convergence issue very complex.
Problems with Traditional Methods

Time to (global) convergence

Problems with collinear data

Problems with models permitting redundancies

Problems with compensation of error among terms
Problems with Traditional Methods: Extrapolation

Former model; here using GMA form

Bad parameters, but good fits because of error compensation

Problem with the “misestimated” system during extrapolation
Example: Regulation of Glycolysis in *Lactococcus lactis*

Bacteria found in yogurt and cheese: *Lactococcus lactis* (top), *Lactobacillus bulgaricus* (blue), *Streptococcus thermophilus* (orange), *Bifidobacterium* spec (magenta).

www.hhmi.org/bulletin/winter2005/images/bacteria5.jpg

Bacterium involved in dairy, wine, bread, pickle production. Relatively simple organization. Here: study glucose regulation.
Goals of Modeling

- Understand pathway; design, operation
- Allow extrapolation to new situations
- Allow prediction for manipulation
- Maximize yield of main product
- Optimize yield of secondary products
- Eventually develop a cell-wide model
Experimental Time Series Data

**Lactococcus Data**

Had modeled these data before

First, difficult to find any solutions

Combination of methods led to good fit

Later, many rather different solutions

Question: Is any of these solutions optimal?

Question: Is the BST model appropriate?

Problems with extrapolation
Dynamic Flux Estimation (DFE)

Inspired by Stoichiometric and Flux Balance Analysis

Extended to dynamic time courses

Study flux balance at each time point

\[ \text{Change in variable @ } t = \text{all influxes @ } t - \text{all effluxes @ } t \]

Linear system; solve as far as possible

Result: values of each flux @ t

Represent fluxes with appropriate models

G. Goel et al., Bioinformatics 2008
Dynamic Flux Estimation (DFE)

Model Free Estimation

- Optimizing and Smoothing
- Numerical Slopes
- System of Fluxes
- Dynamic Flux Profiles
- Linear Algebra

Model Based Estimation

- Parameterized Kinetic Model
- Numerical Flux Representation
- Symbolic Flux Representation
- Parameter Estimation
- Functional Assumptions
- System Topology
- Time Series Data
Dynamic Flux Estimation (DFE)

Diagram (a) shows the metabolic pathways involving glucose, 3PGA, and acetoin. The reaction rates are labeled as $v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12}, v_{13}$.

Diagram (b) represents the dynamic metabolite profiles over time, with graphs for Glucose, PEP, Acetate, FBP, 3PGA, and Lactate.

Diagram (c) depicts the dynamic flux profiles with curves for $v_1, v_2, v_3, v_4, v_5, v_6, v_7$.
Dynamic Flux Estimation (DFE)
Dynamic Flux Estimation (DFE)

Model Free Estimation

Optimizing and Smoothing
Time Series Data
Numerical Slopes
System of Fluxes
Dynamic Flux Profiles
Linear Algebra

Model Based Estimation

System Topology
Parameterized Kinetic Model
Numerical Flux Representation
Symbolic Flux Representation
Parameter Estimation
Functional Assumptions
Open Problems

**Smoothing and Mass conservation:**
Noise in the data leads to loss or gain of mass

**Underdetermined Flux Systems:**
Linear system of flux often not of full rank
Augment DFE with other methods
(e.g., AR or bottom-up estimation)

**Characterization of Redundancies:**
Data collinear or non-informative (pooling?)
Model allows transformation groups (Lie analysis?)
Overriding Challenge

Speed and Convenience

Algorithms for parameter estimation from time series must become much faster and more robust.

They must run reliably and “semi-foolproof” on ordinary PC’s without the need of expensive software.
Efficiently dealing with inverse problems presents new modeling opportunities:

1. Time series data are coming! They contain a lot of implicit information that must be extracted.


3. Important overlooked issue: Error compensation; extrapolation becomes unreliable. DFE promising
Acknowledgements

The Current Crew:

Funding: NIH, NSF, DOE, Woodruff Foundation, Georgia Research Alliance

Information: www.bst.bme.gatech.edu
Further Information

www.bst.bme.gatech.edu