Parameter Estimation and Structure Identification in Metabolic Pathway Systems

Eberhard O.Voit

Georgia Tech

Integrative BioSystems Institute

Atlanta, Georgia

Workshop on
Parameter Estimation for Dynamical Systems
EURANDOM, Eindhoven, The Netherlands, June 8-10, 2009

Overview

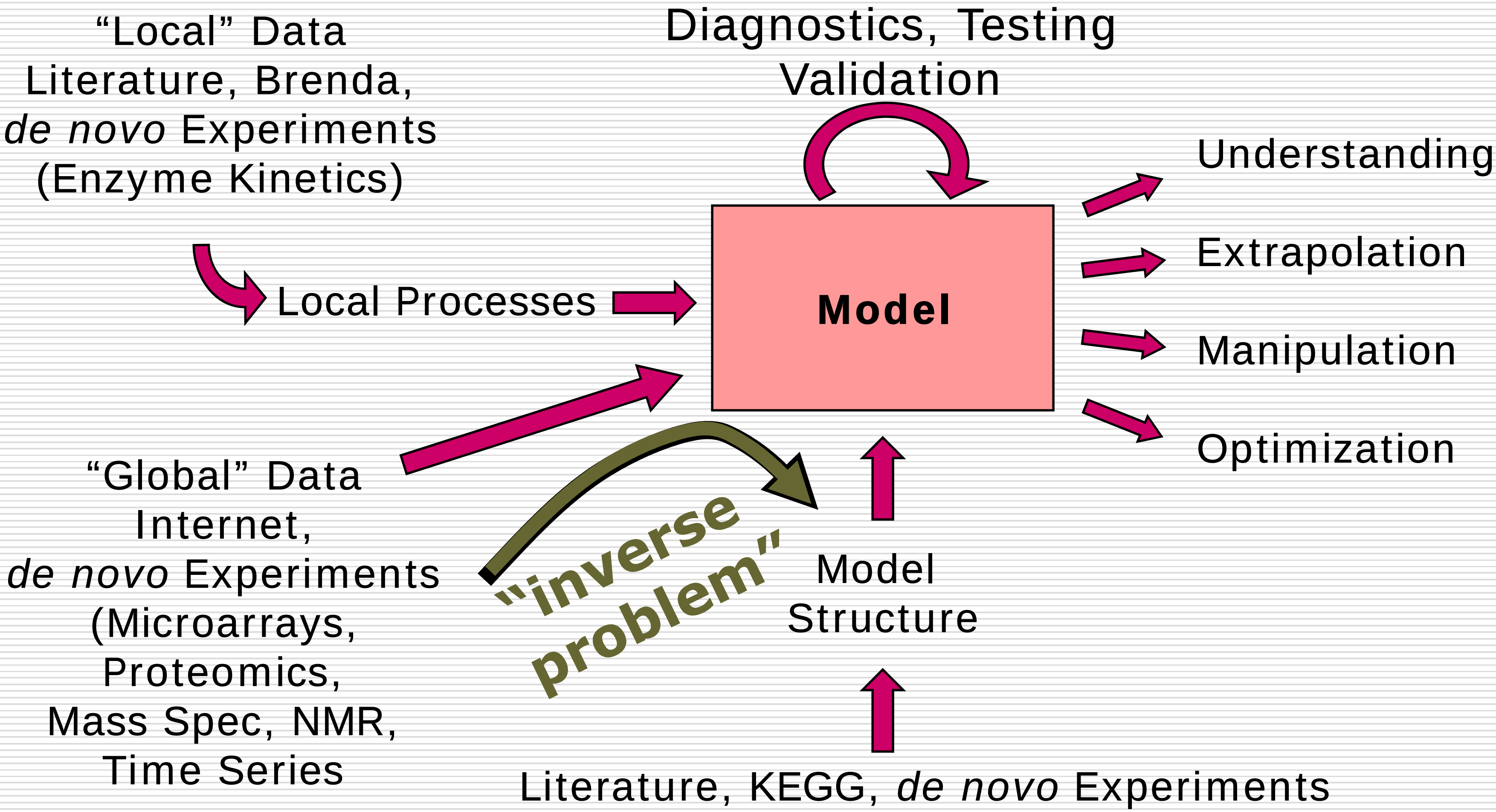
Construction of a Pathway Model

Bottom-up and Top-down Model Estimation

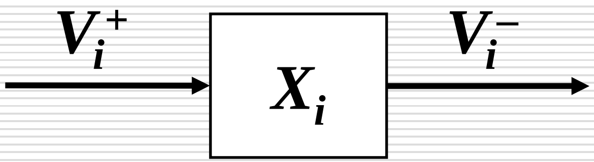
Dynamic Flux Estimation

Open Problems

Construction of a Pathway Model



Formulation of a Dynamical Systems Model



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

$$V_i^+ = V_i^+(\underbrace{X_1, X_2, \ldots, X_n}_{\textit{inside}}, \underbrace{X_{n+1}, \ldots, X_{n+m}}_{\textit{outside}})$$

complicated

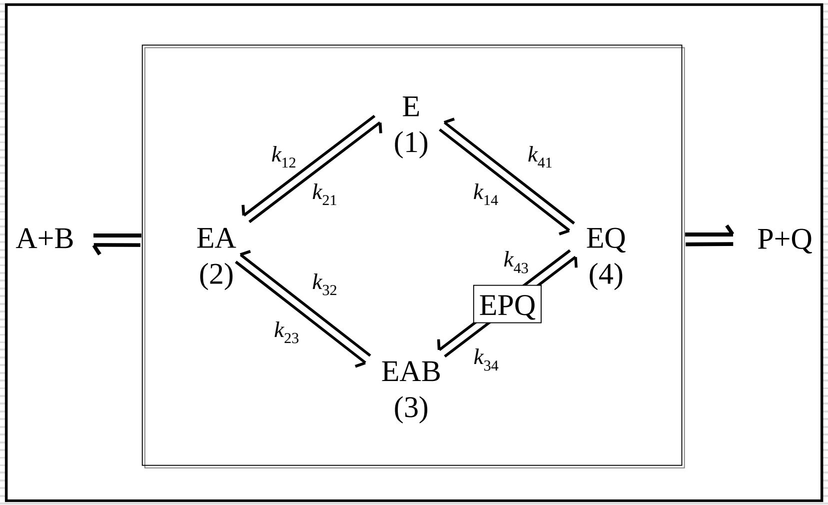
Big Problem: Where do we get functions from?

Sources of Functions for Complex Systems Models

Physics:	Functions come from theory
Biology:	No theory available
Solution 1:	Educated guesses: growth functions
Solution 2:	“Partial” theory: Enzyme kinetics
Solution 3:	Generic approximation

Why not Use “True” Functions?

$$A+B \rightleftharpoons P+Q$$

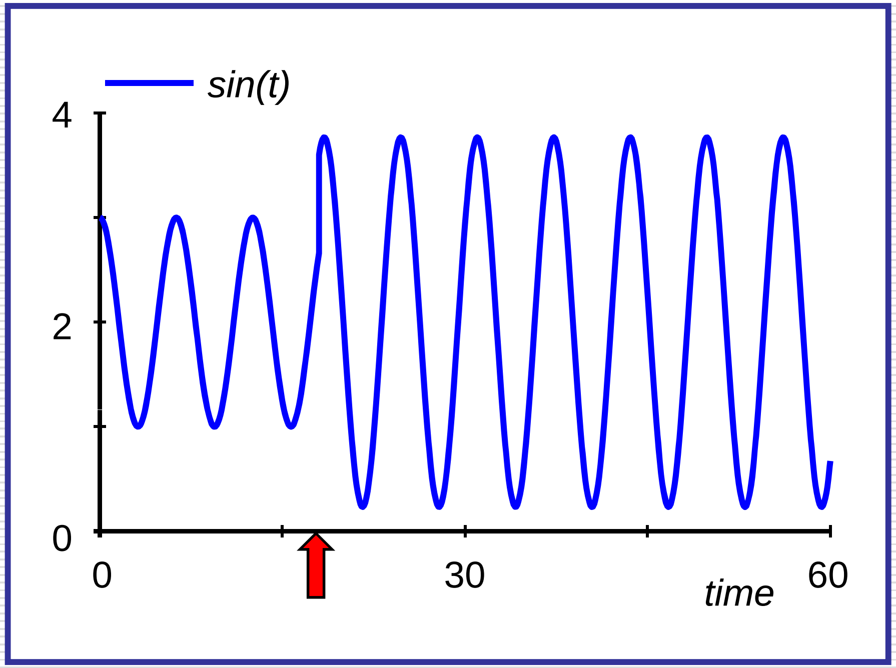


$$v=\frac{\left(\frac{\text{num.1}}{\text{coef. AB}}\right)(A)(B)-\left(\frac{\text{num.1}}{\text{coef. AB}}\times\frac{\text{num.2}}{\text{num.1}}\right)(P)(Q)}{\begin{aligned}&\left(\frac{\text{constant}}{\text{coef. A}}\times\frac{\text{coef. A}}{\text{coef. AB}}\right)+\left(\frac{\text{coef. A}}{\text{coef. AB}}\right)(A)+\left(\frac{\text{coef. B}}{\text{coef. AB}}\right)(B)\\&+\left(\frac{\text{coef. AB}}{\text{coef. AB}}\right)(A)(B)+\left(\frac{\text{coef.P}}{\text{coef. AP}}\times\frac{\text{coef. AP}}{\text{coef. A}}\times\frac{\text{coef. A}}{\text{coef. AB}}\right)(P)\\&+\left(\frac{\text{coef.Q}}{\text{constant}}\times\frac{\text{constant}}{\text{coef. A}}\times\frac{\text{coef. A}}{\text{coef. AB}}\right)(Q)\\&+\left(\frac{\text{coef.AP}}{\text{coef. A}}\times\frac{\text{coef.A}}{\text{coef. AB}}\right)(A)(P)+\left(\frac{\text{coef.BQ}}{\text{coef. B}}\times\frac{\text{coef.B}}{\text{coef. AB}}\right)(B)(Q)\\&+\left(\frac{\text{coef.PQ}}{\text{coef.Q}}\times\frac{\text{coef.Q}}{\text{constant}}\times\frac{\text{constant}}{\text{coef. A}}\times\frac{\text{coef. A}}{\text{coef. AB}}\right)(P)(Q)\\&+\left(\frac{\text{coef.ABP}}{\text{coef. AB}}\right)(A)(B)(P)\\&+\left(\frac{\text{coef.BPQ}}{\text{coef. BQ}}\times\frac{\text{coef.BQ}}{\text{coef. B}}\times\frac{\text{coef.B}}{\text{coef. AB}}\right)(B)(P)(Q)\end{aligned}}$$

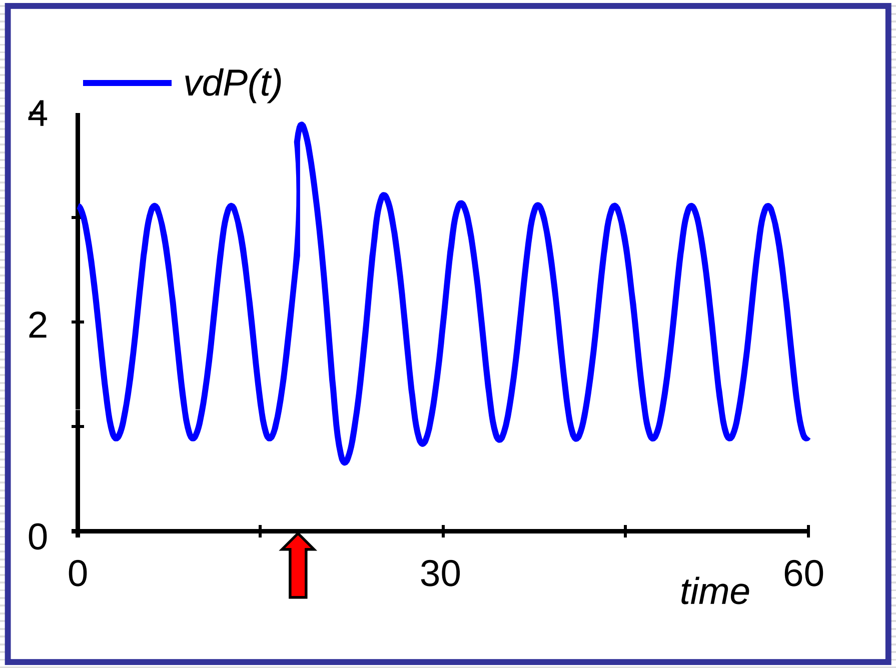
from Schultz (1994)

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*.

Result: S-system

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

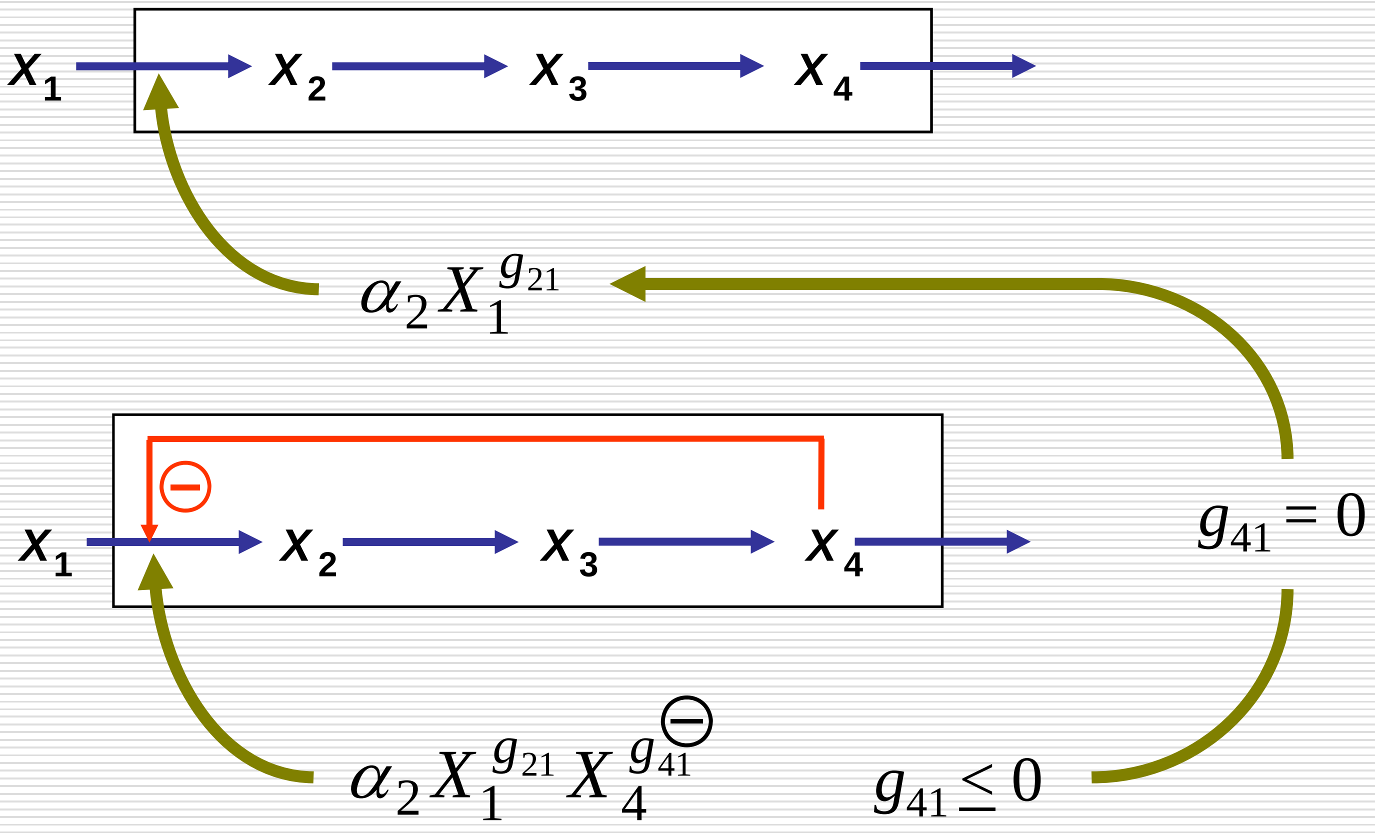
Each term is represented as a product of power-functions.

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

α's and β's are *rate constants*, *g*'s and *h*'s *kinetic orders*.

Important for Estimation & Structure Identification:
Each term contains exactly those variables that have a direct effect; others have exponents of 0 and drop out.

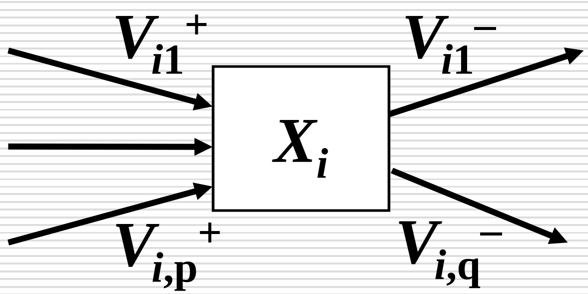
Mapping
Structure ⟺ Parameters



Alternative Formulations Within BST

S-system Form:

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

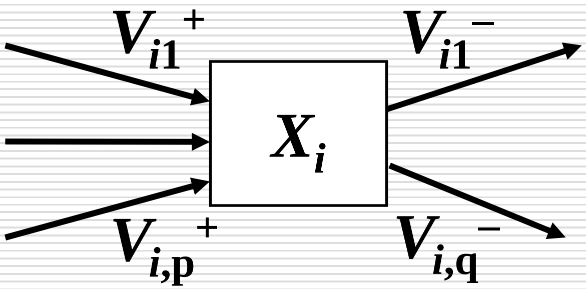


$$\dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^{+} - \sum V_{ij}^{-}$$

Alternative Formulations

S-system Form:

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

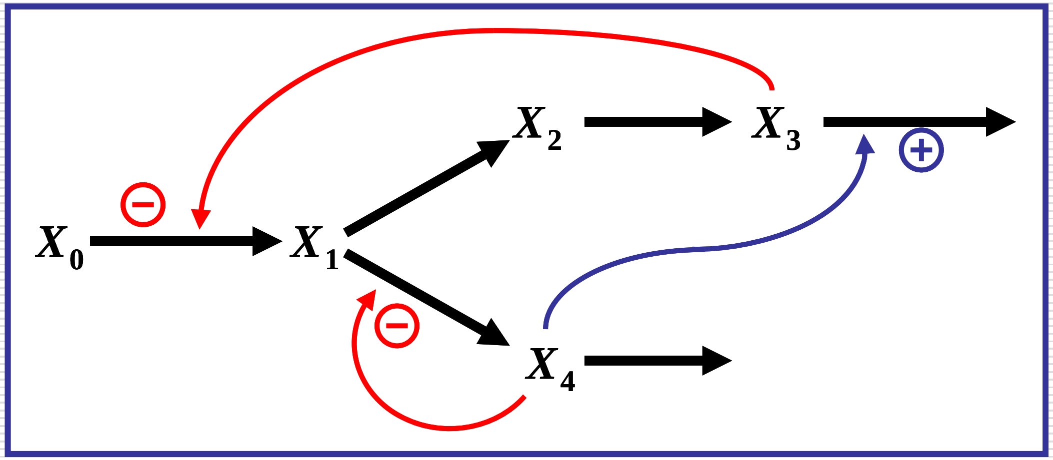


$$\dot{X}_i = \frac{dX_i}{dt} = \sum V_{ij}^{+} - \sum V_{ij}^{-}$$

Generalized Mass Action Form:

$$\dot{X}_i = \sum \pm \gamma_{ik} \prod X_j^{f_{ijk}}$$

Example of Canonical Model Design



GMA / S: $\dot{X}_2 = 8X_1^{0.75} - 5X_2^{0.3}$ $X_2(t_0) = 1$

GMA / S: $\dot{X}_3 = 5X_2^{0.3} - 5X_3^{0.5}X_4^{0.2}$ $X_3(t_0) = 0.5$

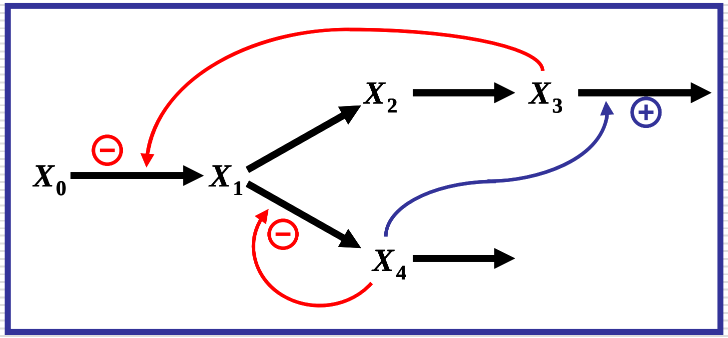
GMA / S: $\dot{X}_4 = 12X_1^{0.5}X_4^{-1} - 4X_4^{0.8}$ $X_4(t_0) = 6$

GMA / S: $X_0 = 1.1$ (constant)

GMA: $\dot{X}_1 = 20X_0X_3^{-0.9} - 8X_1^{0.75} - 12X_1^{0.5}X_4^{-1}$ $X_1(t_0) = 0.8$

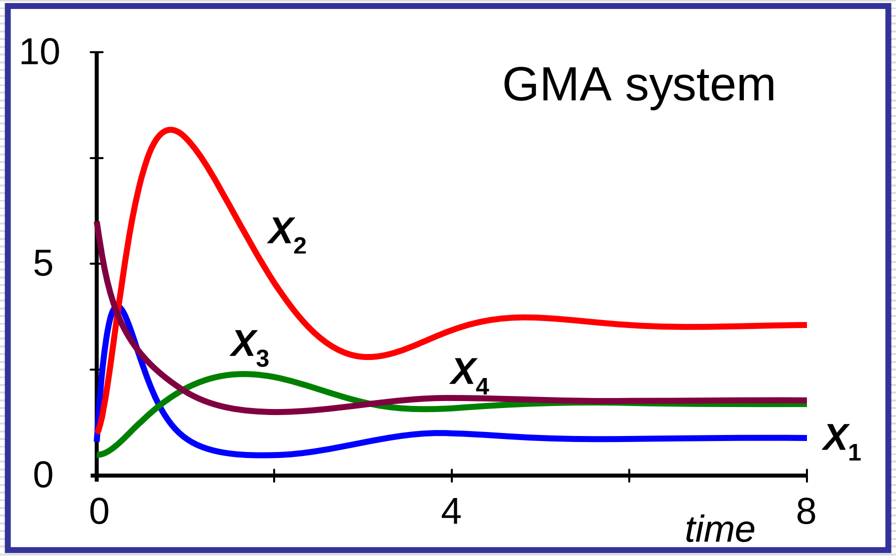
S-system: $\dot{X}_1 = 20X_0X_3^{-0.9} - 19X_1^{0.64}X_4^{-0.45}$ $X_1(t_0) = 0.8$

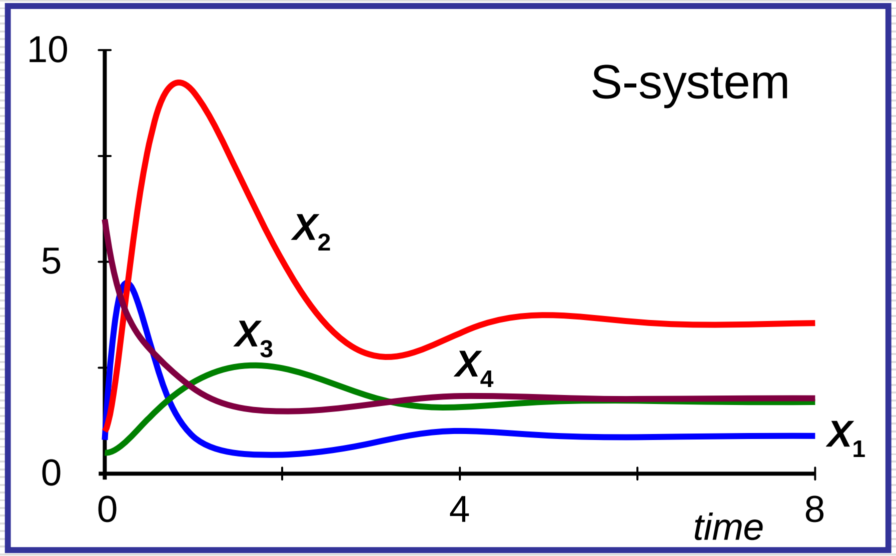
Example of Canonical Model Design



GMA: $\dot{X}_1 = 20X_0X_3^{-0.9} - 8X_1^{0.75} - 12X_1^{0.5}X_4^{-1}$ $X_1(t_0) = 0.8$

S-system: $\dot{X}_1 = 20X_0X_3^{-0.9} - 19X_1^{0.64}X_4^{-0.45}$ $X_1(t_0) = 0.8$





Doable Size

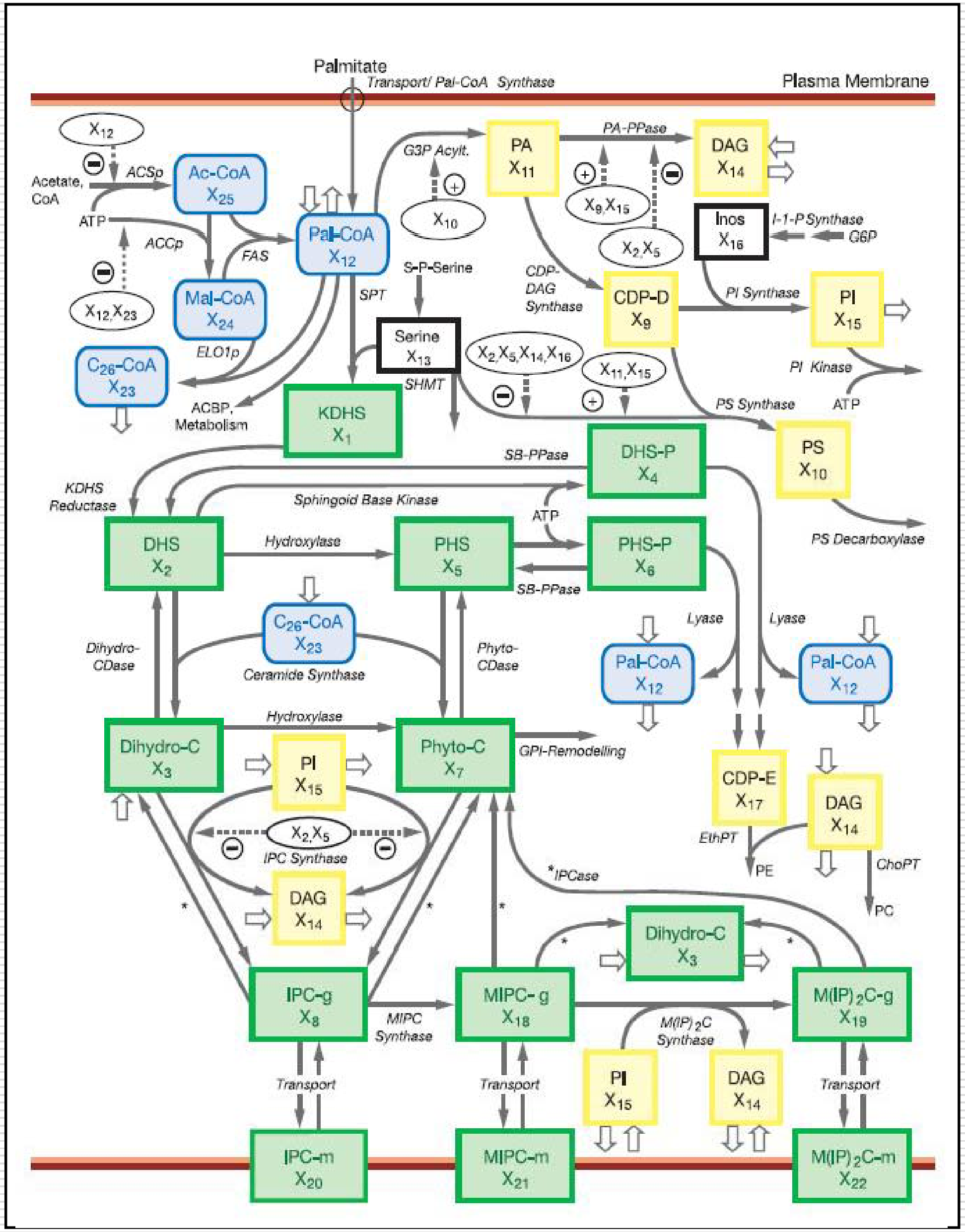
Sphingolipid pathway (purely metabolic)

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, bifurcations, delays...

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; translation 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

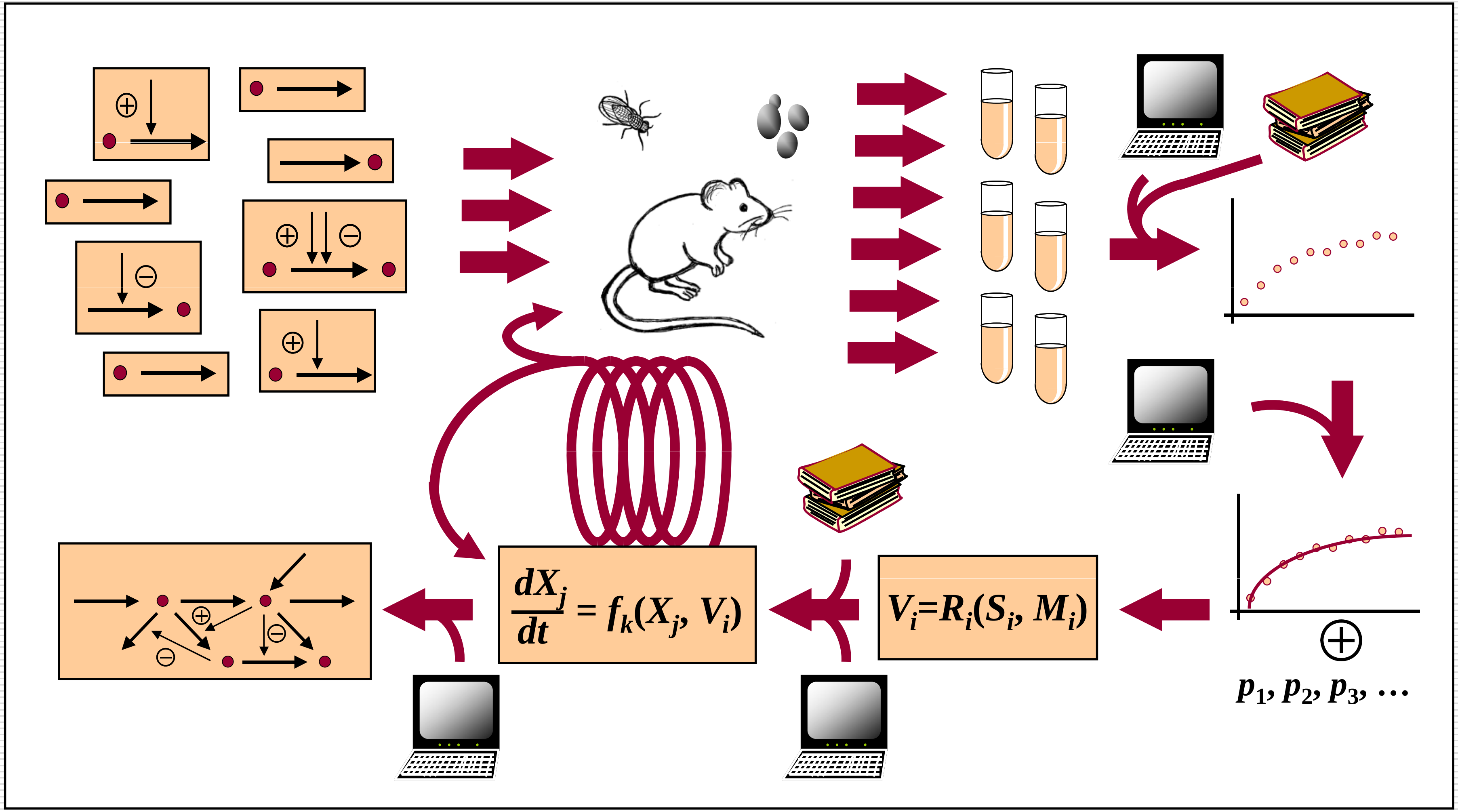
Simplified steady-state computations (for S-systems), including steady-state equations, stability, sensitivities, gains

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

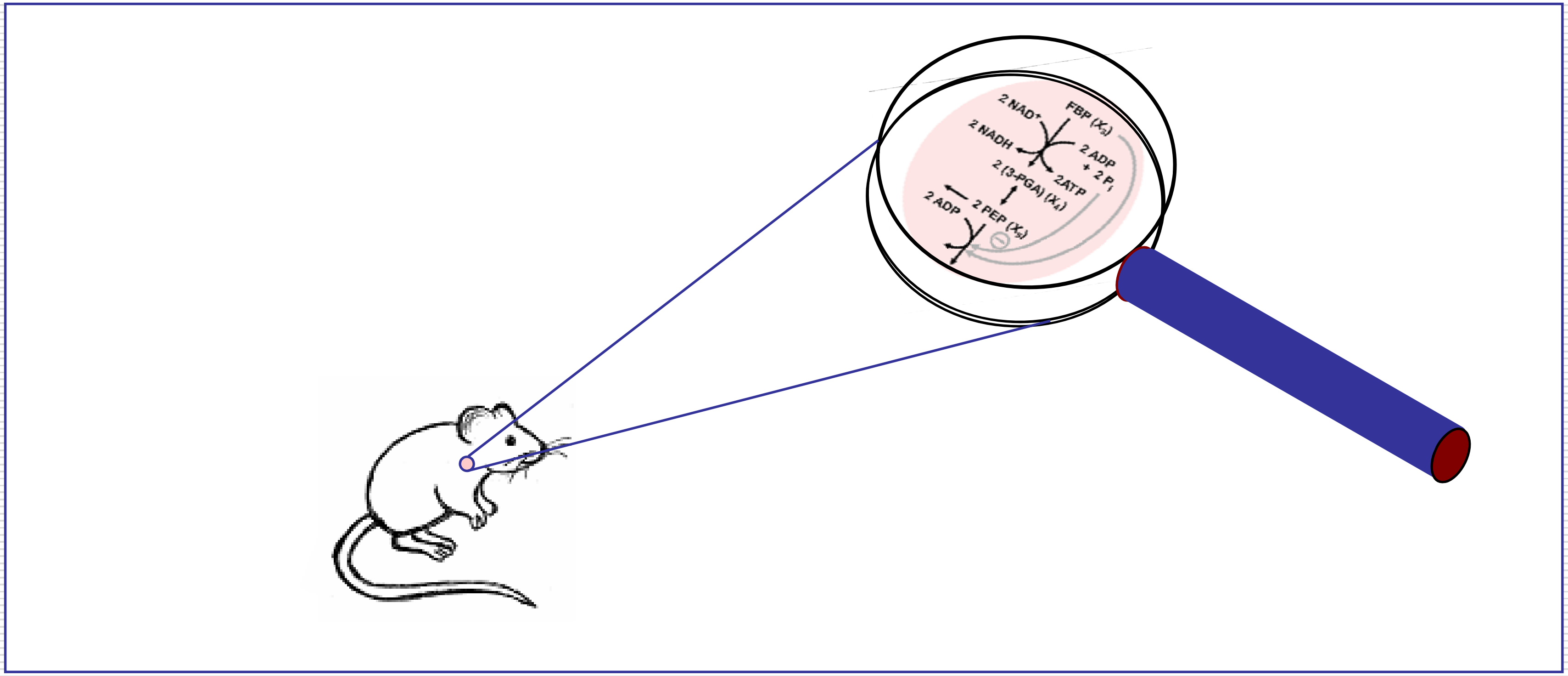
Flow Chart of Traditional Systems Estimation Strategy



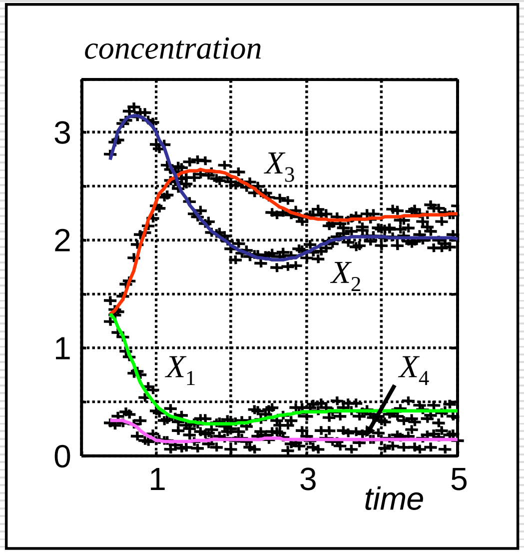
Voit, Drug Discovery Today, 2004

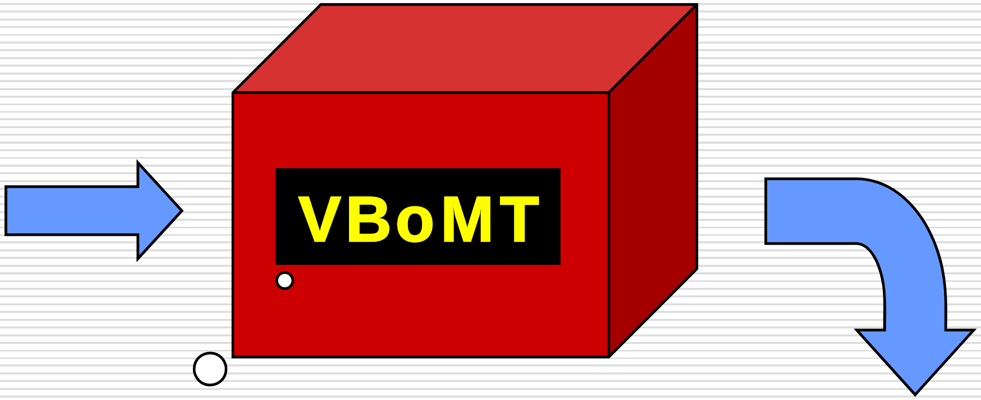
Alternative to Traditional Modeling: Top-Down Modeling

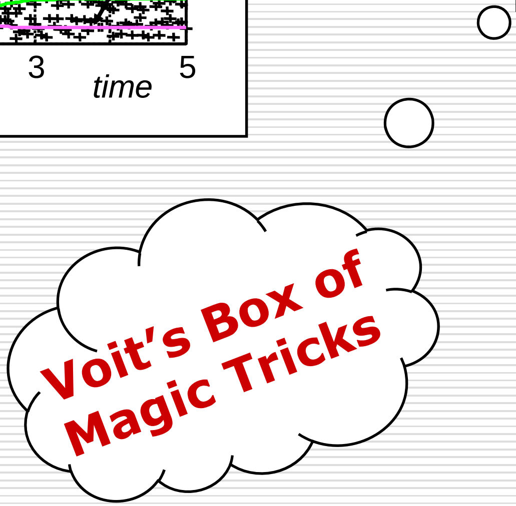
- Use information at the “global” level (*in vivo* time series data) to deduce (per model) structure and regulation at the “local” level (connectivity, signals,…)

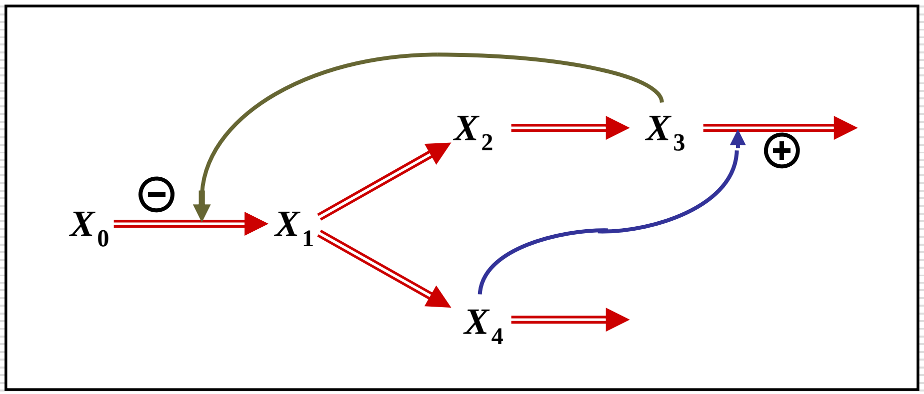


Inverse Problems: Sandbox Example

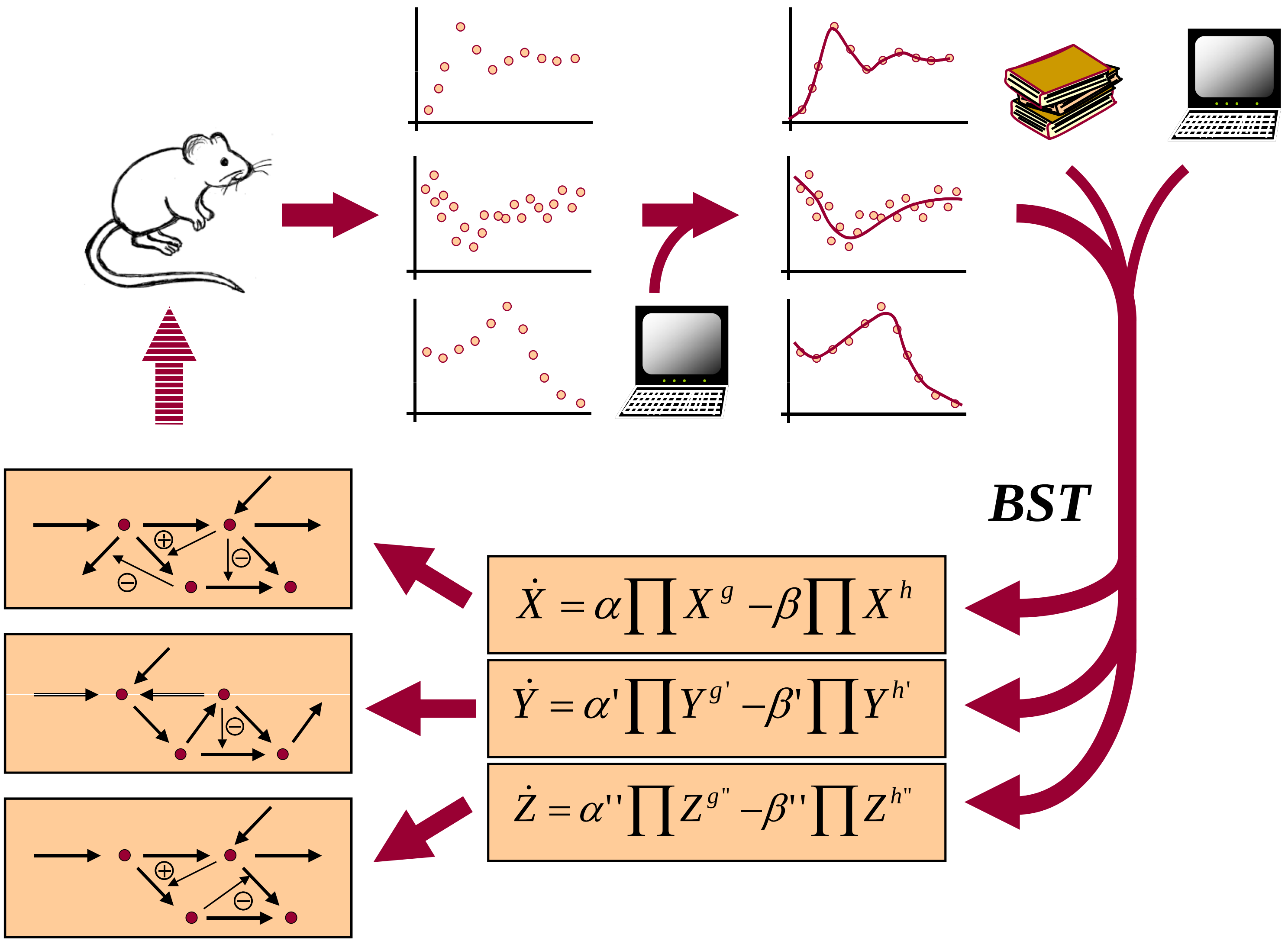








Top-Down "Inverse" Modeling



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 Approaches to Parameter Estimation from Time Series Data

- **Substitution of slopes for differentials; including decoupling of equations (Voit, Savageau, ...)**
- Genetic algorithms (Kikuchi, Tominaga, ...)
- Neural networks + GA's (Almeida, ...)
- Interval methods (Tucker, Moulton, ...)
- Newton flow methods (Tucker, Moulton, ...)
- Simulated annealing (Gonzalez, Mendoza, ...)
- Swarm & ant colony methods (Naval, Mendoza, ...)
- Collocation and hybrid evolution (Tsai, Wang, ...)
- **Alternating regression (Chou, Martens, Voit, ...)**
- Eigenvector optimization (Vilela, Almeida, ...)
- **Dynamic Flux Estimation (Goel, Chou, Voit, ...)**

Chou and Voit, Math. Biosc., 2009

Old Trick: Slope Estimation

$$S(t_k) \approx \dot{X}\Big|_{t_k} = f(X(t_k))$$

$$\vdots$$

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

$$\vdots$$

S-System: $f_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \ldots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \ldots X_n^{h_{in}}$

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

Voit and Savageau, J. Ferment. Techn. 1982

Toward a New Trick

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

estimated from data (smoothing)

measured

Terms become Numbers

→ Guess β_i and h_{ij}

New Trick: Alternating Regression

$$S_i \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \dots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \dots X_n^{h_{in}} \qquad at \quad t_k$$

$$S_i - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \dots X_n^{h_{in}} = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \dots X_n^{g_{in}} \qquad at \quad t_k$$

$$Number = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \dots X_n^{g_{in}} \qquad at \quad t_k$$

$$\log(Number) = \log(\alpha_i) + \sum g_{ij} \log(X_i) \qquad for \quad all \quad t_k$$

Linear regression yields $\hat{\alpha}_i$ and $\hat{g}_{ij}$

Chou, Marten, Voit, Theor. Biol. Med. Model. 2006

Alternating Regression (cont'd)

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

Use $\hat{\alpha}_i$ and $\hat{g}_{ij}$ and compute "α-term"

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

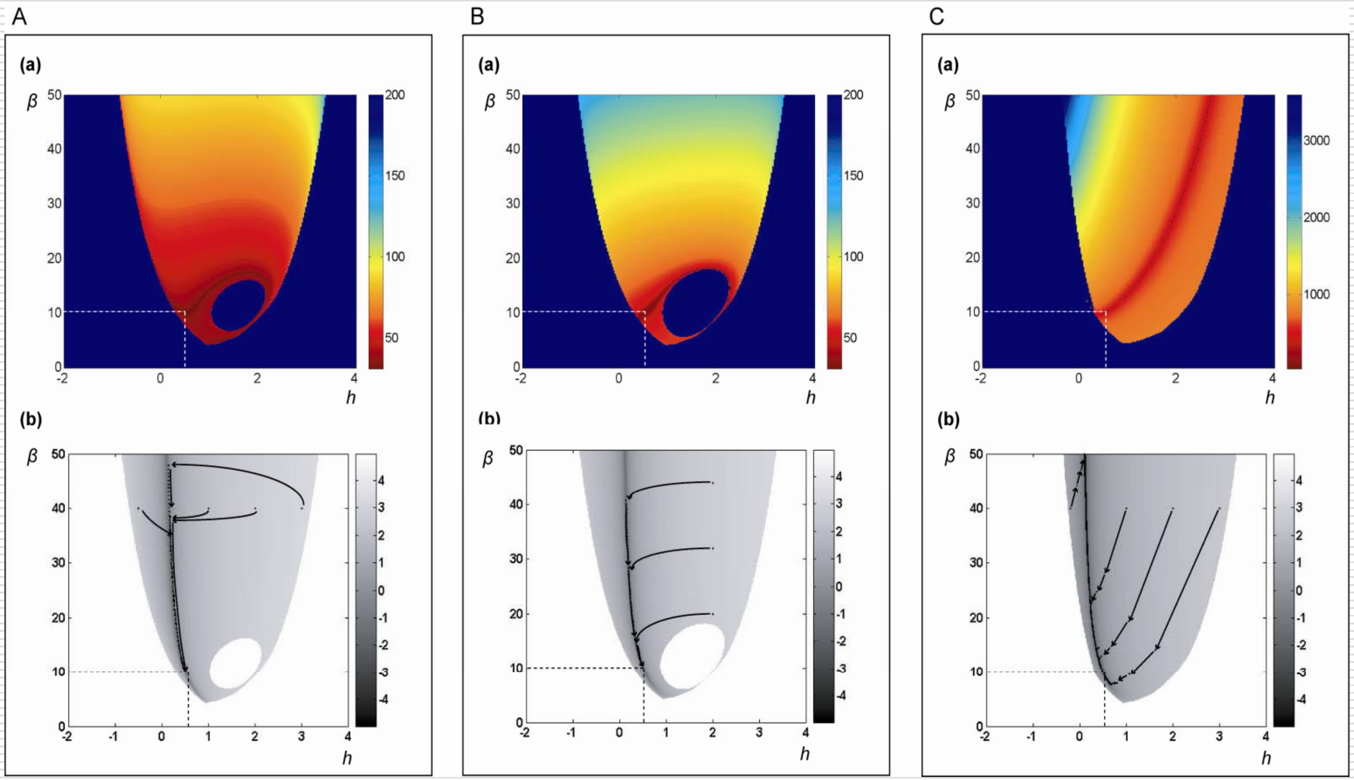
Iterate between α - and β - terms until convergence

Alternating Regression (cont'd)

Results:

Extremely fast, if it converges.

Convergence issue very complex.



Chaotic Lotka-Volterra Model (Vano, …, Sprott, 2006)

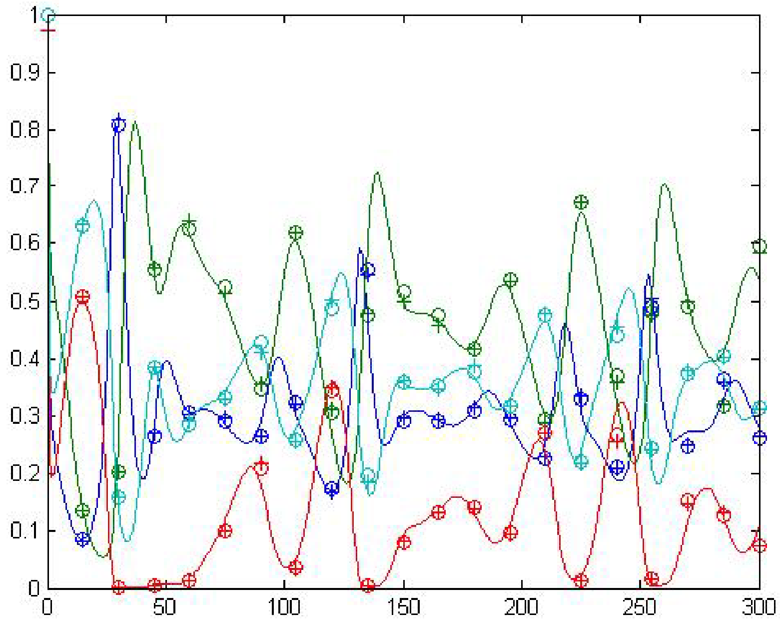
$$\frac{\dot{X}_1}{X_1} = r_1 \cdot (1 - a_{11} \cdot X_1 - a_{12} \cdot X_2 - a_{13} \cdot X_3 - a_{14} \cdot X_4)$$

$$\frac{\dot{X}_2}{X_2} = r_2 \cdot (1 - a_{21} \cdot X_1 - a_{22} \cdot X_2 - a_{23} \cdot X_3 - a_{24} \cdot X_4)$$

$$\frac{\dot{X}_3}{X_3} = r_3 \cdot (1 - a_{31} \cdot X_1 - a_{32} \cdot X_2 - a_{33} \cdot X_3 - a_{34} \cdot X_4)$$

$$\frac{\dot{X}_4}{X_4} = r_4 \cdot (1 - a_{41} \cdot X_1 - a_{42} \cdot X_2 - a_{43} \cdot X_3 - a_{44} \cdot X_4)$$

r_i = (1, 0.72, 1.53, 1.27)
a_{ij} = (1, 1.09, 1.52, 0; 0, 1, 0.44, 1.36;
2.33, 0, 1, 0.47; 1.21, 0.51, 0.35, 1)



Voit, Chou, Int. J. Syst. Synth. Biol. 2009

Typical Problems with Most Methods

Time to (global) convergence

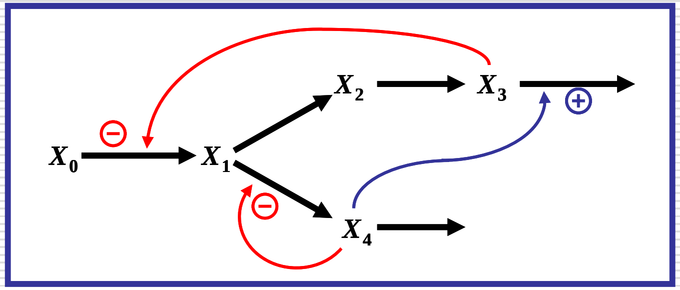
Problems with collinear data

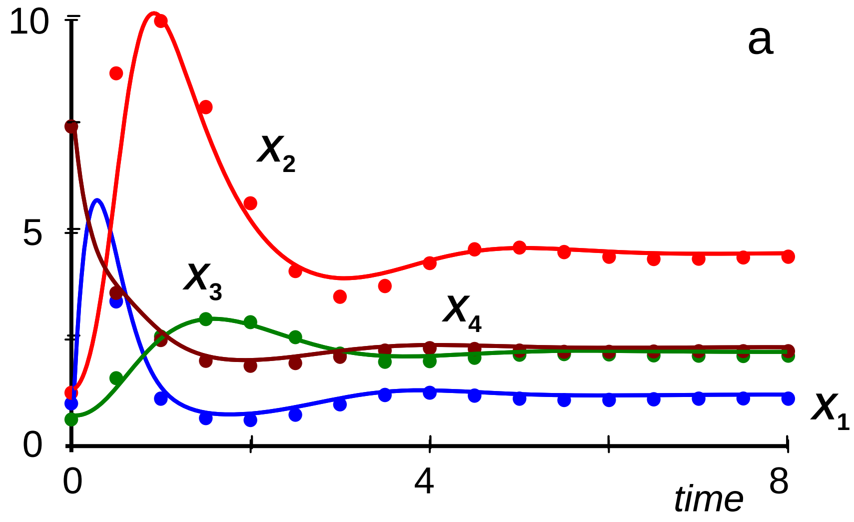
Problems with models permitting redundancies

Problems with compensation of error among terms

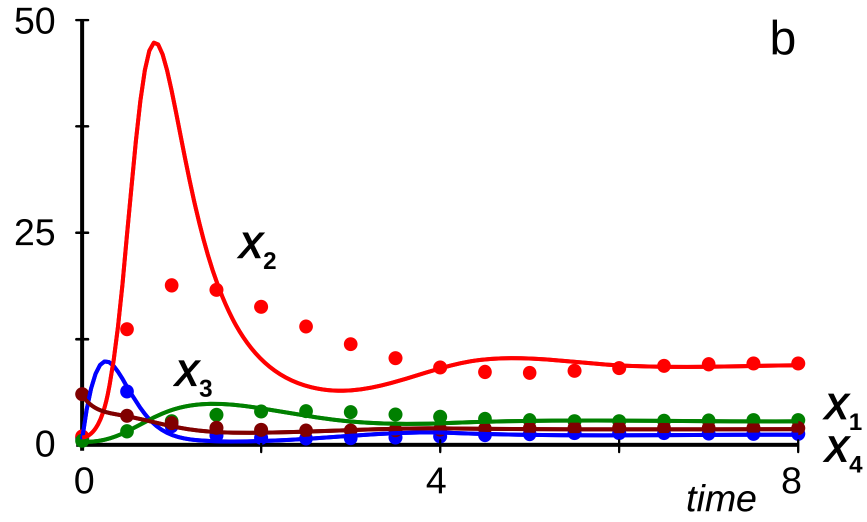
Problem with Traditional Methods: Extrapolation

Former S-system model;
fit with GMA form





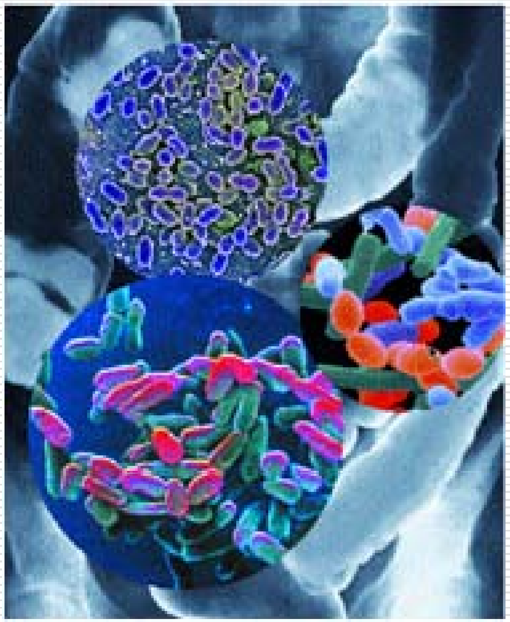
Bad parameters, but good fits because of error compensation



Problem with the "misestimated" system during extrapolation ($2X_0$)

Voit, Lecture Notes in Operations Research 9, 2008

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).

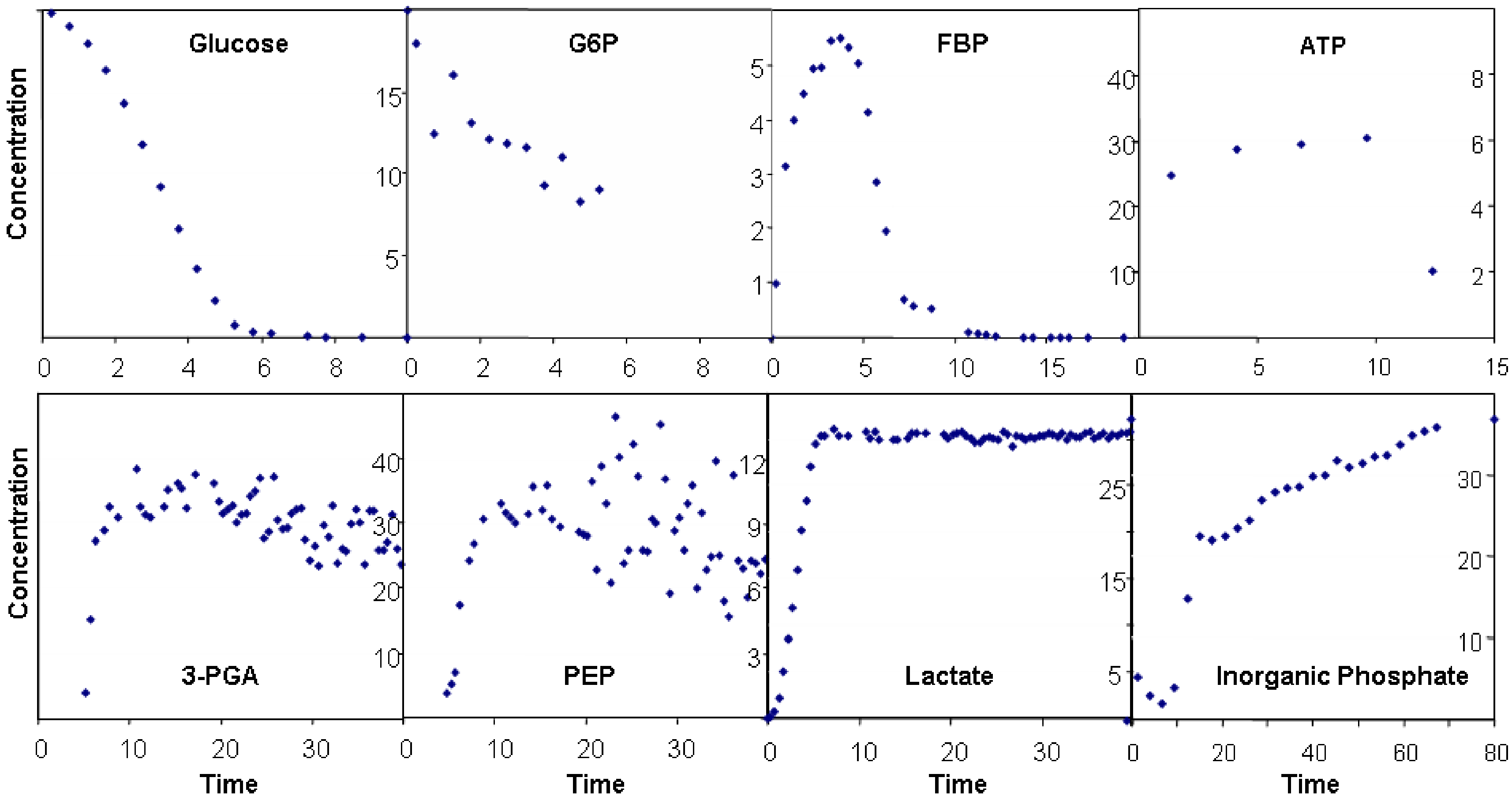
www.hhmi.org/bulletin/winter2005/images/bacteria5.jpg

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

Goals of Modeling

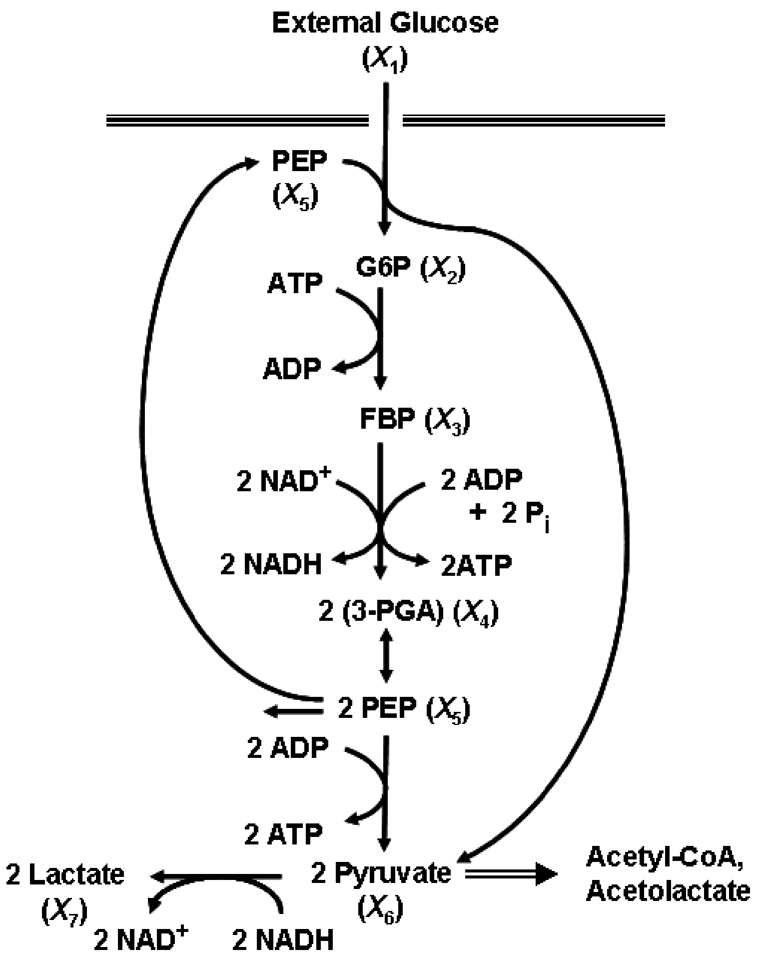
- Understand pathway; design, operation
- Allow extrapolation to new situations
- Allow prediction for manipulation
- Maximize yield of main product
- Optimize yield of secondary products
- Eventually develop a cell-wide model

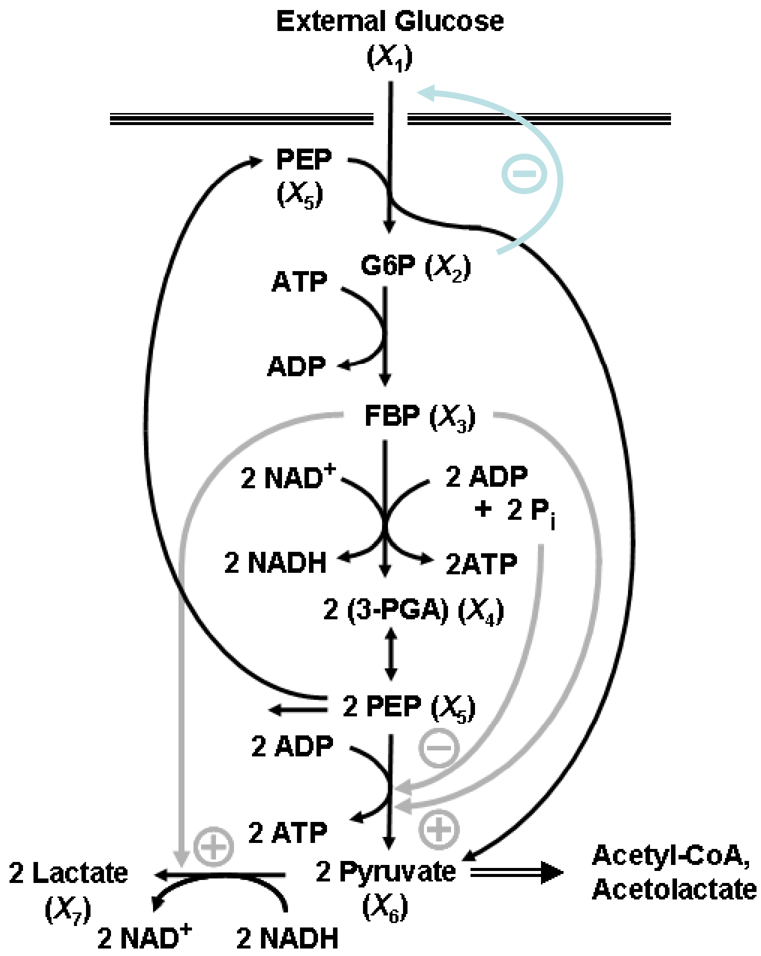
Experimental Time Series Data



Voit et al.: IEE Proc. Systems Biol. 2006; PNAS 2006

Other Information





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

Change in variable @ t = All influxes @ t – All effluxes @ t

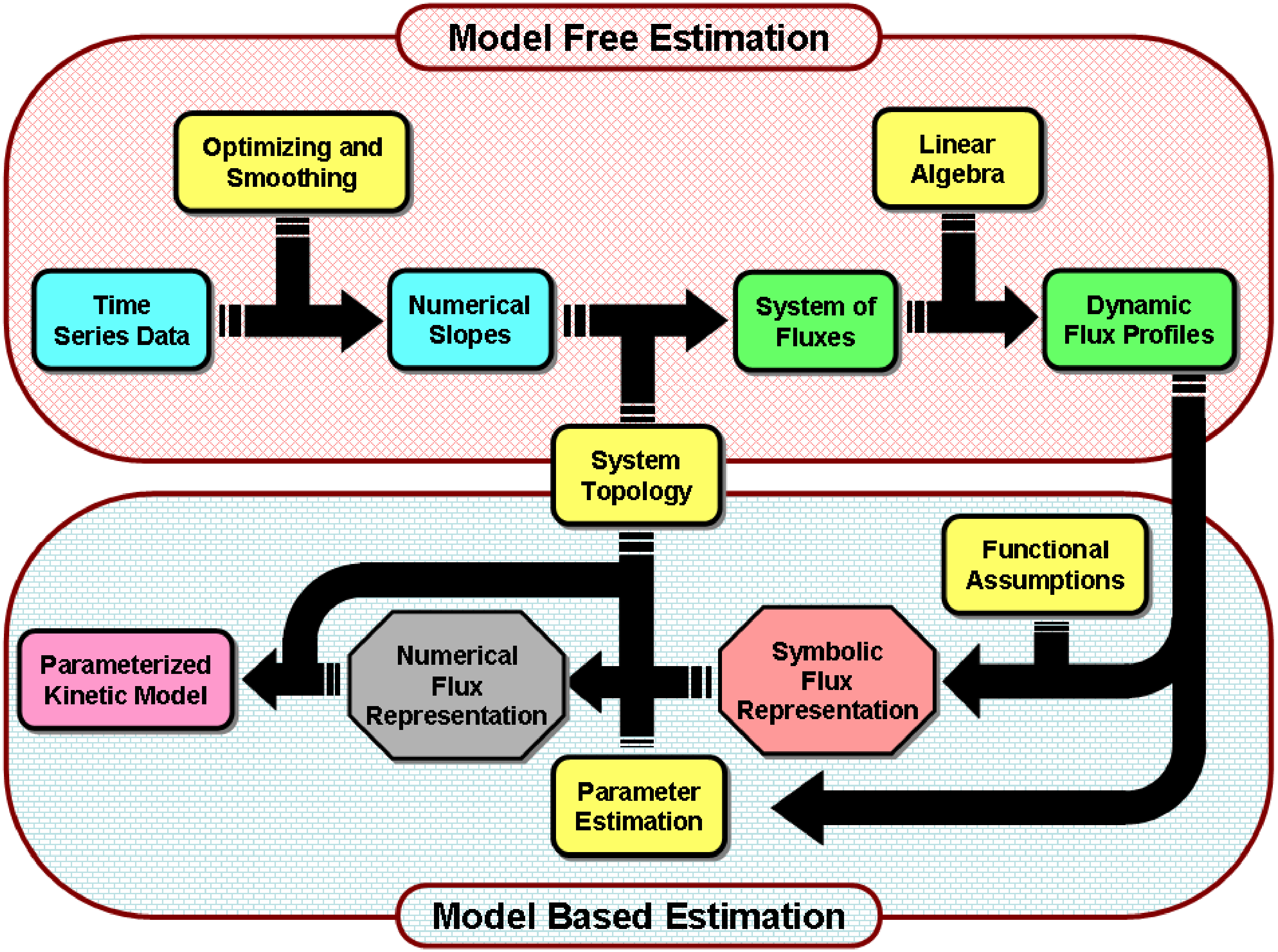
Linear system; solve as far as possible

Result: values of each flux @ time points t_i

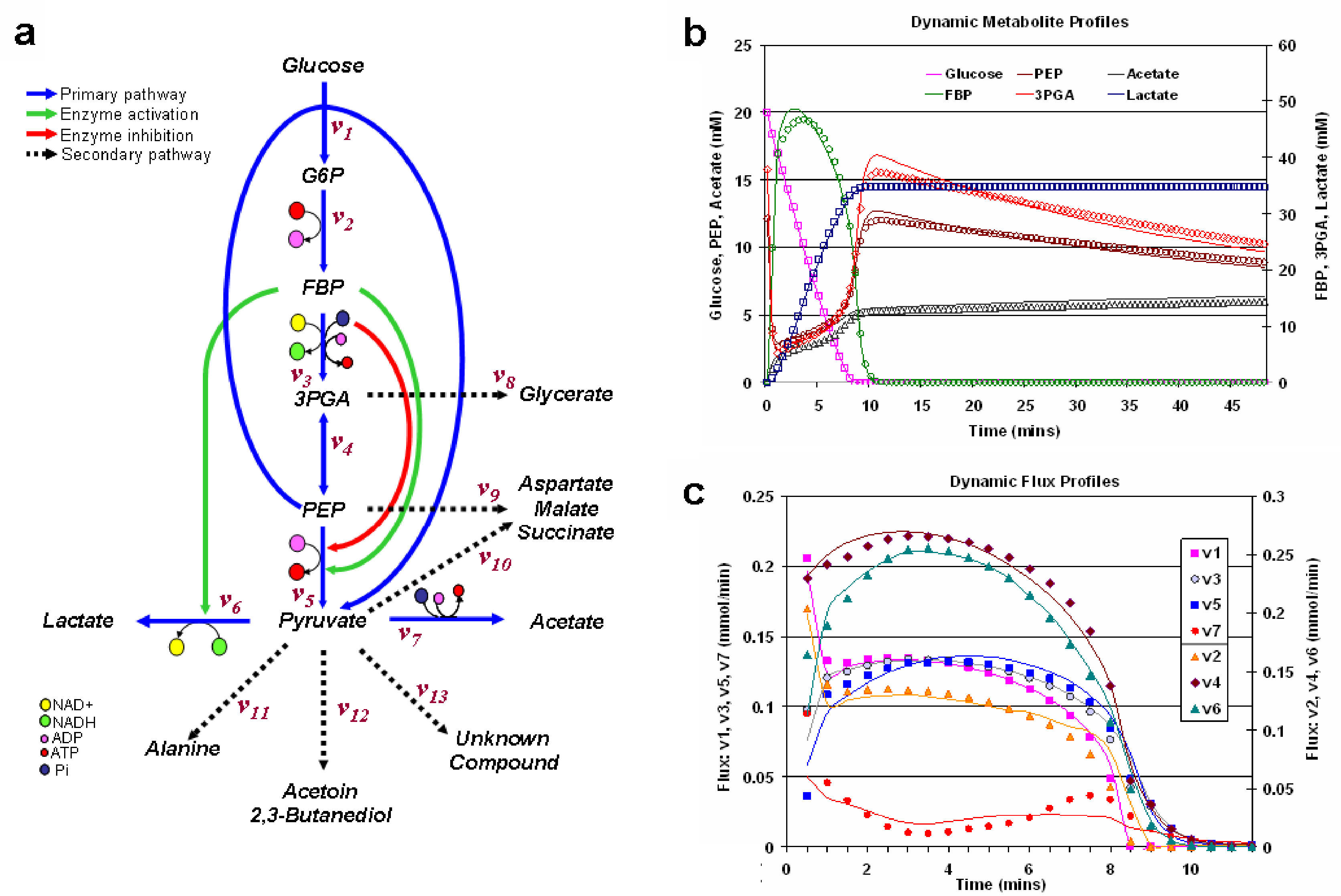
Represent fluxes with appropriate models

Goel, Chou, Voit, Bioinformatics, 2008

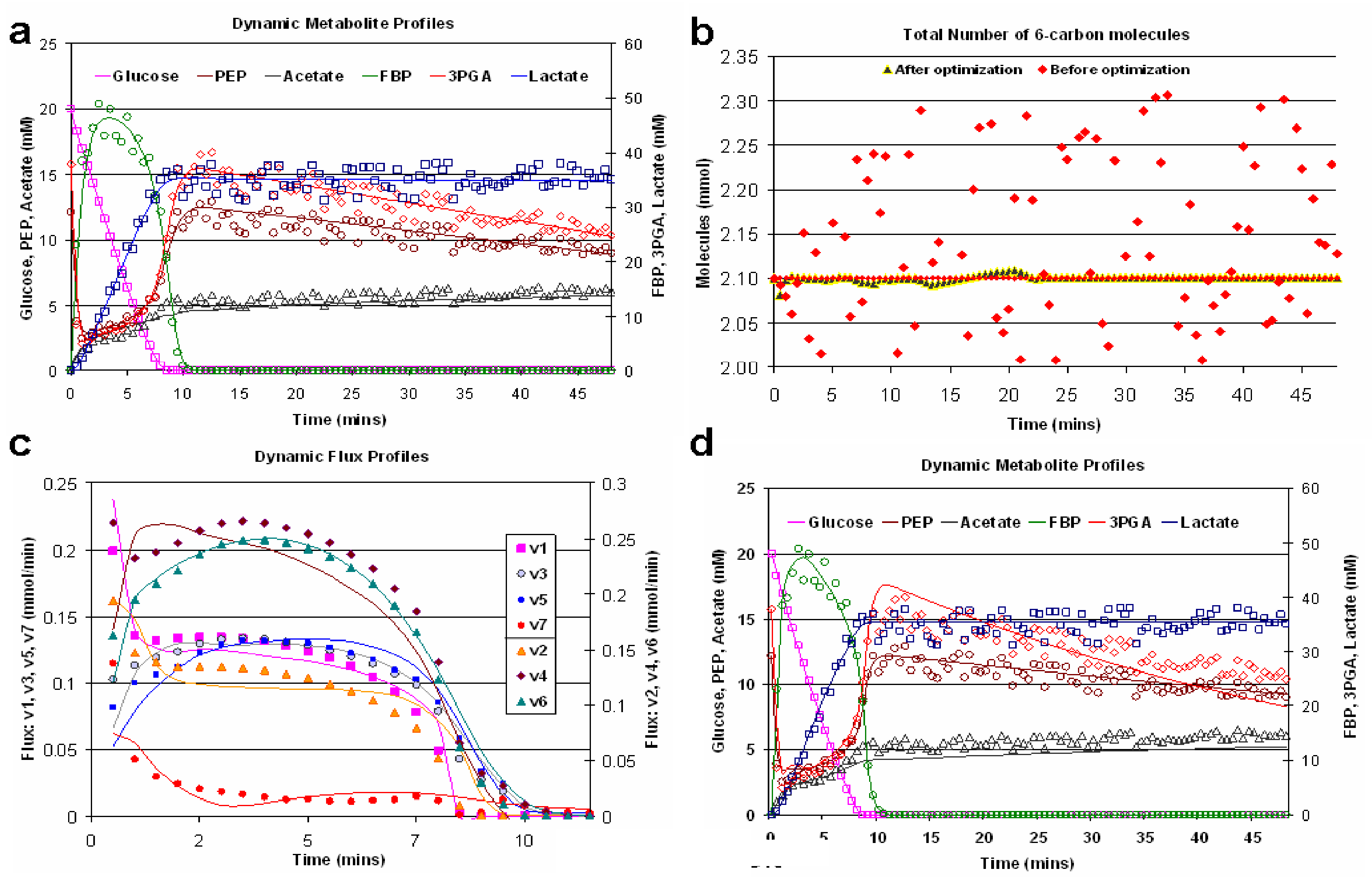
Dynamic Flux Estimation (DFE)



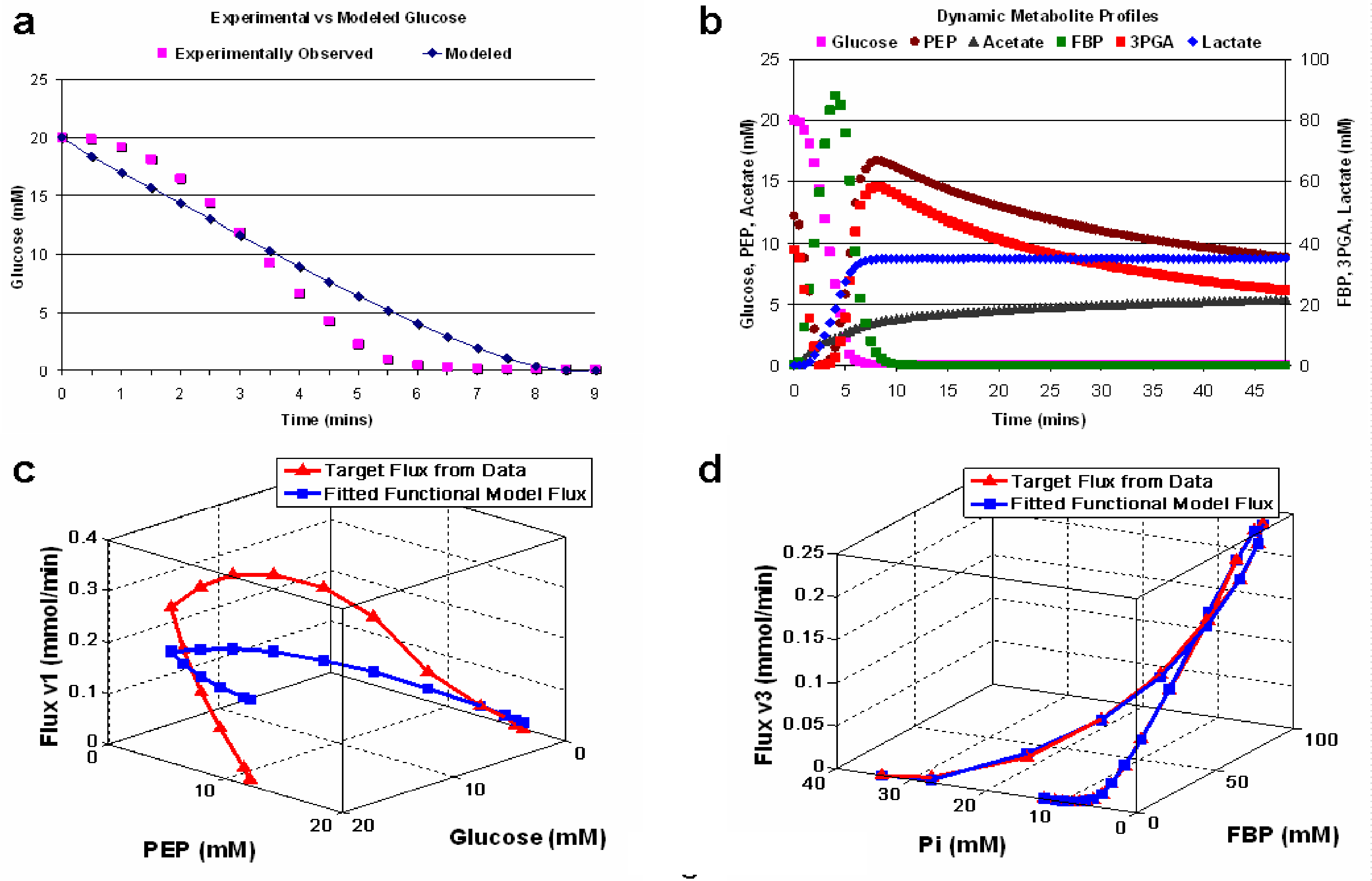
Dynamic Flux Estimation (DFE)



Dynamic Flux Estimation (DFE)



Dynamic Flux Estimation (DFE)



Open Problems

Smoothing and Mass Conservation:

Noise in the data leads to loss or gain of mass

Redundancies / Sloppiness:

Many models fit the data

Underdetermined Flux Systems:

Linear system of fluxes not of full rank

Extrapolation:

System fails for new data

Ill-defined Systems:

Significant information is missing

Smoothing and Mass Conservation

Issue:

Noise in the data leads to loss or gain of mass

Possible Causes:

Experimental measurement errors
Secondary pathways ignored (PPP ~ 5%)
Ethanol evaporates

Possible Remedies:

Identify where mass is lost/gained;
add (degradation, production) reactions
to the model
Constrained smoothing (*e.g.*, with wavelets)

Redundancies / Sloppiness

Issue:

Many models fit the data

Possible Reasons:

1. Data collinear or non-informative
2. High noise permits different models
3. Noise-free data admit different models

Possible Remedies:

1. Pooling of data; set variables constant
2. Monte-Carlo identification of “neutral ensembles”
   More datasets and constraints
3. Lie transformation group analysis

Vilela, Vinga, Grivet, Voit, Almeida, BMC Bioinformatics, 2009

Underdetermined Flux Systems

Issue:

Linear system of flux often not of full rank;

can’t solve uniquely for fluxes

Dominant Cause:

More reactions than metabolites in most pathways

Potential Remedies:

Augment DFE with other methods

bottom-up estimation of some fluxes

Alternating Regression

Prefitting; Flux balance analysis; lin-log

Constraints (maximize growth)

Voit, Goel, Chou, da Fonseca, IET Syst. Biol., 2009

Extrapolation

Issue:

Model fit good, but extrapolation fails

Dominant Cause:

Functional representation of flux profile incorrect

Potential Remedies:

Analyze more data with slightly changed system
Develop better kinetic description
Attempt piecewise representation

Ill-defined Systems

Issue:

Data, time courses missing

Dominant Cause:

Experimental difficulties, *e.g.*, human systems

Potential Remedies:

Order-of-magnitude modeling
Canonical models with default parameter values
Data per expert opinion

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

Summary

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.

2. Technical challenges abound. Important: Efficient, robust, and fast solutions on PC's needed. No single algorithm satisfactory.

3. Important overlooked issue: Error compensation; extrapolation becomes unreliable. DFE promising, but needs auxiliary methods.

4. Many problems remain unsolved.

Acknowledgements

The 2008 Crew:

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

Information: www.bst.bme.gatech.edu

Selected Parameter Estimation References

Voit, E.O. and M.A. Savageau: Power-law approach to modeling biological systems; III. Methods of analysis. *J. Ferment. Technol.* 60 (3), 233-241, 1982.

Voit, E.O.: The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions. *Biometrical J.* **34 (7)**, 855-878, 1992.

Sands, P.J., and E.O. Voit: Flux-based estimation of parameters in S-systems. *Ecol. Modeling* **93**, 75-88, 1996.

Voit, E.O.: *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*, xii + 530 pp., Cambridge University Press, Cambridge, U.K., 2000.

Sorribas, A., J. March, and E.O. Voit: Estimating age-related trends in cross-sectional studies using S-distributions, *Stats. in Med.* **10(5)**: 697-713, 2000.

Voit, E.O.: A maximum likelihood estimator for the shape parameters of S-distributions, *Biometr. J.*, **42 (4)**, 471-479, 2000.

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

Almeida, J.S., and E.O. Voit: Neural-network-based parameter estimation in complex biomedical systems. *Genome Informatics* 14, 114-123, 2003.

Voit, E.O., and J.S. Almeida: Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* 20(11),1670-1681, 2004.

Veflingstad, S.R., J.S. Almeida, and E.O. Voit: Priming nonlinear searches for pathway identification. *BMC Theoretical Biology and Medical Modelling* **1**:8, 2004.

Voit, E.O., S. Marino, and R. Lall: Challenges for the identification of metabolic pathways from time series data. *In Silico Biology* **5**, 83-92 (2005).

He, Q. and E.O. Voit: Estimation and completion of survival data with piecewise linear models and S-distributions. *J. Stat. Comp. Simul.* 75(4), 287-305 (2005).

***key references given in bold**

Selected Parameter Estimation References (cont'd)

Lall, R. and E.O. Voit: Parameter Estimation in Modulated, Unbranched Reaction Chains within Biochemical Systems. *J. Comput. Biol. Chem.* 29, 309-318 (2005).

Marino, S. and E.O. Voit: An automated procedure for the extraction of metabolic network information from time series data. *J. Bioinform. Comp. Biol.* **4**, 665-691, 2006.

Polisetty, P.K., E.O. Voit, and E. P. Gatzke: Identification of metabolic system parameters using global optimization methods. *BMC Theoretical Biology and Medical Modelling*, 3(1):4, 2006.

Chou, I-C., H. Martens, and E.O. Voit. Parameter Estimation in Biochemical Systems Models with Alternating Regression. *BMC Theoretical Biology and Medical Modelling* 3:25, 2006.

Goel, G., I-Chun Chou, and E.O. Voit: Biological Systems Modeling and Analysis: A Biomolecular Technique of the 21st Century. *J. Biomolec. Techn.* **17**, 252-269, 2006.

Chou, I-C., H. Martens, and E.O. Voit: Parameter Estimation of S-distributions with Alternating Regression. *Stat. Operations Res. Transactions (SORT)*, **31(1)**, 55-74. 2007.

Vilela, M., C. Borges, A. T. Vasconcelos, H. Santos, E. O. Voit. and J. S. Almeida: Automated smoother for numerical decoupling of dynamic models. *BMC Bioinformatics* **8**:305, 2007.

Vilela, M., I-C.Chou, S. Vinga, A.T.R Vasconcelos, E.O. Voit, and J.S. Almeida: Parameter optimization in S-system models. *BMC Systems Biol.* **16**; 2:35, 2008.

Goel, G., I-C. Chou, and E.O. Voit: System Estimation from Metabolic Time Series Data. *Bioinformatics* 24, 2505-2511, 2008.

Chou, I-C. and E.O. Voit: Recent Developments in Parameter Estimation and Structure Identification of Biochemical and Genomic Systems. *Math. Biosc.* In press, 2009.

Voit, E.O., G. Goel , I-C. Chou, and L. da Fonseca: Estimation of Metabolic Pathway Systems from Different Data Sources. *IET Systems Biol.* in press, 2009.

Ko, C.-L., E.O. Voit, and F.-S. Wang: Estimating Parameters for Generalized Mass Action Models with Connectivity Information. BMC Bioinformatics. in press, 2009.

Vilela, M., S. Vinga, M. A. Grivet Mattoso Maia, E. O. Voit, J. S. Almeida: Identification of neutral sets of biochemical systems models from time series data. BMC Systems Biology 3:47, 2009.

Voit, E.O. and I-C. Chou: Parameter Estimation in Canonical Biological Systems Models. *Int. J. Syst. Synth. Biol.* in press, 2009.