High-Performance Systems Biology and Associated Combinatorial Scientific Computing Problems

David Alber

Materials and Computational Science Center
National Renewable Energy Laboratory

June 11, 2008

NREL/PR-590-43765

Presented at the 2008 Workshop on Combinatorial Scientific Computing & Petascale Simulations held June 10-13, 2008 in Sant Fe, New Mexico.

Project Participants and Funding

Collaborators:

- NREL SCG: Chris Chang, Peter Graf, and Kwiseon Kim*
- NREL Photobiology: Mike Seibert*
- Summer Student from CU Boulder: David Biagioni

Participating institutions:

- National Renewable Energy Laboratory
- Colorado School of Mines
- Stanford University

Funding through SciDAC (OASCR and OBER)

Metabolism and Metabolic Modeling

Metabolism

- Chemical reactions occurring in living cells
- Reactions catalyzed by enzymes
- Metabolic species (e.g., glucose, pyruvate) produced and consumed

Reaction Modeling

- Several models
- Michaelis-Menten kinetics

High-Performance Systems Biology

In a nutshell

- Model complete metabolism of *Chlamydomonas reinhardtii*
- Develop high-performance software to explore metabolism kinetics

Example problems

- Parameter estimation (data fitting)
- Sensitivity minimization
- Parameter space characterization

Metabolic Model

For metabolic reaction:

$$\frac{dy}{dt} = f(y, k, E)$$

- y – vector of metabolite concentrations
- k – vector of kinetic parameters
- E – vector of enzyme concentrations

Metabolic Model

For metabolic reaction:

$$\frac{dy}{dt} = f(y, k, E)$$

- y – vector of metabolite concentrations
- k – vector of kinetic parameters (mostly unknown)
- E – vector of enzyme concentrations

Metabolic Model

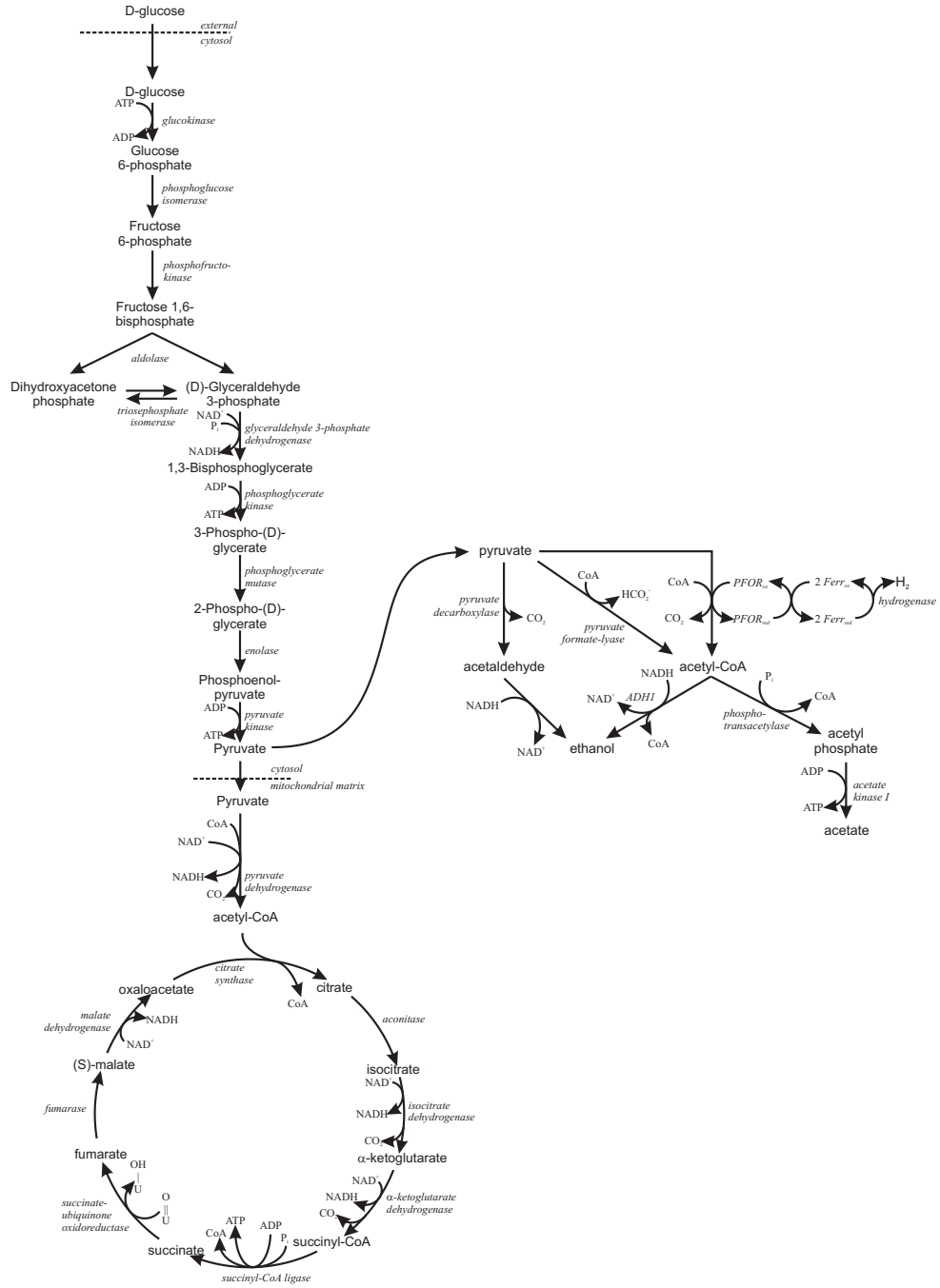
For metabolic reaction:

$$\frac{dy}{dt} = f(y, k, E)$$

- y – vector of metabolite concentrations
- k – vector of kinetic parameters (mostly unknown)
- E – vector of enzyme concentrations

Kinetic parameters:

- Time consuming to determine experimentally
- Essential to understanding metabolic kinetics


D-glucose
external
cytosol
D-glucose
ATP
glucokinase
ADP
Glucose 6-phosphate
phosphoglucose isomerase
Fructose 6-phosphate
phosphofructo-kinase
Fructose 1,6-bisphosphate
aldolase
Dihydroxyacetone phosphate
triosephosphate isomerase
(D)-Glyceraldehyde 3-phosphate
NAD+
Pi
glyceraldehyde 3-phosphate dehydrogenase
NADH
1,3-Bisphosphoglycerate
ADP
phosphoglycerate kinase
ATP
3-Phospho-(D)-glycerate
phosphoglycerate mutase
2-Phospho-(D)-glycerate
enolase
Phosphoenol-pyruvate
ADP
pyruvate kinase
ATP
Pyruvate
cytosol
mitochondrial matrix
Pyruvate
CoA
NAD+
NADH
CO2
pyruvate dehydrogenase
acetyl-CoA
citrate synthase
oxaloacetate
CoA
citrate
malate dehydrogenase
NADH
NAD+
aconitase
(S)-malate
isocitrate
NAD+
fumarase
isocitrate dehydrogenase
NADH
CO2
fumarate
α-ketoglutarate
OH
succinate-ubiquinone oxidoreductase
NAD+
NADH
α-ketoglutarate dehydrogenase
CO2
ATP
ADP
CoA
Pi
succinate
succinyl-CoA
succinyl-CoA ligase
pyruvate
pyruvate decarboxylase
CO2
CoA
HCO2-
pyruvate formate-lyase
CoA
CO2
PFOR_ox
PFOR_red
2 Ferr_ox
2 Ferr_red
H2
hydrogenase
acetaldehyde
NADH
NAD+
ethanol
NADH
NAD+
ADH1
acetyl-CoA
CoA
Pi
phospho-transacetylase
acetyl phosphate
ADP
ATP
acetate kinase I
acetate

Parameter Estimation

- Find set of k such that species concentrations match target values
- Expressed as optimization problem:

$$\min_k g(k),$$

where $g(k) = \|r\|_2^2$ and r is vector of differences between target and simulated values

Sensitivity Minimization

- Find set of k to minimize *sensitivity*
- Applies to organism engineering issues and general "nature of life" questions
- Objective function to minimize:

$$\begin{aligned} h(k) &= \|J_y(k)\|_F^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\partial y_i}{\partial k_j}\right)^2 \end{aligned}$$

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g (adjoint sensitivity analysis)

y_i: simulated metabolite concentration

$\bar{y}_i$: target concentration

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g
(adjoint sensitivity analysis)

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g
(adjoint sensitivity analysis)

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g (adjoint sensitivity analysis)

Sensitivity minimization:

$$h(k) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\partial y_i}{\partial k_j} \right)^2$$

Single entry of ∇h:

$$\frac{\partial h}{\partial k_\ell} = 2 \sum_{i,j} \frac{\partial y_i}{\partial k_j} \frac{\partial^2 y_i}{\partial k_j \partial k_\ell}$$

Computing second derivative term expensive

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g (adjoint sensitivity analysis)

Sensitivity minimization:

$$h(k) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\partial y_i}{\partial k_j} \right)^2$$

Single entry of ∇h:

$$\frac{\partial h}{\partial k_\ell} = 2 \sum_{i,j} \frac{\partial y_i}{\partial k_j} \frac{\partial^2 y_i}{\partial k_j \partial k_\ell}$$

Computing second derivative term expensive

Gradient-based Optimization

Parameter estimation:

$$g(k) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Single entry of ∇g:

$$\frac{\partial g}{\partial k_j} = 2 \sum_{i=1}^{n} (y_i - \bar{y}_i) \frac{\partial y_i}{\partial k_j}$$

No problems computing ∇g (adjoint sensitivity analysis)

Sensitivity minimization:

$$h(k) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\partial y_i}{\partial k_j} \right)^2$$

Single entry of ∇h:

$$\frac{\partial h}{\partial k_\ell} = 2 \sum_{i,j} \frac{\partial y_i}{\partial k_j} \frac{\partial^2 y_i}{\partial k_j \partial k_\ell}$$

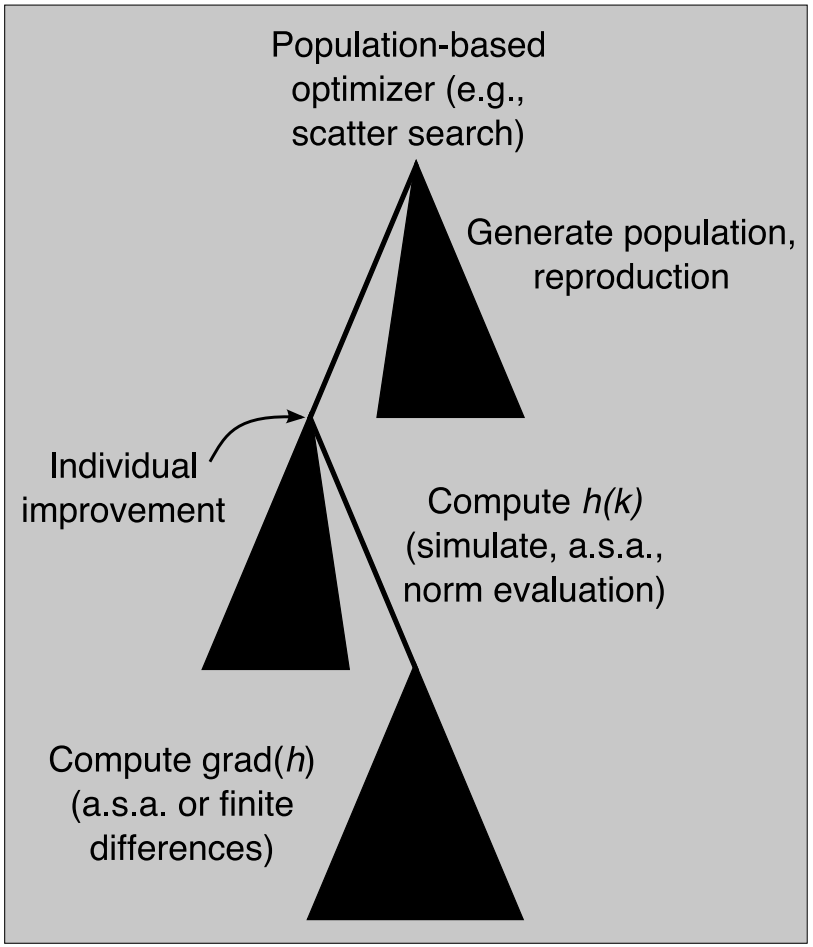
Computing second derivative term expensive

Second Derivative Computation
CSC Problem Slide #1

- Finite differences
- Automatic differentiation
- Collaboration with Paul Hovland
- Collaboration with Radu Serban (formerly at LLNL):
  - Combines adjoint sensitivity analysis and AD (via Tapenade)
  - Parallelized
  - Received code two weeks ago (i.e., nothing yet to show)

Sensitivity Analysis Cost Substantial

- $\approx$1000 dimension parameter space (ultimately)
- Each evaluation of $h(k)$ inexpensive (similar to cost of ∇g)
- Evaluation cost of ∇h adds up



Load Balancing
CSC Problem Slide #2

- The hierarchical parallelism in this application generates load balance issues
- More of the scheduling problem variety
- Want to hear more on this

Kinetics Complications

- Many gaps to fill regarding metabolic kinetic rates
- Requires much experimentally-obtained data
- Relevant experimental data difficult (tedious?) and expensive to generate
- Switch gears and discuss another systems biology (metabolomics) problem

Stoichiometric Matrix

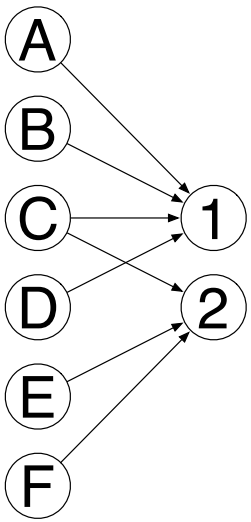
- Stoichiometry refers to numbers in chemical formulae

$$2A + B \longrightarrow C + 3D$$

$$C + 2E \longrightarrow F$$

- Stoichiometric matrix stores stoichiometry for all reactions

$$S = \begin{pmatrix} -2 & 0 \\ -1 & 0 \\ 1 & -1 \\ 3 & 0 \\ -2 & 0 \\ 0 & 1 \end{pmatrix}$$

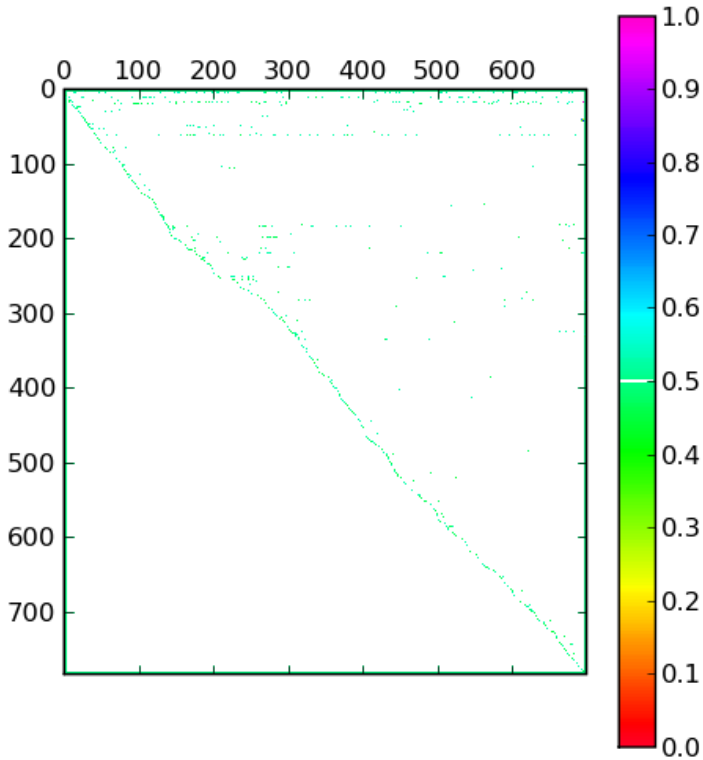


Stoichiometric Matrix and Flux Analysis

- Previously said $dy/dt = f(y, k, E)$
- Can now also say

$$\frac{dy}{dt} = Sv,$$

  where v is vector of *reaction velocities*
- No kinetic parameters to get in the way!
- At steady state, $dy/dt = 0$ (almost)
- Now looking at $Sv = 0$



Problem Definition

- Say we want to maximize concentration of species y_i:

$$\max_{v} y_i$$

- This problem trivial without constraints
- Possible constraints:
  - Thermodynamics (an arbitrarily large reaction rate not possible)
  - $Sv = 0$ at steady state (almost)
  - All $v_i \geq 0$ (reactions do not happen in reverse – not exactly true)

Add Another Layer

- Suppose goal is to dispose of some reactions or pathways
- Which pathways are more critical?
- Conceptual connections to electrical grid modeling?

Molecular Dynamics

Shameless plug for another SciDAC project

Application:

- Study function of CBH I enzyme in “digesting” crystalline cellulose
- Processivity?

Challenges:

- CHARMM does not scale well, has everything needed for application
- Other packages scale well, do not have everything needed
- If CHARMM: load balancing