Constraint-based Modeling: Part II
LP, Lessons Learned, and the Growing Field of CBM

Tuesday, May 4, 2004
Timothy E. Allen / Bernhard Ø. Palsson
BE 203 Lecture
Outline

• Constraints in biology
• Reconstructions and applying constraints
• Constraint-based modeling (CBM): philosophy and overview
• Basics of flux balance analysis (FBA)
• Lessons learned
• CBM: an expanding field
Lessons Learned:
Applications to Genome-scale in silico Reconstructions
Lessons Learned: Applications to Genome-scale in silico Reconstructions

1. Deletion studies (\emph{H. influenzae})
2. Essential amino acids (\emph{H. pylori})
3. Reaction subsets / operons (\emph{E. coli})
4. Gap analysis (\emph{E. coli})
5. Optimal growth predictions / adaptive evolution (\emph{E. coli})
6. Iterative hypothesis generation (\emph{E. coli})
7. Integration of heterogeneous datasets (\emph{E. coli})
Example #1: Gene Deletions & Production Deficiencies

H. Influenzae Central Metabolism

50 Biomass Requirements

Production Capabilities Under Two Environmental Conditions:
1. “in vitro” Minimal Media (fructose)
2. “in vivo” Complete Conditions (multiple carbon sources)

Minimal Substrate Conditions
(fructose)

Carbon-supplemented Conditions
(fructose, glucose, glycerol, galactose, fucose, ribose, and sialic acid)
Example #2: H. Pylori Minimal Requirements

- 8 amino acids required
  - Arginine
  - Leucine
  - Methionine
  - Phenylalanine
  - Valine
  - Thiamin
  - Phosphate
  - Oxygen
  - Sulphate/Cysteine

- Purine sources
  - Adenine
  - Adenosine
  - Guanine
  - Guanosine
  - Hypoxanthine

- Sulfur source
  - Cysteine
  - Sulphate

- Oxygen
  - No substrate level phosphorylation (lacks PYK)

- Glutamate
  - Requires alanine or arginine, only component not dependent on one substrate

* Adenine, Adenosine, Guanine, Guanosine, Hypoxanthine

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu
Example #3: Reaction Subsets (E. coli)

Regulon

extracellular

intracellular

AstE
astB
astD
astA
astC

SpeB
speA

glpK
glpF

ArcD
YdgB

P = 0.003
Correlated Sets / Operons (E. coli)

Correlation of Genes in Correlated Sets and Operons Using Expression Data for E. coli

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu
Example #4: Network Gap Analysis

Models can be used to guide biological discovery.
# 55 Putative Annotations

<table>
<thead>
<tr>
<th>Bnum</th>
<th>EC number</th>
<th>Published Annotation [Serres et al.]</th>
<th>Suggested Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>b3718</td>
<td>3.1.1.17</td>
<td>putative isomerase</td>
<td>gluconolactonase</td>
</tr>
<tr>
<td>b2160</td>
<td>2.7.1.13</td>
<td>putative sugar kinase</td>
<td>dehydrogluconokinase</td>
</tr>
<tr>
<td>b2166</td>
<td>2.7.1.14</td>
<td>putative sugar kinase</td>
<td>sedoheptulokinase</td>
</tr>
<tr>
<td>b2661</td>
<td>1.2.1.19;</td>
<td>succinate-semialdehyde dehydrogenase</td>
<td>aminobutyraldehyde dehydrogenase; succinate-semialdehyde</td>
</tr>
<tr>
<td></td>
<td>1.2.1.24</td>
<td>NADP-dependent</td>
<td></td>
</tr>
<tr>
<td>b4266</td>
<td>1.1.1.6</td>
<td>5-keto-D-gluconate-5-reductase</td>
<td>glycerol dehydrogenase.</td>
</tr>
<tr>
<td>b3003</td>
<td>1.1.1.6</td>
<td>putative oxidoreductase, NAD(P)-binding</td>
<td>glycerol dehydrogenase.</td>
</tr>
<tr>
<td>b2137</td>
<td>1.1.1.5</td>
<td>putative oxidoreductase</td>
<td>acetoin dehydrogenase, Diacetyl reductase</td>
</tr>
<tr>
<td>b2615</td>
<td>2.7.1.23</td>
<td>ORF</td>
<td>NAD+ kinase</td>
</tr>
<tr>
<td>b3718</td>
<td>3.1.1.31</td>
<td>putative isomerase</td>
<td>6-phosphogluconolactonase (Pgl)</td>
</tr>
<tr>
<td>b1511</td>
<td>2.7.1.47</td>
<td>putative sugar kinase (2nd module)</td>
<td>D-ribulokinase</td>
</tr>
<tr>
<td>b1524</td>
<td>3.5.1.2</td>
<td>putative glutaminase</td>
<td>glutaminase A,B</td>
</tr>
</tbody>
</table>

Enzymes acting on network gaps

Suggest alternate substrates for enzymes

Multiple hits for target enzymes

Enzymes in *E. coli* without locus assignments (EcoCyc)

Metabolic model makes growth predictions for knock-out strains (86%).
Regulated metabolic model increases accuracy of predictions (91.4%).
Example #5: Predicting complex biology; adaptive evolution and picking optimal growth states
Using Adaptive Evolution

1. Metabolic reconstruction

2. Constraints imposed on the metabolic network

3. Linear optimization to find 'best solution

4. Phenotype phase plane analysis: Displays all optimal solutions simultaneously

5. Whole Genome Directed Evolution
Evolving Growth Rate on Glycerol

**Growth rate**

**Evolving Growth Rate on Glycerol**

**Day 0**

**Day 0 - 40**

**Day 40**

**Day 60**

University of California, San Diego
Department of Bioengineering


Systems Biology Research Group
http://systemsbiology.ucsd.edu
Genome location of glycerol metabolic and regulatory genes selected for re-sequencing

Fructose 1,6 bisphosphate (FBP) binding site on glpK

FBP binding loop (230-236) : IGGKGGTR

Mutation in FBP binding site

(ggc → gac)

wt

SEVYGQTNIIGGKGGTRIPIS

mut

SEVYGQTNIIGGGKGGTRIPIS

AA #

230

236

2-3 fold decrease in inhibition by 2mM FBP on activity of glpK with mutation

~ 10 fold increase in activity of glpK with mutation (G231D: GLY → ASP)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu
Example #6: Hypothesis generation: transcriptional regulation in *E. coli*

**Step one:** Reconstruct computational model based on available data

**Step two:** Compare new observations to computational predictions

**Step three:** Expand model via hypothesis generation

Initial Model

Prediction & Data Generation

Consistent

Inconsistent

Expanded Model

Iteration

Interpretation

University of California, San Diego
Department of Bioengineering
Model-Centric Hypothesis Generation

- Genome-scale regulatory/metabolic model of *E. coli*
  - 1,008 genes
- Systematic network perturbation analysis
  - ArcA, Fnr, ArcA/Fnr, AppY, OxyR, SoxS
- Generate new rules for model
- Hypotheses generation
Model-driven hypothesis generation

Gene Expression Study
- Added new rules for 78 genes
- Removed old rules for 27 genes
- Changed old rules for 10 genes
- Total of 115 changes in regulatory rules

Prediction & Data Generation

Step two: Compare new observations to computational predictions

Iteration

Step three: Expand model via hypothesis generation

110 new regulatory hypotheses overall

Gene Expression Study
- Phenotypic Predictions
  - 79% (10833/13750) accuracy
- Expression Predictions
  - 98% (100/102) accuracy
  - 66% (100/151) coverage

Gene Expression Study
- Phenotypic Predictions
  - 79% (10828/13750) accuracy
- Expression Predictions
  - 49% (23/47) accuracy
  - 15% (23/151) coverage

Step one: Reconstruct computational model based on available data

iMC1010v1
- Phenotypic Predictions
  - 79% (10828/13750) accuracy
- Expression Predictions
  - 49% (23/47) accuracy
  - 15% (23/151) coverage

iMC1010v2
- Phenotypic Predictions
  - 79% (10833/13750) accuracy
- Expression Predictions
  - 98% (100/102) accuracy
  - 66% (100/151) coverage

Covert et al, Nature (in press) 2004

Systems Biology Research Group
http://systemsbiology.ucsd.edu

University of California, San Diego
Department of Bioengineering
Example #7: Integration of multiple data sets: periodicity in gene expression
Integrating Heterogeneous Datasets

**Genomics**

Glyceraldehyde 3-Phosphate Dehydrogenase

**Transcriptomics**

Succinate Dehydrogenase

**Proteomics**

D-Xylose ABC Transporter

**"Fluxomics"**

GPR ASSOCIATIONS

<table>
<thead>
<tr>
<th>LEVELS</th>
<th>GPR ASSOCIATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>Succinate Dehydrogenase</td>
</tr>
<tr>
<td>Protein</td>
<td>Sdh</td>
</tr>
<tr>
<td></td>
<td>(b0721 and b0722)</td>
</tr>
<tr>
<td></td>
<td>(b0723 and b0724)</td>
</tr>
<tr>
<td></td>
<td>(Sdh)</td>
</tr>
<tr>
<td>mRNA</td>
<td>(Sdh)</td>
</tr>
<tr>
<td>Gene</td>
<td>(b0723)</td>
</tr>
</tbody>
</table>

| Reaction | D-Xylose ABC Transporter |
| Protein | XyIF and XyIG and XylH |
| | (XyIF and XyIG and XylH) |
| mRNA | (b3568) |
| | (b3567) |
| | (b3566) |
| Gene | (b3567) |
| | (b3566) |

| Reaction | Glyceraldehyde 3-Phosphatase Dehydrogenase |
| Protein | GapA or GapC |
| | (GapA) |
| | (GapC) |
| mRNA | (b1779) |
| | (b1416 and b1417) |
| Gene | (b1779) |
| | (b1416 and b1417) |
Integrating “Omics” Data

$q = (q_1 \ldots q_{4290})$ constitutes the "transcription state" of the genome

$t = (t_1 \ldots t_{4290})$ can be calculated on a per codon basis and account for relative tRNA abundance to give the state of the proteome.
Periodicity in genome usage

- Periodicity in *E. coli* expression of ~100 and ~600 genes
- Appear to be distinct 6 regions of genome usage

*Allen, et al., JBact 185:6392 (2003)*
Topobiology of *E. coli* Genome

C. Woldringh, 2001

University of California, San Diego
Department of Bioengineering
Integration of “Omics” Data

• Simultaneous analysis of multiple “-omics” data sets leads to new insights
• Topobiology at the ~200 nm scale seems to be important
• Means of accounting for 3D structural constraints is needed in whole-cell reconstructions going forward
Constraint-based Modeling: An Expanding Field
Development of the *E. coli* Model

"Thirteen years of constraint-based model building of *E. coli*" *J. Bacti*, May 2003

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu
(Several slides deleted due to copyright issues...)

Systems Biology Research Group
Principal Investigator: Bernhard Ø. Palsson, Ph.D.

University of California, San Diego
Department of Bioengineering