Multiscale Molecular Systems Biology: Reconstruction and Model Optimization

Dr. Ronan M.T. Fleming, Systems Biochemistry Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg.

Friday, August 16, 2013
Interagency Modeling and Analysis Group Webinar
Multiscale Systems Biology Collaboration

- **Molecular Systems Physiology Group**
  - [Ines Thiele](#), Luxembourg Centre for Systems Biomedicine.

- **Systems Biology Research Group**
  - [Bernhard Palsson](#), University of California, San Diego.

- **Systems Optimization Laboratory**
  - [Michael Saunders](#), Stanford University.

- **Systems Biochemistry Group**
  - [Ronan Fleming](#), Luxembourg Centre for Systems Biomedicine.
Variables with magnitudes spread over many orders of magnitude.

Chemical formula known for each molecule.

System of biochemical reactions with defined boundary conditions.

Biomedical, biotechnological, environmental, applications.

Multiscale Molecular Systems Biology: Reconstruction and Model Optimization

An abstraction of select biochemical, genetic, and genomic experimental knowledge about a chosen biochemical subsystem

General mathematical model, combined with particular reconstruction, thus creating a computational model.

Numerical optimization problems with a firm grounding in mathematical optimization theory.
Increasing the comprehensiveness of genome scale computational models ....

– Increasing size
  
  • e.g. single microbe versus whole microbial community
    – 1 microbial species ~ 1e3 reactions
    – 1 community (1000 species) ~ 1e6 reactions

– Increasing ratio of fastest to slowest timescale
  
  • e.g. genome scale metabolic model versus integrated model of metabolism and macromolecular synthesis
    – Metabolic reactions
    – Macromolecular synthesis reactions

– Increased simulation fidelity
  
  • e.g. mass conservation alone, versus mass conservation, energy conservation, second law of thermodynamics, reaction kinetics, etc.
... leads to a mathematical and numerical optimization challenge:

• Large scale numerical optimization
  – Reduce computational complexity of algorithms to solve optimization problems

• Multiscale numerical optimization
  – Standard optimization software ideal for \( O(1) \) variables

• Mathematical formulation
  – Biochemical function is an inherently nonlinear process
  – How to formulae a mathematical modeling problem in a form amenable to a polynomial time algorithm
Reconstruction of reaction stoichiometry

Stoichiometric Matrix (denoted S)
Reconstruction of macromolecular synthesis machinery

<table>
<thead>
<tr>
<th># transcription units</th>
<th>249</th>
</tr>
</thead>
<tbody>
<tr>
<td># genes</td>
<td>423</td>
</tr>
<tr>
<td>tRNA/rRNA/misc RNA</td>
<td>86/22/1</td>
</tr>
<tr>
<td>proteins (w/o genes)</td>
<td>228 (34)</td>
</tr>
<tr>
<td># subsystems</td>
<td>27</td>
</tr>
<tr>
<td># reactions</td>
<td>13,694</td>
</tr>
<tr>
<td># components</td>
<td>11,991</td>
</tr>
<tr>
<td># references</td>
<td>+500</td>
</tr>
</tbody>
</table>

Integration of metabolism with macromolecular synthesis

Conversion of integrated reconstruction of metabolism and macromolecular synthesis into a computational model

Canonical steady-state modeling

Implicit representation of an enzymatic reaction:

Explicit representation of an enzymatic reaction:

If metabolic reaction is used, then protein & mRNA need to be produced

If flux through metabolic reaction increases, the synthesis rate of protein and mRNA needs to increase accordingly

Conversion of integrated reconstruction of metabolism and macromolecular synthesis into a computational model

• Increasing scope of molecular processes represented

• However, molecular processes are intrinsically on different timescales scales...
Computational modeling requires numerical optimization involving large, sparse & stiff stoichiometric matrices: numerical analysis challenge

Many metabolic moieties in one macromolecule

Reaction rates over many orders of magnitude

~80,000 columns

~60,000 rows
Robust flux balance analysis of multiscale biochemical reaction networks

Yuekai Sun¹*, Ronan MT Fleming²,³, Ines Thiele²,³ and Michael A Saunders⁴

maximize \[ c^T \nu \]
subject to \[ Sv = 0, \]
\[ Cv \leq d, \]
\[ \nu_l \leq \nu \leq \nu_u, \]
\[ A + 10^9 B \rightarrow C + D \]
\[ 0.0001 \leq \frac{\nu_1}{\nu_2} \leq 10000. \]

Reformulation involves a trade off between computational efficiency and reliability.

Robust flux balance analysis of multiscale biochemical reaction networks

Yuekai Sun¹*, Ronan MT Fleming²,³, Ines Thiele²,³ and Michael A Saunders⁴

Table 1 FBA results for ME76664 before and after lifting

<table>
<thead>
<tr>
<th></th>
<th>Simplex</th>
<th></th>
<th>Barrier</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Iterations</td>
<td>48603</td>
<td>58288</td>
<td>56490</td>
<td>9985</td>
</tr>
<tr>
<td>CPU time</td>
<td>242</td>
<td>292</td>
<td>384</td>
<td>93</td>
</tr>
<tr>
<td>Infeasibilities</td>
<td>$1.3 \times 10^{-4}$</td>
<td>$2.9 \times 10^{-6}$</td>
<td>$1.4 \times 10^{-1}$</td>
<td>$3.4 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

FBA results for the *E. coli* Metabolic-Expression model ME76664 using CPLEX primal simplex and barrier solvers. Iterations, time, and sum of infeasibilities before and after lifting. The iterations in columns 4 and 5 include about 100 for the barrier solver and the remainder for the simplex crossover.
Robust flux balance analysis of multiscale biochemical reaction networks

Yuekai Sun¹*, Ronan MT Fleming²,³, Ines Thiele²,³ and Michael A Saunders⁴

Figure 2 Flux variability analysis of the E. coli Metabolic-Expression model. Minimum and maximum flux for iAF1260 (which only accounts for metabolic reactions) versus the minimum and maximum flux for the Metabolic-Expression model. Each colored box corresponds to a different reaction in metabolism. The boxes are always longer on the axis for the metabolic model (iAF1260) than on the axis for the Metabolic-Expression model. This demonstrates that increasing the comprehensiveness of the model toward whole cell modeling leads to a substantial shrinkage of the steady state solution space. (Fluxes are plotted in mmol · g⁻¹ · hr⁻¹).
New insights from multiscale systems biology models: 
*mechanics of the genotype-phenotype relationship*

- The genetic code has redundancy but no ambiguity
  1. can be multiple codons per amino acid
  2. multiple tRNA can read the same codon
  3. tRNA can read multiple synonymous codons
- Codon usage bias i.e. frequency of synonymous codons differs between organisms, within genomes, and along genes.

---

### New insights from multiscale systems biology models: mechanics of the genotype-phenotype relationship

**Genetic code of E. coli (RNA perspective)**

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU</td>
<td>Phe (2)</td>
<td>UCU</td>
<td>UAU</td>
<td>UGU</td>
</tr>
<tr>
<td>UU</td>
<td>Cys (1)</td>
<td>UGU</td>
<td>UGG</td>
<td>Trp (1)</td>
</tr>
<tr>
<td>UU</td>
<td>Ser (5)</td>
<td>UAC</td>
<td>Ochre</td>
<td>Opal</td>
</tr>
<tr>
<td>UU</td>
<td>Leu (8)</td>
<td>UAA</td>
<td>UGA</td>
<td>Amber</td>
</tr>
<tr>
<td>UU</td>
<td>Cys (1)</td>
<td>UGC</td>
<td>CGA</td>
<td>CGG</td>
</tr>
<tr>
<td>UU</td>
<td>Tyr (3)</td>
<td>UUC</td>
<td>CGU</td>
<td>Arg (7)</td>
</tr>
<tr>
<td>UU</td>
<td>Val (7)</td>
<td>GCC</td>
<td>GGU</td>
<td>Gly (6)</td>
</tr>
<tr>
<td>UU</td>
<td>Ala (5)</td>
<td>GCA</td>
<td>GGA</td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Glu (4)</td>
<td>GCG</td>
<td>GGG</td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>AAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Ser (5)</td>
<td>AGU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asp (3)</td>
<td>GAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>AAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Gln (4)</td>
<td>CAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>His (1)</td>
<td>CAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Pro (3)</td>
<td>CCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>AAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>CAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>ACA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>ACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Ser (5)</td>
<td>GGU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Ala (5)</td>
<td>GCU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>GAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Glu (4)</td>
<td>GAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Tyr (3)</td>
<td>UAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Pro (3)</td>
<td>UCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>ACA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>GAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asp (3)</td>
<td>GAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Gln (4)</td>
<td>CAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>His (1)</td>
<td>CCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Pro (3)</td>
<td>CCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>ACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>CAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asp (3)</td>
<td>GAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Tyr (3)</td>
<td>UAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Pro (3)</td>
<td>UCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>ACA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>CAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asp (3)</td>
<td>GAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Tyr (3)</td>
<td>UAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Pro (3)</td>
<td>UCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Thr (4)</td>
<td>ACA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asn (4)</td>
<td>CAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Asp (3)</td>
<td>GAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Arg (7)</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UU</td>
<td>Lys (6)</td>
<td>GCA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*E. coli* has 86 tRNA molecules.

Number of distinct tRNA molecules per amino acid are given in parenthesis.

- Leu = Leucine
  - 6 different synonymous
  - 8 different tRNA
  - In wild type *E. coli*, CUG is the dominant synonymous codon.

---

New insights from multiscale systems biology models: mechanics of the genotype-phenotype relationship

The biased strains were generated using the following algorithm:

**Input:** model, sequence for each gene in model, number of iterations m

**Output:** model_biased

**Algorithm:**

1. Choose randomly a codon, $c_1$
2. Identify possible synonymous codons: $c_s = \{c_1, c_{s2}, \ldots, c_{sk}\}$
3. Choose randomly one codon from $c_s$: $c_{si}$
4. Replace all instances of $c_1$ with $c_{si}$
5. Update ME-matrix for all genes based on new gene sequence:
   - (a) Transcription reactions.
   - (b) mRNA degradation reactions.
   - (c) Translation reactions (tRNA molecule will be updated based on codon recognition).
6. Repeat 1 through 5 m times, $m = 100$.

New insights from multiscale systems biology models: mechanics of the genotype-phenotype relationship

- Changes in codon usage affect
  - ability to grow
  - maximal possible growth rate in silico.
- Analysis of numerical properties of flux balance analysis solutions used to derive causal molecular mechanistic hypothesis connecting codon usage with growth rate
- Limit to growth was ribosomal RNA operon transcription rate in wild type, but leucyl-tRNA transcription rate in biased strains

**Glucose, Aerobic**

<table>
<thead>
<tr>
<th>Glucose, Aerobic</th>
<th>Lactate</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>B1</td>
</tr>
<tr>
<td>B2</td>
<td>B2</td>
</tr>
<tr>
<td>B3</td>
<td>B3</td>
</tr>
<tr>
<td>B4</td>
<td>B4</td>
</tr>
<tr>
<td>B5</td>
<td>B5</td>
</tr>
<tr>
<td>B6</td>
<td>B6</td>
</tr>
<tr>
<td>B7</td>
<td>B7</td>
</tr>
<tr>
<td>B8</td>
<td>B8</td>
</tr>
<tr>
<td>B9</td>
<td>B9</td>
</tr>
<tr>
<td>B10</td>
<td>B10</td>
</tr>
<tr>
<td>EQ1</td>
<td>EQ1</td>
</tr>
<tr>
<td>EQ2</td>
<td>EQ2</td>
</tr>
<tr>
<td>EQ3</td>
<td>EQ3</td>
</tr>
<tr>
<td>EQ4</td>
<td>EQ4</td>
</tr>
<tr>
<td>EQ5</td>
<td>EQ5</td>
</tr>
</tbody>
</table>

**CUG (LeuP/Q/T/V/W)**
- WT, B2, B9

**CUU (LeuU)**
- B1

**CUC (LeuU)**
- B6, B7, B8, B10

**UUA (LeuZ)**
- B3

**CUA (LeuW)**
- B4

**UUG (LeuX/Z)**
- B5

Increased tRNA demand may be met by augmenting supply
- e.g. modification of a tRNA to expand its set of read codons
- In *E. coli* MAS39, a second leucyl-tRNA (tRNA\text{leu}^W) is able to read CUU due to a uridine-5-oxyacetic acid modification.
- It remains to be experimentally established if *E. coli* MG1655 tRNA\text{leu}^W can also read CUU.

**Growth rate (relative to wild type)**

New insights from multiscale systems biology models: iterative annotation of gene function

(a & b) In vivo transcriptome measurements confirm the in silico transcriptomics predictions for differential expression of genes when growing on L-Arabinose or cellobiose minimal medium.

(c) scanning of promoter and upstream regions of essential genes identified high-scoring motifs

(d) scan of remaining genome for AraR motif identified different genes, within a single transcriptional unit, with sequences similar to
  • a sugar-binding protein for an arabinose ABC transporter
  • a permeases of an ABC transporter

• iterative workflow: reconstruction-> model-> prediction -> new annotation-> better model -> ...

• Metabolic models for increasing number of species
• We envisage a demand for integrated models of metabolism and macromolecular synthesis for all of these species
U01 Aims 2012-2017: Systems Biology Research Group, UCSD

Prototyping reconstruction & validation procedure on individual organisms, e.g. *E. coli* & *T. maritima*

Development of software workflows for multiscale model reconstruction and systematic validation using transcriptomic data

- Develop software for reconstruction of biochemical networks spanning multiple cellular subsystems.
- Develop techniques for quantitative prediction and validation of transcript abundance in an integrated model of metabolism, macromolecular synthesis and regulation
  - Comparison with new experimental data
  - e.g. for *Geobacter* spp.
    - biogeochemical cycling of carbon and metals
    - bioenergy applications.


- Development of quad precision versions of large-scale LP/QP/NLP solvers (SQOPT, SNOPT), and their linear sparse-matrix algebra ‘engine’ (LUSOL)
  - re-implement in Fortran 2003 with quad precision variables
  - redesign of key components, e.g., storage allocation
- Software: www.stanford.edu/group/SOL/multiscale/software.html

<table>
<thead>
<tr>
<th>Decade</th>
<th>Variables</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>single</td>
<td>single</td>
</tr>
<tr>
<td>1b</td>
<td>single</td>
<td>double</td>
</tr>
<tr>
<td>2a</td>
<td>single</td>
<td>single</td>
</tr>
<tr>
<td>2b</td>
<td>double</td>
<td>double</td>
</tr>
<tr>
<td>3a</td>
<td>single</td>
<td>single</td>
</tr>
<tr>
<td>3b</td>
<td>double</td>
<td>double</td>
</tr>
<tr>
<td>3c</td>
<td>quad</td>
<td>quad</td>
</tr>
</tbody>
</table>

Table 2: History of Scientific Computing

Increasing the reliability, while maintaining efficiency of numerical optimization solvers.
Increasing size of *E. coli* reconstructions

- I: Transcription, II: mRNA degradation;
- III: translation;
- IV: protein maturation,
- V: protein folding; VI: metallo-ion binding;
- VII: protein complex formation;
- VIII: ribosome assembly;
- IX: RNA processing; X: rRNA modification;
- XI: tRNA modification; XII: tRNA charging

- compartmentalized reconstruction (distinct periplasm)
- extensive cell wall metabolism (phospholipids, murein, LPS)
- reaction thermodynamics

- alternate carbon utilization
- quinone characterization
- elemental and charge balancing

- fatty acid metabolism
- expanded cellular transport systems
- used genome as a scaffold

- cell wall constituent biosynthesis
- cofactor biosynthesis
- growth-dependent biomass objective function

- amino acid and nucleotide biosyn.

---


- Development of hypergraph network flow algorithms for optimization with biological networks
  - Standard network flow algorithms
    - Complexity $O(n)$
    - Designed specifically for graphs, whereas biological networks are hypergraphs
  - Standard linear optimization algorithms
    - Complexity $O(n^3)$
    - Used for, amongst many other things, optimization over biochemical networks
      - e.g. Flux Balance Analysis
- Open research question: Does there exist an algorithm of lower computational complexity than a standard linear optimization solver, specifically for biochemical hypergraph flow problems?

$$\begin{align*}
\text{maximize} & \quad c^T v \\
\text{subject to} & \quad Sv = 0, \\
& \quad Cv \leq d, \\
& \quad v_l \leq v \leq v_u,
\end{align*}$$
Biochemical reaction network models are increasing in scope

However, the currently available approaches for large scale modeling of biochemical reaction networks only explicitly represent reaction flux, not molecular abundance.

Open question: does there exist a numerically scalable approach to model fluxes and concentrations explicitly?
Consider a biochemical network with \( m \) molecular species and \( n \) reversible chemical reactions.

Define forward and reverse stoichiometric matrices, \( F, R \in \mathbb{Z}_{\geq 0}^{m,n} \), respectively, where \( F_{ij} \) denotes the stoichiometry\(^1\) of the \( i^{th} \) molecular species in the \( j^{th} \) forward elementary reaction and \( R_{ij} \) denotes the stoichiometry of the \( i^{th} \) molecular species in \( j^{th} \) reverse elementary reaction.

We assume that every elementary reaction conserves mass, that is, there exists at least one positive vector \( l \in \mathbb{R}^m_{\geq 0} \) satisfying \((R-F)^T l = 0\) where \( R - F \) represents net reaction stoichiometry.

Let \( c \in \mathbb{R}^m_{>0} \) denote a variable vector of molecular species concentrations.

Assuming constant non-negative elementary kinetic parameters \( k_f, k_r \in \mathbb{R}_{\geq 0}^n \), we assume elementary reaction kinetics for forward and reverse elementary reaction rates as \( v_f(k_f, c) \equiv \exp(\ln(k_f) + F^T \ln(c)) \) and \( v_r(k_r, c) \equiv \exp(\ln(k_r) + F^T \ln(c)) \), respectively.
The deterministic dynamical equation for time evolution of molecular species concentration may then be expressed as

\[
\frac{dc}{dt} = (R - F)(v_f(k_f, c) - v_r(k_r, c)),
\]

\[
= (R - F)(\exp(\ln(k_f) + F^T \ln(c)) - \exp(\ln(k_r) + F^T \ln(c))) \equiv v(c)
\]

Assuming a non-equilibrium steady state, \( \frac{dc}{dt} = 0 \) and \( v_f(k_f, c) \neq v_r(k_r, c) \), one is then interested in the nonlinear, nonconvex set of steady state molecular species concentrations \( \{c | v(c) = 0\} \).

- **Modeling challenge**
  - High dimensional
    - many molecular species and reactions
  - Nonlinear
    - nonlinear relationship between molecular species abundance and reaction rate
  - Need algorithms with low polynomial time complexity, guaranteed convergence, certificate of infeasibility
    - mathematical model & algorithm formulation problem
  - Multiscale
    - molecular species concentrations vary over many orders of magnitude
      - e.g. transcript abundance versus metabolite abundance
  - Paucity of kinetic parameters
Nonlinear relations between reaction rates and metabolite concentrations satisfied at optimum of a convex optimization problem

Amenable to solution with polynomial time algorithms
A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks

R.M.T. Fleming\textsuperscript{a,*}, C.M. Maes\textsuperscript{b}, M.A. Saunders\textsuperscript{c}, Y. Ye\textsuperscript{c}, B.Ø. Palsson\textsuperscript{d}

\textbf{A B S T R A C T}

We derive a convex optimization problem on a steady-state nonequilibrium network of biochemical reactions, with the property that energy conservation and the second law of thermodynamics both hold at the problem solution. This suggests a new variational principle for biochemical networks that can be implemented in a computationally tractable manner. We derive the Lagrange dual of the optimization problem and use strong duality to demonstrate that a biochemical analogue of Tellegen’s theorem holds at optimality. Each optimal flux is dependent on a free parameter that we relate to an elementary kinetic parameter when mass action kinetics is assumed.

\textbf{Theorem 1.} Let $v_\ast$ be any set of optimal exchange fluxes from problem (FBA). Define $b = -S_e v_\ast$, and let $c$ be any vector in $\mathbb{R}^n$. The convex equality-constrained problem

\[
\begin{align*}
\text{minimize} & \quad \phi \equiv v_f^T(\log(v_f)+c-e) + v_r^T(\log(v_r)+c-e) \\
\text{subject to} & \quad Sv_f - Sv_r = b : y
\end{align*}
\]  

(EP)

is then feasible, and its solution $(v^\ast_f, v^\ast_r)$ is a set of thermodynamically feasible internal fluxes. The combined vector $(v^\ast_f, v^\ast_r, v^\ast_e)$ is thermodynamically feasible and optimal for problem (FBA). The associated chemical potentials $u$ may be obtained from the optimal Lagrange multiplier $y^\ast \in \mathbb{R}^m$ for the equality constraints according to $u = -2\rho y^\ast$. 

- Constraint on ratio of concentrations, not absolute concentration.
- Biochemical reaction directions are an evolved subset of thermodynamically feasible directions.
Consistent Estimation of Gibbs Energy Using Component Contributions

Elad Noor¹, Hulda S. Haraldsdóttir²,³, Ron Milo¹*, Ronan M. T. Fleming²,³*

A

B

C

\[ \exp\left(-\frac{S_j^T \cdot u^o}{RT}\right) = \frac{k_{f,j}}{k_{r,j}} \]

\[ \text{ATP}^3^- + \text{FGAM}^2^- + \text{Gln} + \text{H}_2\text{O} \rightleftharpoons \text{ADP}^2^- + \text{P}^2^- + \text{FPRAM}^2^- + \text{Glu}^1^- + 2\text{H}^+ \]

\[ \Delta G^o = 40.9 \text{ kJ/mol} \]

\[ \text{ATP}^3^- + \text{Gln} + \text{H}_2\text{O} \rightleftharpoons \text{ADP}^2^- + \text{P}^2^- + + \text{Glu}^1^- + 2\text{H}^+ \]

\[ \Delta G^o = -164.0 \text{ kJ/mol} \]

\[ \text{FGAM}^2^- \rightleftharpoons \Delta G^o = 204.9 \text{ kJ/mol} \]

\[ \text{FPRAM}^2^- \]

Figure 3. A diagram illustrating how the component contribution method projects the stoichiometric vector onto the different spaces. (A) The reaction vector \( x \) is decomposed into the two components \( x_R \) and \( x_N \), where the reactant contribution and group contribution methods are used for the relevant components. Later, \( x_N \) is decomposed into \( x_{NR} \) and \( x_{NN} \). The same projection is shown graphically in (B) where the green plane represents the range of \( S \) and the normal to that plane represents the null space of \( S^T \). (C) An example for a reaction which decomposes into two non-zero components. In this case, the component \( x_{NN} \) is equal to 0, which means that the reaction is covered by the component contribution method.
Mass conserved elementary kinetics is sufficient for the existence of a non-equilibrium steady state concentration

R.M.T. Fleming\textsuperscript{a,b,\ast}, I. Thiele\textsuperscript{a,c}

When does there exist a non-equilibrium steady state concentration?

\textbf{Theorem 1.} Let the dynamical equation for mass conserved elementary kinetics be

\[
\dot{c} \equiv \frac{dx}{dt} = S \cdot (K_f \cdot \exp(F^T \cdot \ln(c))) - K_r \cdot \exp(R^T \cdot \ln(c)),
\]

where $c \equiv c(t) \in \mathbb{R}^m$ is the molecule concentration at time $t > 0$, $\dot{c} \in \mathbb{R}^m$ is the time derivative of concentration, $K_f = \text{diag}(k_f)$, $K_r = \text{diag}(k_r)$ and $k_f, k_r \in \mathbb{R}^{n \times n}_{\geq 0}$ are non-negative forward and reverse kinetic parameters. $F, R \in \mathbb{R}^{m \times n}_{\geq 0}$ are forward and reverse stoichiometric matrices. $S \equiv -F + R$ is a consistent stoichiometric matrix defined by the existence of at least one strictly positive vector $l \in \mathbb{R}^m_{> 0}$, such that $S^T \cdot l = 0$. Assuming a finite and strictly positive initial concentration $c_0 \equiv c(0) \in \mathbb{R}^m_{> 0}$, then there exists at least one finite and non-negative steady state concentration $x^*_\geq 0$, such that $\dot{c} = 0$. 

• Scalable algorithms for multiscale reconstruction and modeling
  • Software to enable high fidelity reconstruction of biochemical networks
    • e.g. checking for consistency with known biochemistry
    • e.g. suggesting extension to existing reconstruction to account for known biochemical function
  • **Multiscale mass conserved elementary kinetic modeling**
    • **Forward problem**
      • given kinetic parameters, compute a non-equilibrium steady state
    • **Inverse problem**
      • given reaction stoichiometry, experimental boundary conditions thermodynamic constraints and net reaction directions consistent with biochemistry, search among consistent kinetic parameters
      • Gradient-based search of kinetic parameters in multiscale models.
        • existing algorithms do not search for kinetic parameters using gradient based methods due to a perception that the merit function for such a problem contains local minima
        • High risk, high gain project: does a formulation of the problem exist which is amenable to solution with a polynomial time algorithm guaranteed to reach a global optima?
Sharing models

• Network reconstructions used to generate computational models are always made available with the accompanying paper
  – Systems Biology Markup Language (SBML)
    • all monoscale models
    • for multiscale models, standardised representation that is scalable needs to be developed.
    • multiscale models still distributed, but representation not yet standardised.

http://systemsbiology.ucsd.edu/Downloads
http://thielelab.eu/
http://www.stanford.edu/group/SOL/multiscale/models.html
Open source software

- **rBioNet** - a COBRA toolbox extension for reconstructing high-quality biochemical networks (*Thorleifsson & Thiele, Bioinf, 2011*)

- **von Bertalanffy 1.0** - a COBRA toolbox extension to thermodynamically constrain metabolic models (*Fleming & Thiele, Bioinf, 2011*)

- **COBRA Toolbox v2.0** - quantitative prediction of cellular metabolism with constraint-based models (*Schellenberger et al, Nat Protoc, 2011*)

- **COBRApy** - COnstraints-Based Reconstruction and Analysis for Python (*Ebrahim et al, BMC Syst Biol, 2013*)

- **fastFVA** – a tool for computationally efficient flux variability analysis (*Gudmunsson & Thiele, BMC Bioinf, 2010*)

- **robustFBA** - Robust flux balance analysis of multi-scale biochemical reaction networks (*Sun et al, BMC Bioinf, 2013*)

See [http://www.stanford.edu/group/SOL/multiscale/software.html](http://www.stanford.edu/group/SOL/multiscale/software.html) for links to software
Open source software continued:

- **LUSOL, LUMOD** Routines for dense and sparse LU factorization.

- **PDCO** A primal-dual interior point method for large-scale optimization with convex objective and linear constraints.

- **PNOPT** Proximal Newton-type methods for minimizing composite functions (unconstrained optimization of the sum of smooth and nonsmooth functions).

- **Need help with cobra methodology?**
  - openCOBRA Google group
    - [https://groups.google.com/forum/#!forum/cobra-toolbox](https://groups.google.com/forum/#!forum/cobra-toolbox)
    - >2009
    - 430 posts
    - 300 members
    - Anyone can view content.
    - Anyone can apply to join.

See [http://www.stanford.edu/group/SOL/multiscale/software.html](http://www.stanford.edu/group/SOL/multiscale/software.html) for links to software
Thanks

• Colleagues in the Multiscale Systems Biology Collaboration

• Funding 2012-2017
  – National Institutes of Health (NIH) through the Multiscale Modeling Physiome Initiative, grant 1U01GM102098-01.
  – Department of Energy (DOE) through the Scientific Discovery Through Advanced Computing Program, grant DE-SC0010429.

• Program coordination
  – Interagency Modeling and Analysis group