Computational Challenges in Large-Scale Pathway Modeling

Frank Tobin
Scientific Computing and Mathematical Modeling
GlaxoSmithKline

September 22, 2004
Agenda

- **Biological pathways**
  - simple example of a pathway
  - simple example of pharmaceutical interest

- **Building a mathematical model of biological networks**

- **Computational challenges**
Motivation

• Build as complete a model of as much of a cell or organism as possible
  ♦ E. coli is the archetypical prototype

• Figure out what to do with it once we get it
  What if we had a perfect model? Then what?
What is a Pathway?

For the purposes of this talk:

A network of interaction biological entities represented as a directed graph.

So network and pathway are equivalent under this definition.
Pharmaceutical Interest in Pathways

- Predicting culture conditions for overproduction of biopharmaceuticals and drug targets, bioengineering of target assays, enzymes, receptors, etc.
- Understanding compound modes of action
- Identifying novel behaviors and new behaviors of known pathways
  - clues to new intervention approaches
  - selecting and prioritizing of new targets
- Identifying and validating bio-markers
  - animal ⇔ human correlation
- Interpreting and integrating system biology data:
  - transcriptomics, proteomics and metabolomics and other ‘omics’
A Simple Pharmaceutical Pathway Example

- Risperidone is a psychotropic agent used for treating schizophrenia or psychosis
- 2.1% of patients develop extrapyramidal symptoms:
  - involuntary movements
  - tremors and rigidity
  - body restlessness
  - muscle contractions
  - changes in breathing and heart rate
- Hypothesis for the extrapyramidal symptoms:
  Dopamine receptor antagonism
Receptor Binding:
Formation of active complex:
Risperidone conversion to 9-hydroxyrisperidone
Binding to D₂ and 5-HT₂ receptors

DA + D₂ ⇔ DA•D₂
DA•D₂ + T ⇔ DA•D₂•T
R → OH
R + D₂ ⇔ R•D₂
R + HT₂ ⇔ R•HT₂
OH + D₂ ⇔ OH•D₂
OH + HT₂ ⇔ OH•HT₂

DA: Dopamine  
D₂: Receptor  
T: Transmitter  
R: Risperidone  
OH: 9-hydroxyR  
HT₂: Receptor

Missing from Yamada Model
Incorrectly specified in Yamada
Non-antagonized system
Risperidone dosing and clearance
Risperidone metabolism
Risperidone antagonism
Yamada model for Risperidone PK


1-compartment PK model for Risperidone concentration

\[ c(t) = A(c_0, k_a, k_{cl}) \left[ \exp(-k_{cl} t) - \exp(-k_a t) \right] \]
The ODE Model Approach

**Biological model**

\[
\mathbf{x}' = \mathbf{f}(\mathbf{x}, \lambda) + D(t)
\]

**Mathematical model**

**Numerical simulations**

**ODEs**

\[
\begin{align*}
\frac{d[R]}{dt} &= -k_1^R [R] \\
\frac{d[OH]}{dt} &= -k_1^{OH} [OH] \\
\frac{d[R]}{dt} &= k_1^R \cdot [R] \\
\frac{d[OH]}{dt} &= k_1^{OH} \cdot [OH] \\
\frac{d[DA]}{dt} &= k_2^{DA} \cdot [DA][D_2] - k_3^{DA}[DA] \\
\frac{d[DA_2]}{dt} &= k_2^{DA_2} \cdot [DA_2][T] - k_3^{DA_2}[DA_2] \\
\frac{d[R_2]}{dt} &= k_2^{R_2} \cdot [R_2][D_2] - k_3^{R_2}[R_2] \\
\frac{d[OH_2]}{dt} &= k_2^{OH_2} \cdot [OH_2][D_2] - k_3^{OH_2}[OH_2] \\
\frac{d[R_2T]}{dt} &= k_2^{R_2T} \cdot [R_2][HT_2] - k_3^{R_2T}[R_2] \\
\frac{d[OH_2T]}{dt} &= k_2^{OH_2T} \cdot [OH_2][HT_2] - k_3^{OH_2T}[OH_2]
\end{align*}
\]

**Dosing**

**Clearance**

**GlaxoSmithKline**
# Daily Dosing Differs from a Single Dose

## Plasma Concentration

**Plasma conc. (ng/ml)**

<table>
<thead>
<tr>
<th>time (h)</th>
<th>0</th>
<th>24</th>
<th>48</th>
<th>72</th>
<th>96</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average OH conc.**

**Average R conc.**

**OH simulation**

**R simulation**

**OH - single dose**

**R - single dose**

**daily dose**

---

R, OH exptl data from Ishigooka et al., Clin Eval 19, 93-163 (1991)
Effect of multiple dosing on receptor occupancy

![Graph showing the effect of multiple dosing on receptor occupancy](image)

- **Mean receptor occupancy (%)**
- **Dose (mg/day)**
- **Therapeutic Range**

The graph compares the mean receptor occupancy for single dose (1dose) and five doses (5 doses) for 5-HT₂ and D₂ receptors. The therapeutic range is indicated by the shaded area.
Daily dosing causes differences in predicted side-effects

Multiple dosing results in increased ESRS shift, increasing with daily dose administered
Receptor Occupancy as a function of cumulative dosing

**D₂**

Cumulative changes in occupancy

First 24 hours identical between single and multiple doses

**5-HT₂**

Occupancy

Only R

R + OH

Scientific Computing and Mathematical Modeling
Real Pathways are More Complex
Mathematical Complexity

- Consider a small, relatively unsophisticated bacterium: Escherichia coli
  - \( \approx 2000 \) genes
  - \( 2500 \) proteins
  - at least several hundred small molecules
  - \( 3 \) interactions per entity \( \times 5000 \) entities
  - \( 3 \) parameters per equation
  - \( \approx 15\,000 \) equations with \( 45\,000 \) parameters!

\[
X' = F(X; \lambda) \quad \text{continuous, discrete, stochastic}
0 = G(X; \lambda) \quad \text{analytic constraints}
0 = H(X; \lambda) \quad \text{non} - \text{analytic constraints}
\]

- Now add on spatial change - 15 000 PDEs!
The Modeling Process

1A Building the model -- forward problem
- Static
- Kinetic
  - Rate law determination
  - Parameter determination

1B Reconstructing the model -- inverse problem

2 Validating the model
- Experimental data comparison
- Plausible biology from analytic analysis/simulation
- Examining and assertions testing results

3 Simulation
- Hypothesis testing
- Hypothesis generation
**Static Model**

- Only connectivity (topology) of the interactions
- Visualised as connection or interaction graph
- Used for initial model verification and testing
- Types
  - Metabolic
  - Gene Regulation
  - Gene-Product, and Protein-Protein Interactions

### Metabolic network

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucose</td>
<td>HPr-P</td>
</tr>
<tr>
<td>2.7.1.69</td>
<td>HPr R12</td>
</tr>
<tr>
<td>D-glucose 6-phosphate</td>
<td>ATP</td>
</tr>
<tr>
<td>5.3.1.9</td>
<td>ADP</td>
</tr>
<tr>
<td>2.7.1.11</td>
<td>D-fructose 1,6-bisphosphate</td>
</tr>
</tbody>
</table>

### Gene-Product interactions network

### Genetic network

- **Repressor**
- **Activator**
- **OPERON**
Kinetic Model

- **First phase:** Kinetic models - time dependency incorporated
  - Kinetic behaviour (rate laws) added to static model
  - May or may not obey mass action kinetics

- **Second phase:** Kinetic constants determined from experimental data

- **Third phase** Mathematical model - equations generated
  - Time variation of all concentrations and fluxes can be simulated
  - Model analyses possible: sensitivity, linear stability theory, asymptotic analysis, etc.

**Example: Inhibition of a Ligand-Receptor Complex Formation**

### Static model
- Ligand
- Receptor
- Inhibitor

### Kinetic Model
- \( R + L \leftrightarrow R \cdot L \)
- \( R + I \leftrightarrow R \cdot I \)

### Mathematical Model

\[
\begin{align*}
[R]' & = -k_1[R][L] + k_2[RL] - k_3[R][I] + k_4[RI] \\
[RL]' & = k_1[R][L] - k_2[RL] \\
[RI]' & = k_3[R][I] - k_4[RI] \\
[L]' & = -k_3[R][L] + k_2[RL] \\
[I]' & = -k_3[R][I] + k_4[RI] \\
L_0 & = [L] + [RL] \\
I_0 & = [I] + [RI] \\
R_0 & = [R] + [RL] + [RI]
\end{align*}
\]

### Numerical Simulation
The Resulting System
Very Large, Flawed, and Damned Useful!

- The resulting system of equations:
  \[ x' = F(x, l) \]
  \[ 0 = G(x, l) \] algebraic relationships
  \[ 0 = H(x, l) \] analytic constraints
  \[ 0 = I(x, l) \] non-analytic constraints

- Very large dimensionalities in:
  - the number of species, \( X \)
  - the number of interactions
  - the number of parameters, \( \lambda \)
  - the number of constraint equations

- Uncertainty, error, ambiguity, approximations, etc
As the pathways grow large, the nature of the problems change.

- **Building the model**
  - knowledge management
  - knowledge updating
  - incomplete knowledge
  - **Automation**
  - Updating the model - versioning

- **Analysis of the model**
  - Too much for a human to peruse
  - Theory gaps
  - **Automation**

- **Analysis of the simulation results**
  - Too much for a human to peruse
  - New techniques
  - **Automation**
Automation

- No human intervention whatsoever
  - None, nada, zip!
  - If it takes a human to setup, run or analyze - its not automated

- Robust algorithms
  - Graceful failure
  - Knowledge of domain of applicability
  - Pathological data happens very often - Murphy is omnipresent

- Not as easy as it main seem at first

- Many existing algorithms are not automatable in current usage
The Model Understanding Roadmap

Biology

Static model
Kinetic model
Dynamics model

Analysis
graph theory
analytic theory
analytic theory

Understanding

New experiments
Find model errors
“Gaps”

Simulations

Analyzing Results

Automate
Exhaustive Analysis
Theory Gap for Large Systems

- Large but not infinite dimensionality is the problem
- Analytical and numerical determination:
  - Finding ‘true’ null states - there may be a great number
  - Finding linear null states - there may be a great number
  - Asymptotic behaviors
  - Controllability, predictability, integrability, ...
  - Steady state, non-linear behaviors
  - Bifurcation analyses
  - Perturbed behaviors - drug dosing, environment, mutants, etc.
  - ...
- How to calculate in a computationally efficient manner
- Can’t afford to calculate everything
- Need to a priori determine which are to be done
Continuousness / Stochasticity / Discreteness / Ambiguity

- Continuous approximation breaks down
  - Need to use master equations or some other form of involving stochasticity
  - May need to dynamically switch as system evolves
- Some processes are truly discrete
  - Consider cellular automatons, Petri Nets, discrete events, etc.
- Some parts of the model are only known qualitatively
  - Qualitative simulation techniques.
- Uncertainty and variation in the system
  - Initial conditions
  - Rate constants and rate laws
  - Population variations
  - Interval or fuzzy integration
- Multiscale - time, length, concentration, etc.
- Constraints - DAEs

The challenge: one hybrid integrator
Parameter challenges

• The larger the model:
  ♦ the more parameters compared to the experiments

• Static guessing - filling in the gaps
  ♦ guessing gene function by analogy
  ♦ looking for missing reactions - i.e. enzyme

• Kinetic guessing - integrating kinetic islands - guessing plausible rate laws and parameters
  ♦ Analogy approaches, similarity across species (‘multiple alignment’)
  ♦ From flux analysis?

• Do we need to know all parameters? Accuracy?
Parameter challenges

- Determine parameters of rate laws from an optimization to fit experimental kinetics data
  - noisy and incomplete data
  - ill-posed, possibly severely

- How do we scale this up as the model gets bigger?
  - One huge model fitting? - Can we even afford this approach?
  - One sub-systems at a time fitting?
  - Hierarchical fitting? - Stitching together pieces individually calibrated does not a priori mean the model is calibrated

- What’s the best way to optimize?
  - Is $L_2$ the best objective function?
  - Constraints - incorporating and coming up with better ones

- How do we know how well we’ve done?
Inverse Problems and Biological Plausibility

• What makes a model more biological than another?
  ♦ thermodynamic constraints
  ♦ numerical integrity - semi-definite solutions
  ♦ asymptotic behaviors
  ♦ stability properties
  ♦ information theory constraints
  ♦ physico-chemical constraints
  ♦ environmental constraints
  ♦ evolution constraints
  ♦ flux distributions
  ♦ mass and energy balance

• Parameter determination needs also
Visualization Challenges

- Visualization in a large graph with *too much detail*
  - Analysis of results - what’s interesting?
  - Drill down, hyperbolic viewers, database driven for large models
  - Visualizing fluxes in a meaningful way

- How do you visualize huge networks?

- Tools needed for panning, zooming, drill-down, scalable, incrementally updatable from a database, etc.
- Pathway editors for input
- Animation - visualizing temporal fluxes

Experiments By T. Munzer, UBC, for visualizing Web connections
Discovering “New” Biology

Assumption: if we didn’t know anything any biology per se, could we rediscover it from the model?

Caveat: if we can find “old” biology, then presumably we could find “new” biology
Discovering “New” Biology

- Finding new cooperative or emergent phenomena:
  - pathways and “distinguishable” sub-systems
  - cycles and “clocks”
  - oscillatory systems
  - regulatory systems
  - “states” or “modes” of the system

- The resulting biology acts as plausible checks on the model

- Some ideas:
  - Persistent - pathway behavior is or is not independent of initial conditions
  - Conditional - pathway is active only for certain initial conditions - the nub of course is how to identify this
  - Model ⇒ graph ⇒ matrix ⇒ permutation matrix reordering ⇒ structure ⇒ biology?
  - Pattern recognition approaches. Model comparison? Different organisms/species?
  - Some type of flux or domain decomposition?
How do you know they’re right? Assertions checking

• Provide a means to formally represent biology that went into the model
  ♦ aspects of computer language parsing, AI-knowledge representation, inference

• Purposes
  ♦ as a formal computer language for incorporation into software
  ♦ for automation of the biology knowledge comparisons against data
  ♦ allow checking model accuracy
  ♦ used as criteria for optimisation - e.g. parameter determination of rate laws

• Consequences of the assertions
  ♦ require certain behaviours to be present in the model
  ♦ expect, but not require some behaviours
  ♦ search for speculative behaviours
  ♦ provide diagnostic tools for examining the quality of the data
**Assertions - Bacterial Aerobicity Example**

Different genes are expressed under different environments conditions - temperature, media composition, pH, and oxygen. Regulatory systems control expression, but assertions can be used to ensure the basic regulatory processes of the model are accurate.

# Find the time when the system changes from anaerobic to aerobic behaviour and then
# make sure that the key regulations appear to be happening

\[
\text{Regulation\_time} = \text{time} > \text{change\_time('ANAEROBIC', 'AEROBIC')} \\
\text{AND} \quad '\text{ArcA-P} >> '\text{ArcA}' \quad \# \text{positive regulation (activation) of ArcA by ArcA-P} \\
\text{OR} \quad '\text{FNR-ox} >> '\text{FNR-red}' \quad \# \text{FNR repressed aerobically}
\]

#Then, if regulation appears to be happening, for each protein behaving aerobically:

\[
\text{ForEach aerobic\_protein=aerobic(*)} \\
\{
\quad b = \text{flux\_value(aerobic\_protein)}; \\
\quad c = \text{gene\_name(aerobic\_protein)}; \\
\quad \text{if (regulation\_time AND (b > 0))}
\quad \{
\quad \quad \text{Success Action:}
\quad \quad \quad \text{Message ('AerobicityState' confirmed by the expression profile of gene %s',c)}
\quad \quad \text{Failure Action:}
\quad \quad \quad \text{Message ('Gene %s does not have the expected 'AerobicityState' expression pattern',c)}
\quad \quad \text{Status = WARNING} \quad \# \text{indicate a non-fatal problem}
\quad \}
\}
\]
The Pathway Modeling Factory Concept
What Else Is There? Much, Much More!

Only limited by our imaginations

"Come on, people. We need a creative epiphany right now. Who has one?"
Acknowledgements