Lecture 1: Biomolecular Modeling

Richard M. Murray
Caltech CDS/BE

Goals:
- Motivate the role of dynamics and feedback in the cell
- Provide mathematical tools for analyzing and predicting the dynamics of transcriptional regulation in the cell
- Work through case study for the lac operon (model control system)

References
**Architecture of the Cell**

**Prokaryote (bacterium)**
- Simplest type of cell: cell wall, cytoplasm, flagella/pili
- DNA is circular, single copy
- Simple cell division (mitosis): DNA duplicated in each daughter cell
- Examples: *E. coli, Bacillus subtilis*
  - *E. coli*: 1 µm long, 20 min cycle

**Eukaryote: single and multi-cell**
- Much more complicated structure
- DNA resides in nucleus; proteins are formed in the cytoplasm
  - DNA organized in chromosomes
- Organelles implement cell functions
  - Nucleus: contains DNA
  - Mitochondria: energy production

http://en.wikipedia.org/Cell (biology)
Central Dogma: DNA to Proteins

Transcription: DNA to mRNA
- Double stranded DNA contains nucleotide sequence (A, C, T, G) on a sugar (deoxyribose) backbone
- Watson-Crick pairing: A:T, C:G
- RNA polymerase transcribes DNA sequence to RNA sequence (A, C, U, G sequence on a ribose backbone)

Translation: mRNA to protein
- (Eukaryotes) pre-mRNA is spliced and processed to form mature mRNA
- mRNA is translated by ribosomes into a chain of amino acids using the genetic code (3 bp code for 1 aa)
- Amino acid chain (polypeptide chain) folds into a protein

Regulation: control gene expression
- Proteins bind to DNA, RNA and proteins to modulate expression
Model Control System: The Lac Operon

Key idea (Jacob & Monod, 1950s): produce proteins when you need them
- Bacterial growth dependent on nutrient environment
- Lactose only consumed if glucose is not present
- Q: How does E. coli decide when to produce proteins?
  - A: Lac “operon” (≈ control system)
  - KPT07: “Lac operon is the hydrogen atom of regulation”

How does it work?
- Proteins for digesting lactose are controlled by binding of repressor and activator
- CAP (activator): recruits RNA polymerase when bound to cAMP, whose concentration is controlled by absence of glucose (+glucose ⇒ no cAMP)
- Repressor: blocks transcription by causing DNA looping unless it is bound to allolactose, a byproduct of lactose metabolism
- Feedback: if lactose present, then create proteins required to turn on lac operon, which creates the proteins required to metabolise lactose
**E. Coli Genome**

**DNA facts**
- Typical gene is about 1000 bp (300 aa) long
- E. coli genome is 4.6 Mb, with ~4000 genes
- Human genome is about 4 Gb, with ~25,000 genes

**Lac control circuitry**
- Main circuitry is about 6000 bp (genes plus regulatory regions)
- Can see structure of the promoter in DNA sequence

Phillips, Kondev, Theriot (2008)
Operon Layout and Molecular Census

The regulatory landscape

- Repressor binds in a region that blocks the binding of RNAP
- If CAP is present (activated by cAMP, which is present in the absence of glucose), it recruits RNAP
- Can assign energies to each of these events and work out statistical mechanics (see PKT)

Phillips, Kondev, Theriot (2008)
Repressor and DNA looping

Repressor acts through DNA looping

- Looking can occur between binary binding site (O1) and secondary sites on DNA (O2, O3)

\[
\text{Repression} = \frac{p_{\text{bound}}(R = 0)}{p_{\text{bound}}(R)}
\]

- Modeling via statistical mechanics: compute \( p_{\text{bound}} \) of RNA binding to DNA for different situations
- Experiments: relocate secondary binding sites and see what happens
- Still get some repression even if there is only one operator site present (no looping?)

http://www.ks.uiuc.edu/Research/vmd/

Phillips, Kondev, Theriot (2008)
Rates of Transcriptional Regulation

Primary timescales:
- mRNA production: 10-30 bp/sec
- Protein production: 10-30 aa/sec
- Protein folding: (depends)

DNA, protein from PKT08, mRNA production from Vogel & Jensen

Other important rates
- mRNA half life: ~100 sec
- Protein half life: ~5 x 10^4 sec
- Protein diffusion (along DNA): up to 10^4 bp/sec
- Typically assume that activators and repressors reach equilibrium state much more quickly than other time scales

Half life times from Yildirim and Mackey, 2003; Protein diffusion from Blainey et al, PNAS 2006.

Phillips, Kondev, Theriot (2008)
From Numbers to Equations: Dynamical Modeling

Modeling philosophy: Ask questions first, build model later
- Many different models possible for the same system; no such thing as “the model”
- The model you use depends on the questions you want to answer
- Never build a model without first posing the questions

Analysis and design based on models
- A model provides a prediction of how the system will behave
- Feedback can give counter-intuitive behavior; models help sort out what is going on
- Models don’t have to be exact to be useful; they just need to help explain (and predict)
Dynamic Modeling Approaches

Possible approaches to modeling

- Molecular dynamics - keep track of vibration of molecules and detailed reaction dynamics
- Monte Carlo/Stochastic simulation - extend ideas from statistical mechanics to include time
- Continuum/partial differential equations - keep track of evolution in space and time
- Reduced order models - ordinary differential equations that capture bulk properties

Choice of model depends on the questions you want to answer

- Modeling takes practice, experience and iteration to decide what aspects of the system are important to model at different temporal and spatial scales
- Good analysis models make testable predictions and produce “surprising” results
- Good design models capture enough of the important behavior to give allow good design decisions
Statistical Mechanics Viewpoint

 Provides an *equilibrium* view of the world

- Computes the “steady state” probability that a situation occurs
- Based on energy arguments; allows study of complex situations

**Partition function**

- Sum up the weights for desired states (bound) versus all possible states

\[
\sum_{\text{states}} \left( \begin{array}{c} \text{states} \\ \text{states} \end{array} \right) + \sum_{\text{states}} \left( \begin{array}{c} \text{states} \\ \text{states} \end{array} \right)
\]

\[
P_{\text{bound}} = \frac{\sum_{\text{states}} \left( \begin{array}{c} \text{states} \\ \text{states} \end{array} \right)}{\sum_{\text{states}} \left( \begin{array}{c} \text{states} \\ \text{states} \end{array} \right) + \sum_{\text{states}} \left( \begin{array}{c} \text{states} \\ \text{states} \end{array} \right)}
\]
The Master Equation: Detailed Events and Rates

Key idea: transition rates between microstates

- Enumerate micro-states corresponding to the system of interest
- Define the system “state” in terms of the individual probabilities of each microstate at each instant in time
- Dynamics are given by rate of change of the probability of each individual state

Transition rates depend on detailed model of molecular dynamics

- Example: lattice structure captured by Botzmann model

\[ k_{\text{des},i} = \nu \exp \left( - \frac{E_{\text{des},0} + i\Delta E}{k_b T} \right) \]

(\( \nu \) = vibrational frequency, 10^{13})

- For DNA, RNA, protein interactions, can look at probability of collision + binding energy to reason about rates; also possible to identify via experiments
Simulating the Master Equation for Chemical Reactions

Stochastic Simulation Algorithm (SSA)

- $N$ states ($X_1$, $X_2$, ..., $X_N$) where $X_i$ is the number of copies of species $S_i$ in the system
- $M$ reaction channels $R_i$ that define changes in the state; $\nu_i(X) = \text{change in } X_i \text{ for } R_i$
- Propensity function: $a_i(X) \, dt = \text{probability that a reaction } R_i \text{ will occur in time interval } dt$
- SSA (“Gillespie algorithm”): determine how many reactions occur in a given time step and execute; then update propensity functions and repeat
- Choose time steps to be small enough that propensity functions are roughly constant

Example

$X_\infty + Y_1 \longrightarrow 2Y_1$ with propensity $a_1$

$Y_1 + Y_2 \longrightarrow 2Y_2$ with propensity $a_2$

$Y_2 \longrightarrow Z$ with propensity $a_3$

- Propensity function: $a_1(X) = c_1 \, X_\infty \, Y_1$ to account for multiple copies of each species

Tools

- SBML - modeling language + many tools
- StochKit (Linda Petzold) - C++ libraries
- SimBiology (MATLAB) - includes SSA

http://www.caam.rice.edu/~caam210/reacllec.html
Chemical Kinetics: The Law of Mass Action

Alternative approach: keep track of concentrations

- If number of molecules is large, we can keep track of concentration of each species
- No longer track individual events; assume an average rate of events and use ODEs

\[
\begin{align*}
    X_\infty + Y_1 &\rightarrow 2 Y_1 \quad \text{with rate } k_1 \\
    Y_1 + Y_2 &\rightarrow 2 Y_2 \quad \text{with rate } k_2 \\
    Y_2 &\rightarrow Z \quad \text{with rate } k_3
\end{align*}
\]

- Keep track of change due to each reaction

\[
\begin{align*}
    d[Y_1]/dt &= k_1[X_\infty][Y_1] - k_2[Y_1][Y_2] \\
    d[Y_2]/dt &= k_2[Y_1][Y_2] - k_3[Y_2]
\end{align*}
\]

More complicated reactions:

\[
2 A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} A_2 B
\]

\[
\begin{align*}
    d[A]/dt &= 2k_{-1}[A_2 B] - 2k_1[A]^2 \cdot [B] \\
    d[B]/dt &= k_{-1}[A_2 B] - k_1[A]^2 \cdot [B] \\
    d[A_2 B]/dt &= k_1[A]^2 \cdot [B] - k_{-1}[A_2 B]
\end{align*}
\]

\[
\frac{dx}{dt} = Sv(x)
\]

- \(x\) = concentration vector
- \(S\) = stochiometry matrix
- \(v(x)\) = Rate vector

Stochiometry matrix
Reaction rates
Michaelis-Menten Kinetics

Enzymatically controlled reactions

- Enzyme E acts as a catalyst for reaction
  \[ E + S \xrightleftharpoons[k_-]{k_+} ES \to E + P \]

- Dynamics follow from mass action kinetics
  \[
  \begin{align*}
  \frac{d[E]}{dt} &= k_- [ES] - k_+ [E] \cdot [S] + r[ES] \\
  \frac{d[S]}{dt} &= k_- [ES] - k_+ [E] \cdot [S] \\
  \frac{d[ES]}{dt} &= k_+ [E] \cdot [S] - k_- [ES] - r[ES] \\
  \frac{d[P]}{dt} &= r[ES]
  \end{align*}
  \]

- If we assume enzyme dynamics are fast compared to other species, we can simplify the resulting dynamics:
  \[
  [ES] = \frac{[E] \cdot [S]}{K_m}
  \]
  \[
  K_m = \frac{k_- + r}{k_+}
  \]
  \[
  \frac{d[P]}{dt} = r[E]_{tot} \cdot \frac{[ES]}{[E] + [ES]} = V_{max} \cdot \frac{[S]}{K_m + [S]}
  \]

Phillips, Kondev, Theriot (2008)
Phosphorylation

Post-translational modification of proteins
- Addition of a phosphate (PO$_4$) group to a protein
- Modifies shape of protein; affects binding to other proteins (Xp “active”)

Kinase
- Binds ATP and catalyzes a transfer of phosphate group to protein
- Enzymatic reaction (kinase is available again after reaction)

Phosphatase
- Removes phosphate group
- Non-specific enzyme

**Phosphorylation**

E + ATP $\rightleftharpoons$ E:ATP
E:ATP + X $\rightleftharpoons$ E:ATP:X
E:ATP:X $\rightarrow$ E:ADP:Xp
E:ADP:Xp $\rightleftharpoons$ E:ADP + X
E:ADP $\rightarrow$ E + ADP

**Dephosphorylation**

Xp + F $\rightleftharpoons$ Xp:F
Xp:F $\rightarrow$ X + F

Consider adding hydrolysis more explicitly in dephosphorylation
Transcriptional Regulation

Prokaryotic regulation occurs via binding of transcription factors (proteins) to DNA

- Activator: recruits RNA polymerase
- Repressor: blocks binding of RNAP
- Model dynamics by tracking binding events

Reduced order model: Hill functions

- Assume that binding dynamics are fast compared to transcription

Hill function

\[
\frac{dm_Y}{dt} = f(X) - \gamma m_Y
\]

\[
\frac{dY}{dt} = \beta m_Y - \delta Y
\]

Degradation

Translation

Activation

\[
g_A([TF]) = \frac{f^{-1} + ([TF] / \kappa)^n}{1 + ([TF] / \kappa)^n}
\]

Repression

\[
g_R([TF]) = \frac{1 + ([TF] / \kappa)^n / f}{1 + ([TF] / \kappa)^n}
\]

Buchler, Gerland and Hwa (PNAS, 2005)
Questions:

- In the absence of glucose, what concentration of lactose is required for the lac operon to become “active”?
- Focuses on “bistability”: lac operon has two stable equilibrium points:
  - low lactose: machinery off
  - high lactose: machinery on

Model

- Ordinary differential equation for rates of transcription, translation and degradation of β-galactosidase (β-gal) and allolactose
- Assume levels of lactose outside the cell is constant and level of permease (from lacY) is constant to simplify the model
- Takes into account time delays in producing proteins (RBS transcription + protein translation)
Model Derivation: β-Gal Production

\[
\frac{dM}{dt} = \alpha_M \frac{1 + K_1 (e^{-\mu_T M} A(t - \tau_m))^n}{K + K_1 (e^{-\mu_T M} A(t - \tau_m))^n} - \tilde{\gamma}_M M
\]

\[
\frac{dB}{dt} = \alpha_B e^{-\mu_T B} M(t - \tau_B) - \tilde{\gamma}_B B
\]

- **M** = lacZ mRNA concentration
- **B** = β-gal concentration
- **A** = allolactose concentration

**mRNA production**
- Production rate related to \( p_{\text{bound}} \) via a modified Hill function, \( \text{ala} \)
  \[
  \frac{dM}{dt} = \alpha (1 - p_{\text{bound}}(A)) + \epsilon - \gamma M
  \]
- RNA degradation via exponential decay
- Account for time delay in translation of RBS via \( \tau_M \):
  - Use allolactose concentration, \( A \), at time \( t - \tau_M \)
  - Exponential factor to account for dilution due to cell division

**Protein production**
- Assume rate of production is proportional to amount of mRNA
- Include protein degradation via exponential decay
- Add time delay to account for time to produce functional protein, \( \tau_B \)
Model Derivation: Allolactase Dynamics

\[
\frac{dA}{dt} = \alpha_A B \frac{L}{K_L + L} - \beta_A B \frac{A}{K_A + A} - \tilde{\gamma}_A A
\]
\[
\frac{dL}{dt} = \alpha_L P \frac{L_e}{K_{L_e} + L_e} - \beta_{L_1} P \frac{L}{K_{L_1} + L} - \beta_{L_2} B \frac{L}{K_{L_2} + L} - \tilde{\gamma}_L L
\]
\[
\frac{dP}{dt} = \alpha_P e^{-\mu \tau_P} M (t - \tau_P) - \tilde{\gamma}_P P
\]

A = allolactose concentration
L = internal lactose concentration
P = permease concentration
B = β-gal concentration

Allolactose
1. Converted from lactose with Michaelis-Menten-like kinetics (Huber et al)
2. Converted back to glucose and galactose
3. Degradation

Lactose (internal)
1. Transported to interior of cell by permease
2. Loss back to external environment
3. Converted to allolactose by β-gal
4. Degradation

Permease
1. Produced by lacY gene (after delay)
2. Degradation
Determining the Constants

Yildirim, Santillan, Horike and Mackey:

- $\mu$ - dilution rate, based on 20 minute cell division time
- $\alpha_x$ - production rate, based on steady state values
- $\gamma_x$ - decay rate, based on half life experiments
- $\tau_M$ - time delay to produce RBS, based on RNA elongation rates
- $\tau_B$ - time delay to translate protein, based on protein length and translation speed
- $n$ - Hill coefficient (no justification!)
- $K$ - based on basal rate of production (Yagil & Yagil)
- $K_1$ - based on dissociation constant (Yagil & Yagil)
- $K_x$ - measured by Wong, Gladney and Keasling (97)
- $\beta_A$ - loss of allolactase, through conversion to glucose and galactose. Measured by Hubert et al (75)

Note: repressor binding model is pretty ad hoc...

\[
\frac{dM}{dt} = \alpha_M \frac{1}{K + K_1 (e^{-\mu \tau_M} A(t - \tau_m))^n} - \gamma_M M
\]
\[
\frac{dA}{dt} = \alpha_A B \frac{L}{K_L + L} - \beta_A B \frac{a}{K_A + A} - \gamma_A A
\]
Some Predictions

Bistable behavior (saddle node bifurcations)

- Can have single or multiple equilibrium points depending on parameters
- Bifurcation plot: change in stability versus params
  - Note: possible hysteresis from saddle node
- Parametric stability plot: stability regions
- Simulations: nearby initial conditions can lead to different steady state solutions
- Use to predict behavior (for future experiments)
Comparison to Experiment

**β-gal activity for $L_e = 8 \times 10^{-2}$ mM**
- Experimental data from Knorre (1968) for *E. coli* ML30 (○) and Pestka et al. (1984) for *E. coli* 294 (●)
- Model simulation using constants from Table 1 (slide 16) with $\mu = 2.26 \times 10^{-2}$ min$^{-1}$ and $\gamma_x (= ??)$ fit to data

**Oscillation in β-gal w/ phosphate feeding**
- Periodic phosphate feeding from Goodwin (1969)
- Simulation used $\mu = 2.26 \times 10^{-2}$ min$^{-1}$ and $\gamma_x (= ??)$. Other parameters unchanged.

- Q: how should we assess these data?
Modeling Noise in Chemical Reactions

Fundamental assumption (Gillespie, 1992)
- Treat species concentrations as a random variable $X(t)$
  \[ a_j(x) \, dt = \text{the probability, given } X(t) = x, \text{ that one } R_j \text{ reaction will occur in the next infinitesimal time interval } [t, t + dt] \]

Chemical Master equation (exact)
- Describes the evolution of the probability distribution for the microstate $X(t)$
  \[ \frac{\partial P(x, t | x_0, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(x - v_j)P(x - v_j, t | x_0, t_0) - a_j(x)P(x, t | x_0, t_0)] \]

Chemical Langevin equation (CLE)
- Assume that time steps are small enough but not too small
  - (i) Time increment $dt$ small enough that propensity functions are approx constant
  - (ii) Time increment $dt$ large enough that number of events is much larger than 1
- Under these assumptions, can derive stochastic ODE (Gillespie, 2000):
  \[ X_i(t + dt) = X_i(t) + \sum_{j=1}^{M} v_{ji}a_j(X(t))dt + \sum_{j=1}^{M} v_{ji}a_j^{1/2}(X(t))N_j(t)(dt)^{1/2} \]

Remarks
- Mass action kinetics are deterministic version of mean of the CLE
Cell Noise (Elowitz et al, 2002)

Noise in cells
- Put RFP and GFP under identical promoters; *should* get yellow
- Results: get range of colors

Extrinsic Noise:
- global to a single cell, but varies from one cell to the next (e.g. cell volume, plasmid copy number)

Intrinsic Noise:
- inherent stochasticity in gene expression (e.g. what order reactions occur in)

\[
\dot{x}_i = E(t) \cdot f_i(x_i, I_i(t))
\]

Coefficient of variation: \[
CV(X) = \frac{\text{std}(X)}{\text{mean}(X)}
\]
- Normalized measure of noise, commonly used in biology
Modeling Summary

Core processes in biological circuits
- Transcription, translation, binding
- Activation and repression
- Enzymatic reactions, (phosphorylation)

Quantitative modeling approaches
- Statistical thermodynamics
- Chemical master equation, SSA
- Chemical Langevin equation
- Mass action, Michaelis-Menten, Hill functions and ODEs

Lac operon
- Model system for transcriptional regulation
- Activation via CAP:cAMP
- Repression via Lac Repressor (DNA folding)

Next
- Control analysis techniques
- More complex biological circuits