Stochastic Gene Expression: Modeling, Analysis, and Identification

Mustafa Khammash

University of California, Santa Barbara

Outline

- Randomness at the molecular level
- Exploiting the randomness
- Stochastic modeling framework
- Simulation and Analysis Methods
- Identification from Noise
- Conclusions

Randomness at the Molecular Level

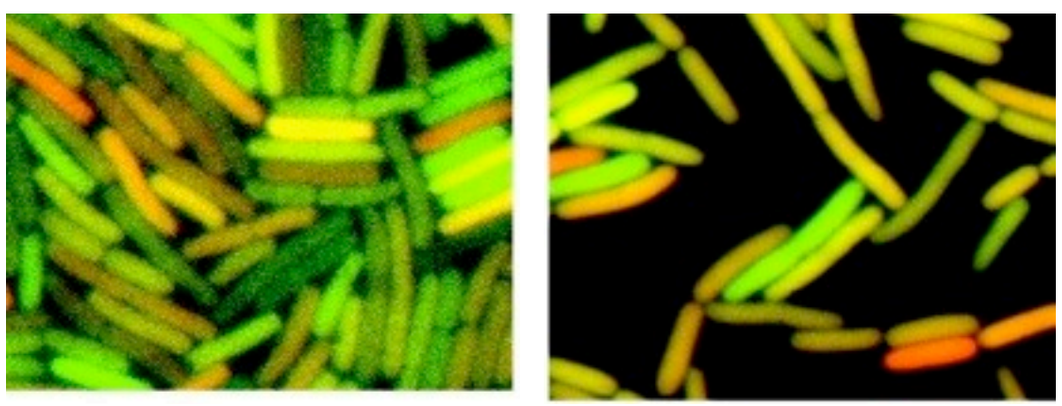
Stochastic Influences on Phenotype

Fingerprints of identical twins

J. Raser and E. O'Shea, Science, 1995.

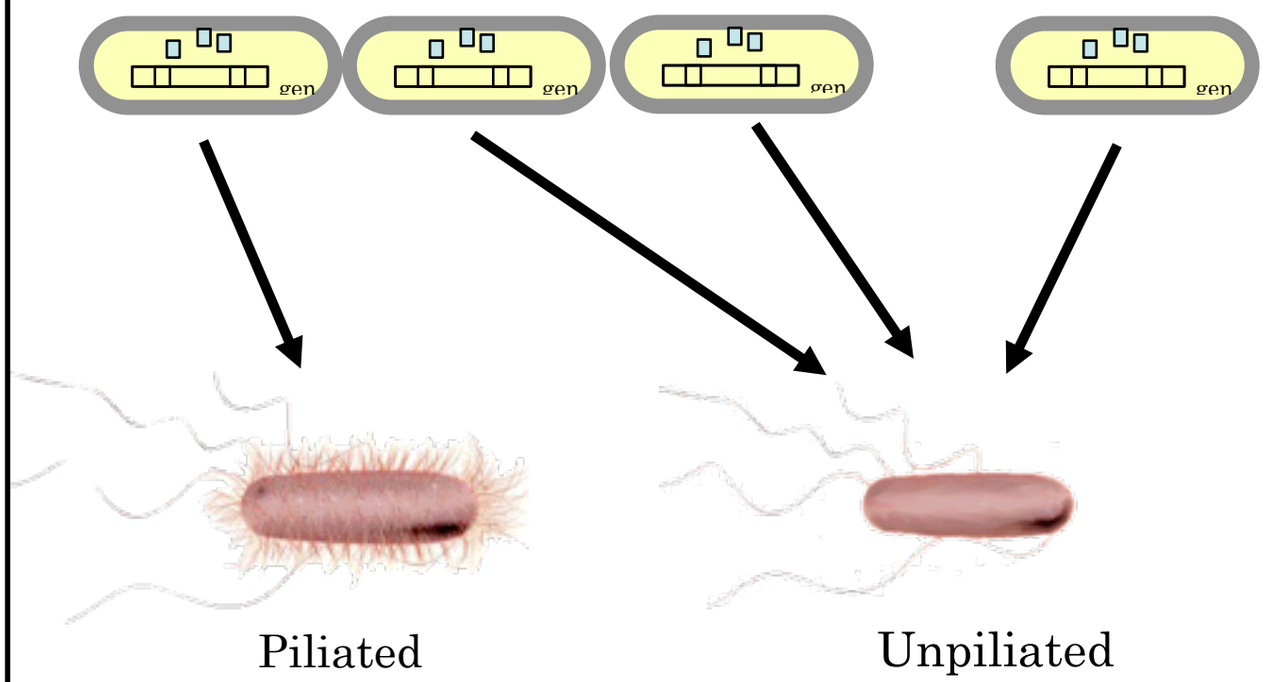
Cc, the first cloned cat and her genetic mother

J. Raser and E. O'Shea, Science, 1995.

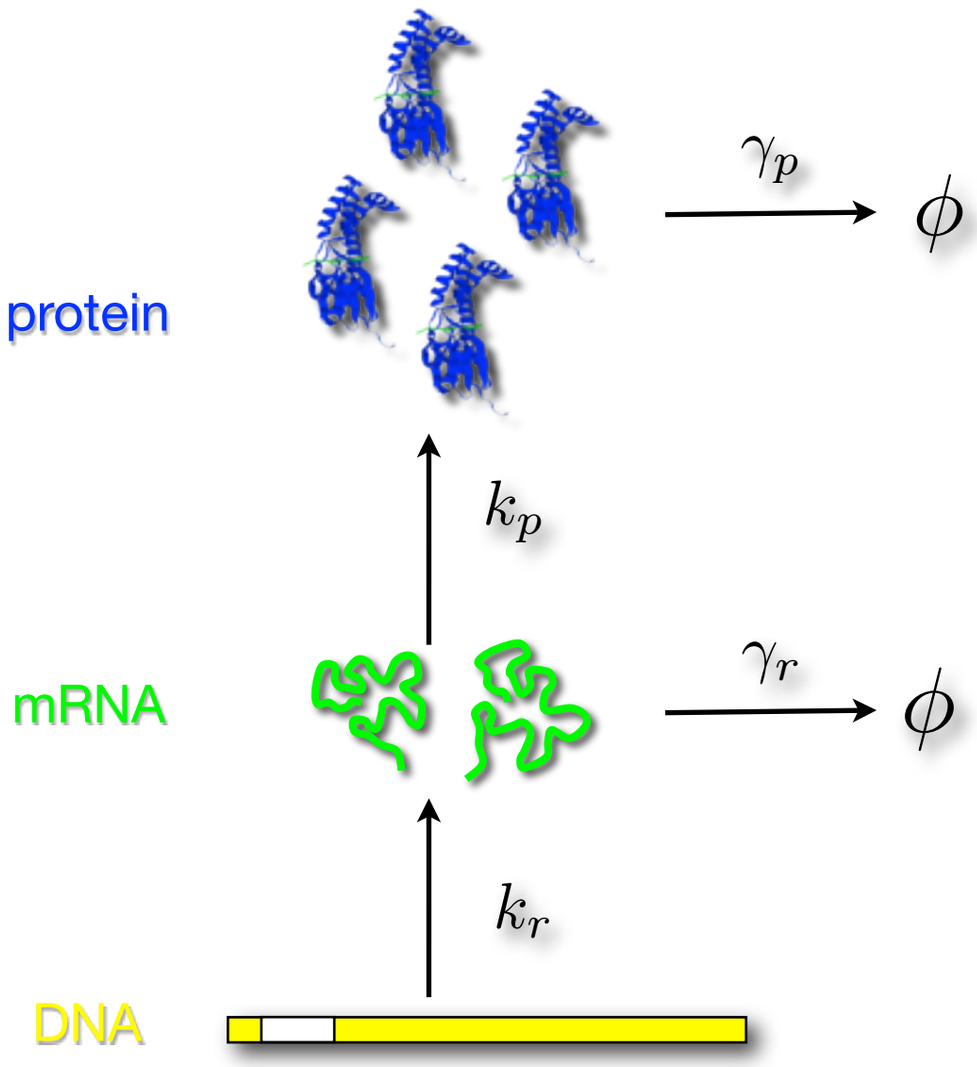


variability in gene expression

Elowitz et al, Science 2002



Capturing Randomness in Gene Expression Models

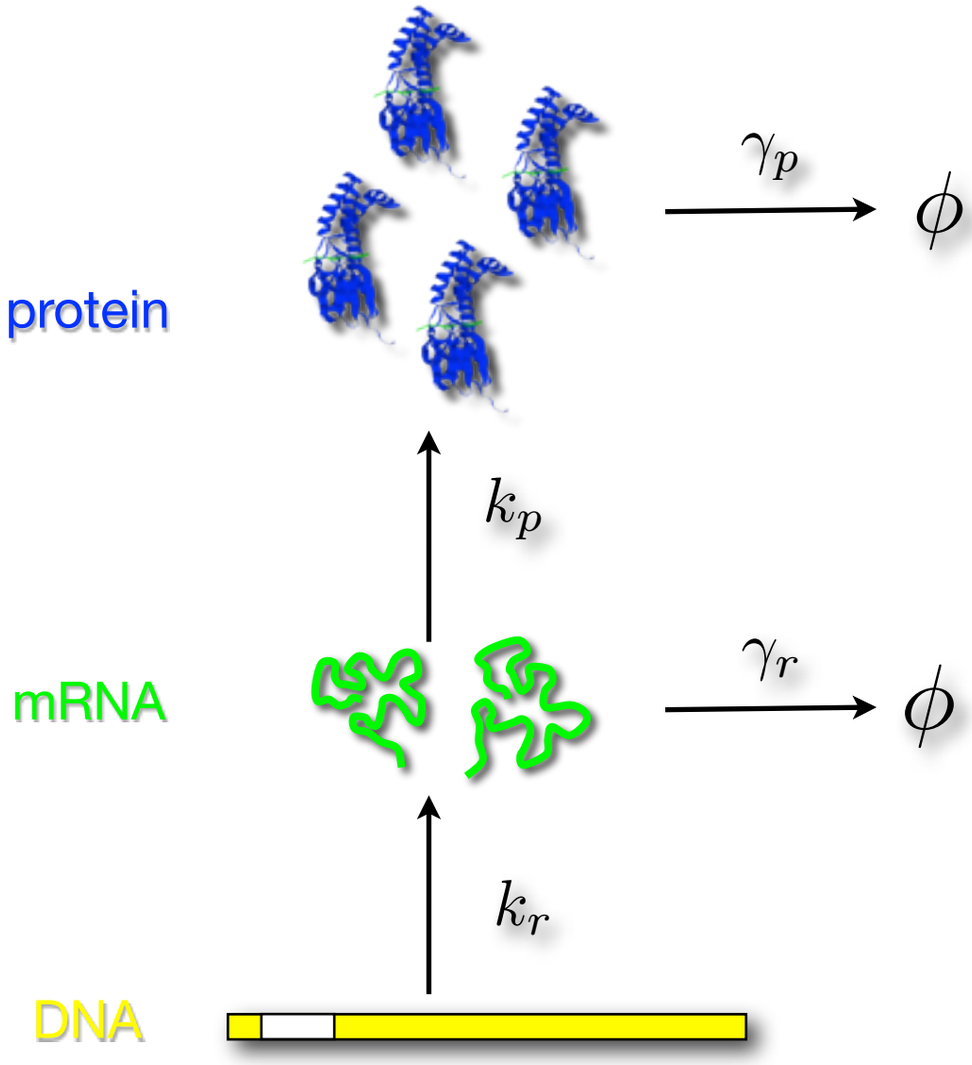


Deterministic model

$$\frac{d[mRNA]}{dt} = -\gamma_r[mRNA] + k_r$$

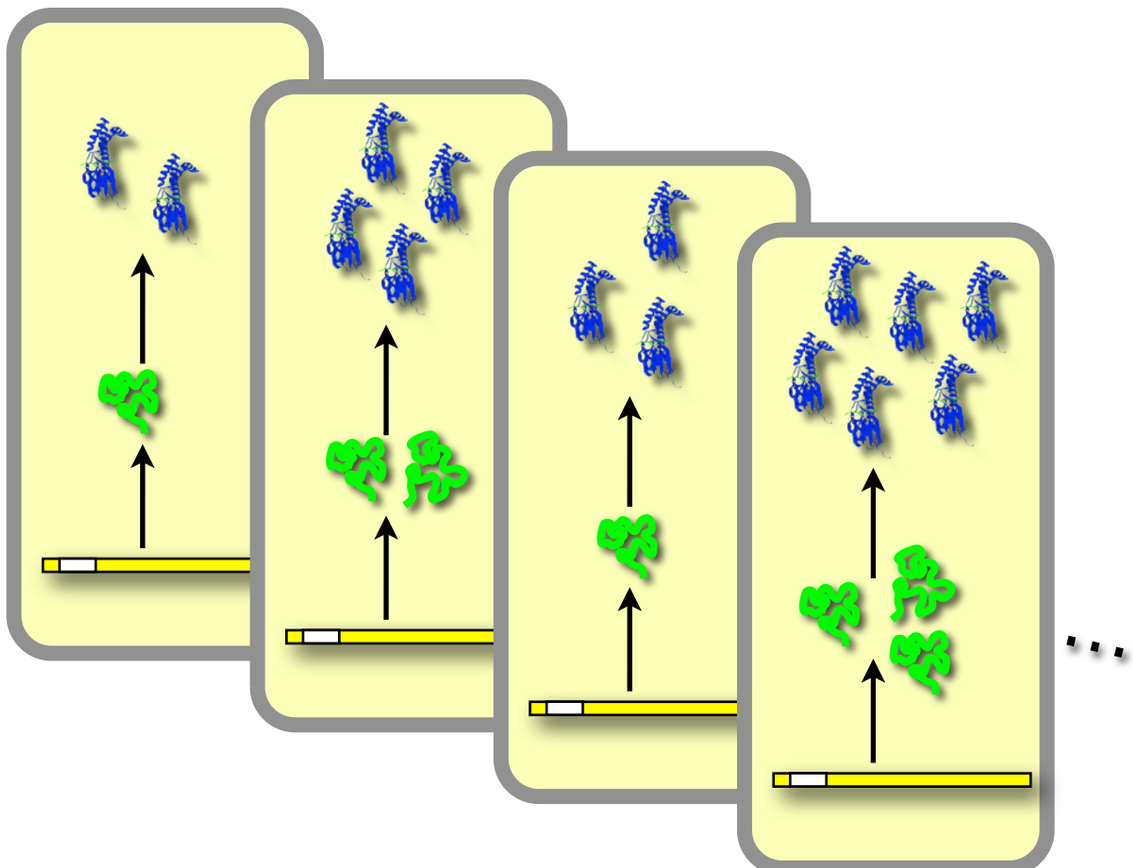
$$\frac{d[protein]}{dt} = -\gamma_p[protein] + k_p[mRNA]$$

Capturing Randomness in Gene Expression Models

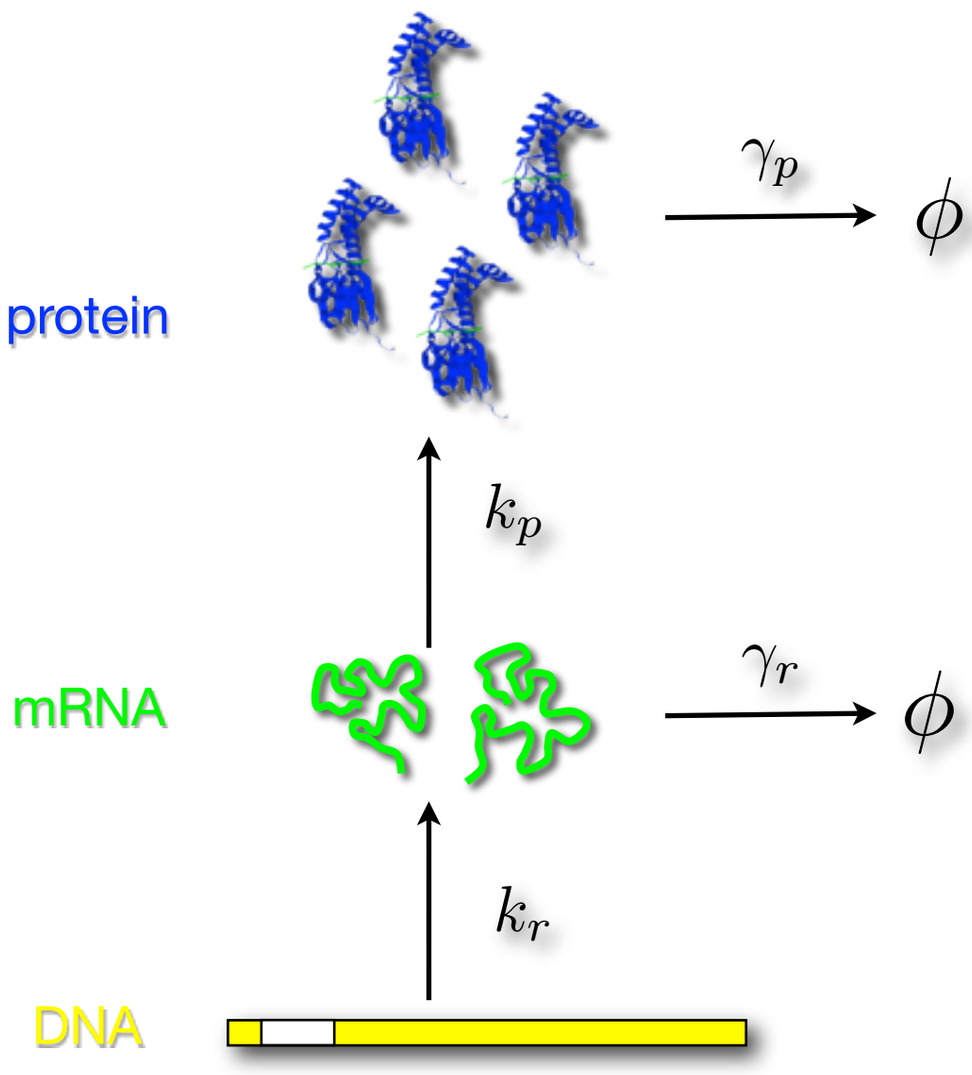


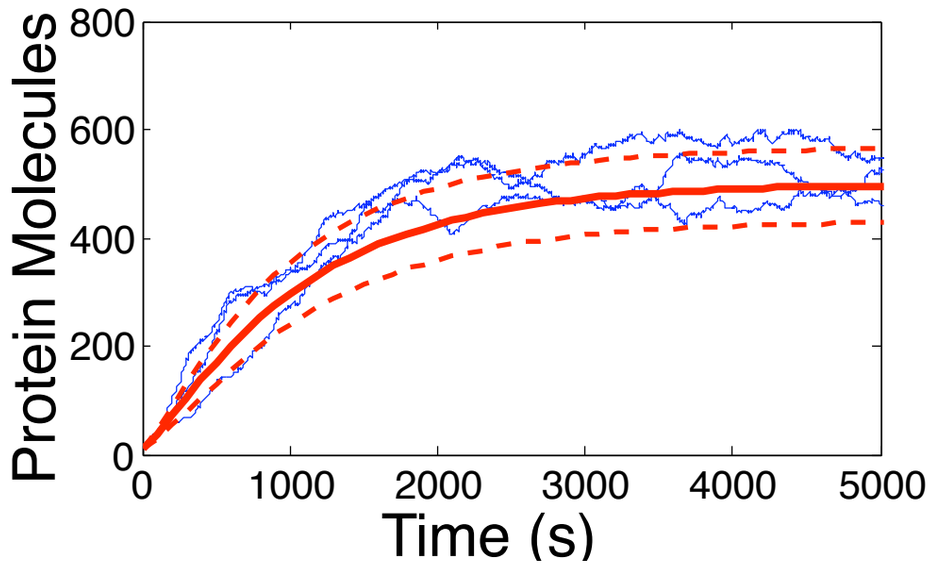
Stochastic model

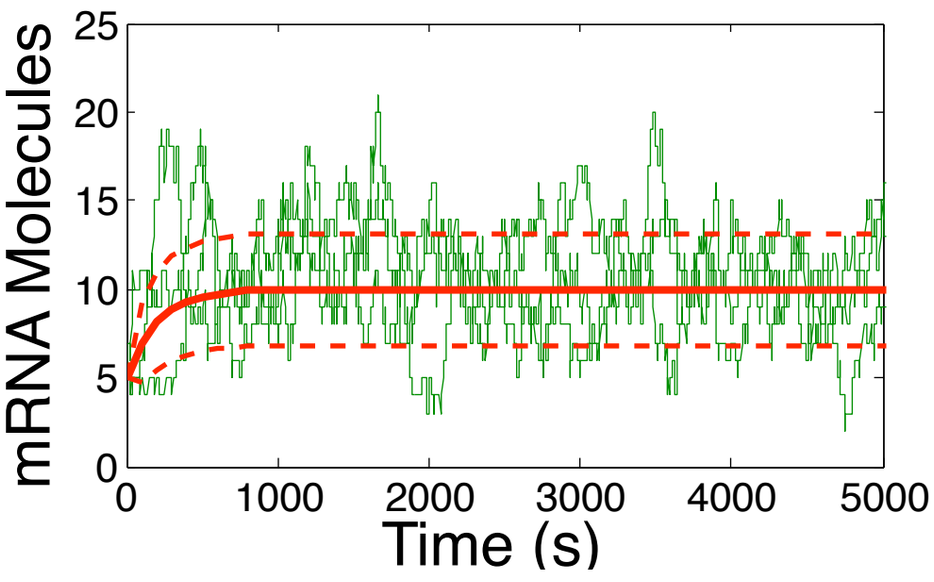
- Probability a single mRNA is transcribed in time dt is $k_r dt$.
- Probability a single mRNA is degraded in time dt is $(\#mRNA) \cdot \gamma_r dt$



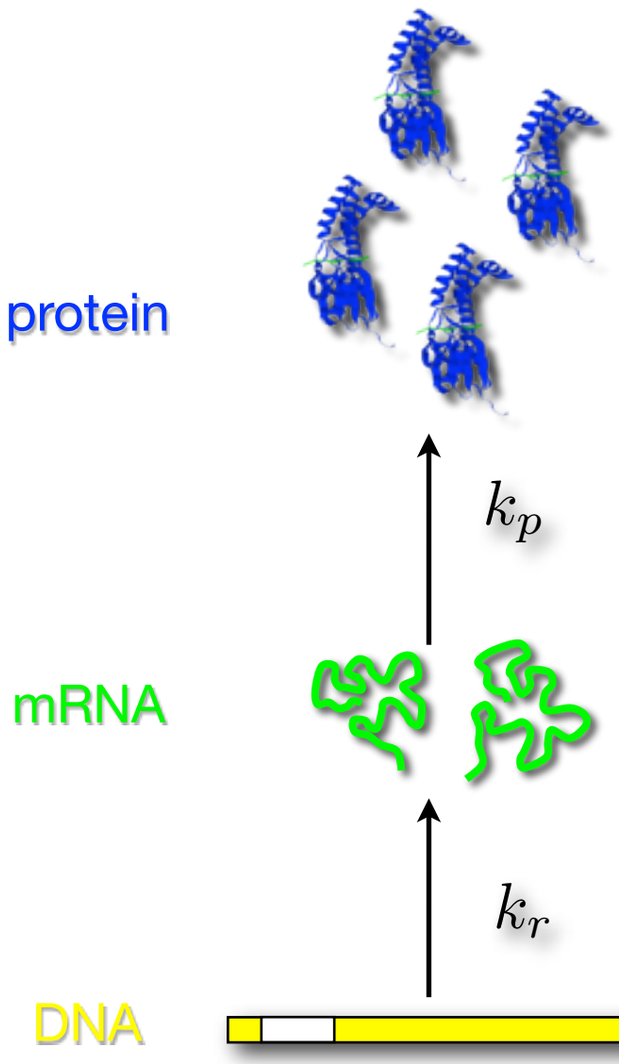
Fluctuations at Small Copy Numbers







Fluctuations at Small Copy Numbers



$$E(p) = \frac{k_r k_p}{\gamma_r \gamma_p}$$

$$\eta(p) = \frac{1}{\sqrt{E(p)}}\left(1 + \frac{kp}{\gamma_p + \gamma_r}\right)^{1/2} \qquad \text{(protein)}$$

$$E(r) = \frac{k_r}{\gamma_r}$$

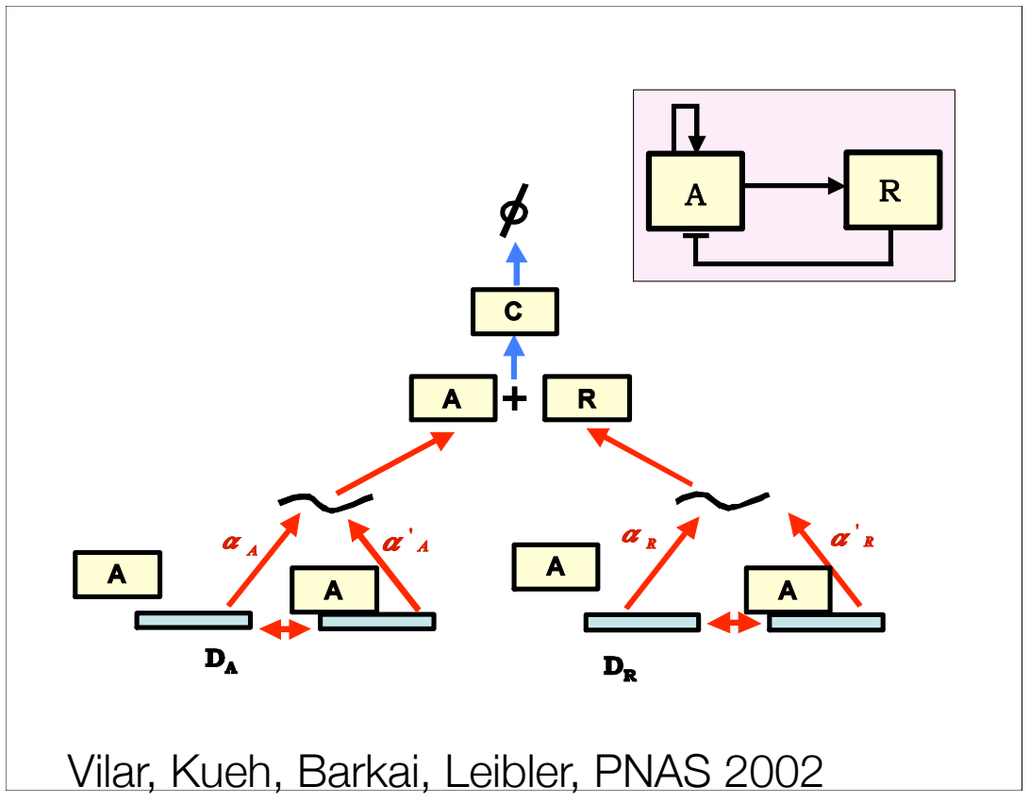
$$\eta(r) = \frac{1}{\sqrt{E(r)}} \qquad \text{(mRNA)}$$

$$C_v = \text{coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}}$$

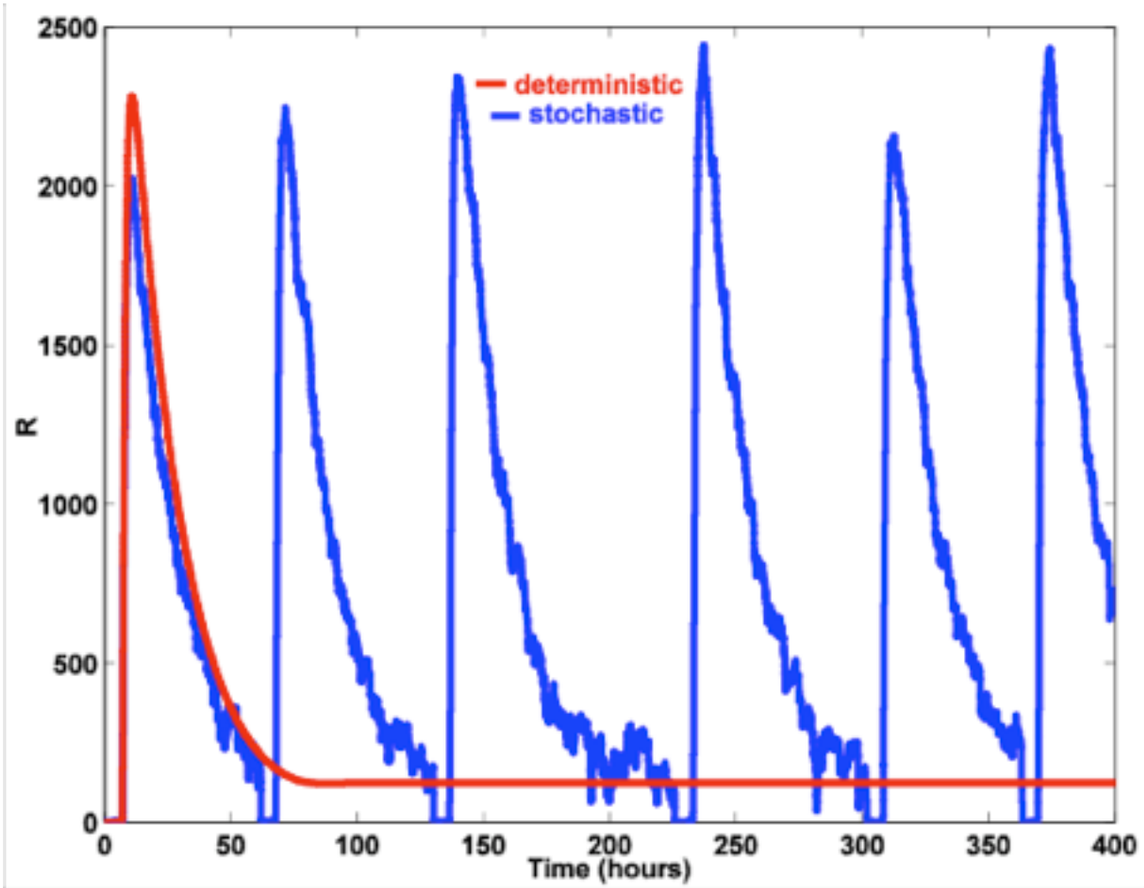
Exploiting the Randomness

Noise Induced Oscillations

Circadian rhythm



Vilar, Kueh, Barkai, Leibler, PNAS 2002



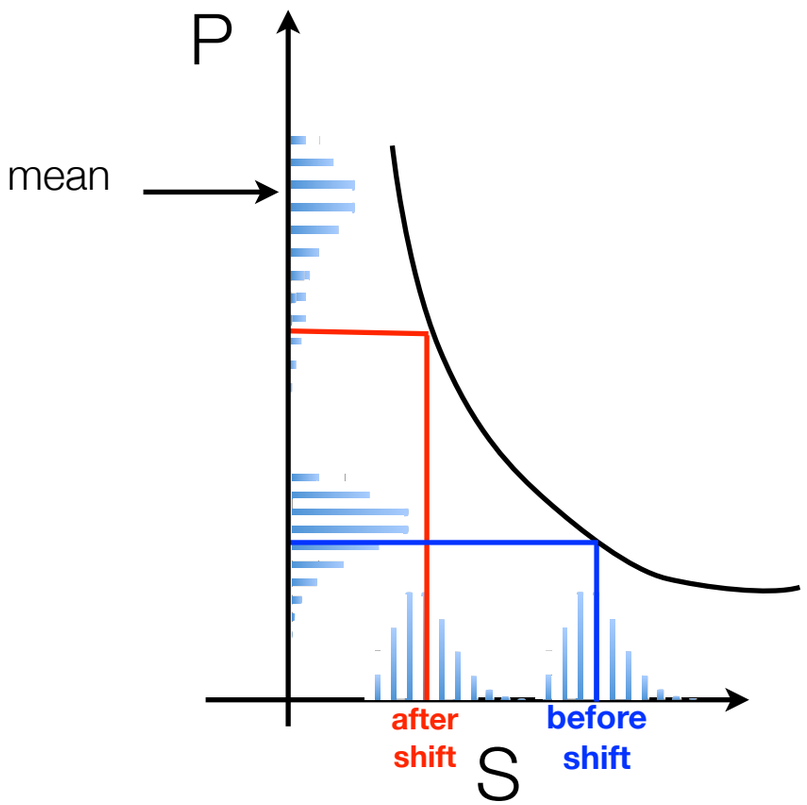
- Oscillations disappear from deterministic model after a small reduction in deg. of repressor
- (Coherence resonance) Regularity of noise induced oscillations can be manipulated by tuning the level of noise [*El-Samad, Khammash*]

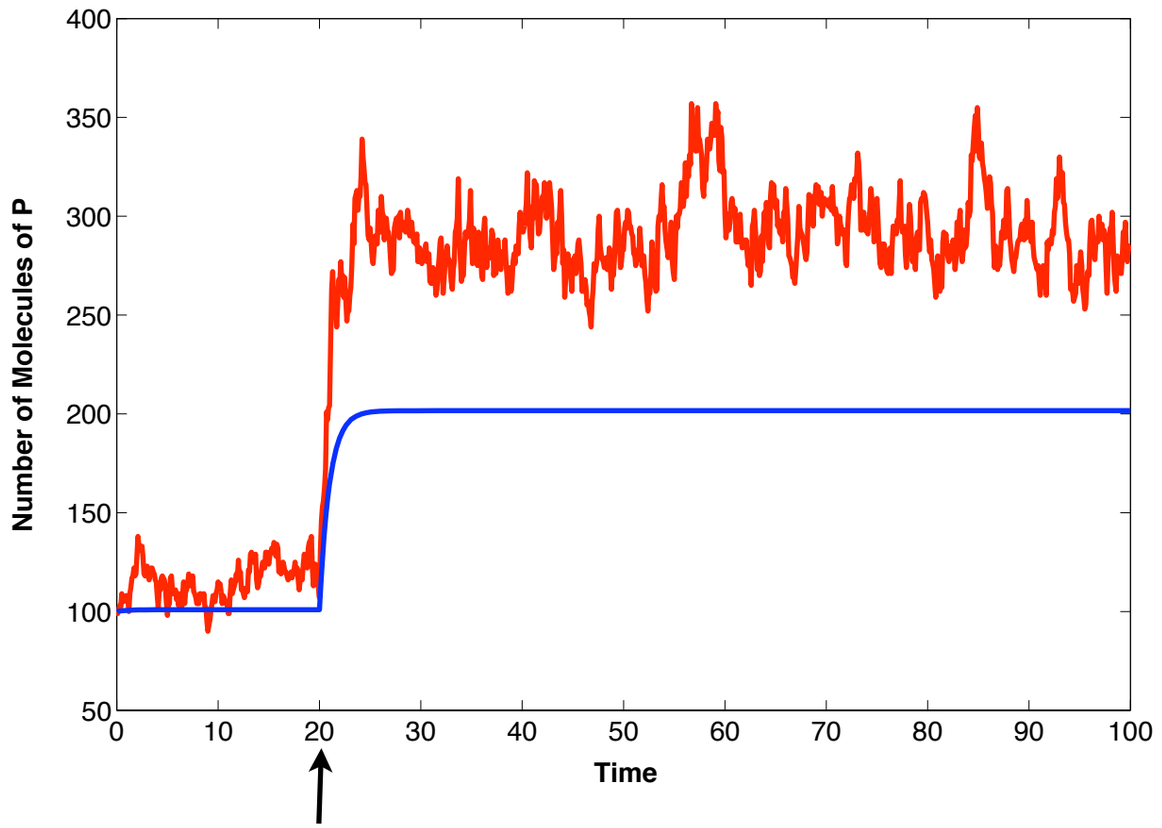
Stochastic Focusing: Fluctuation Enhanced Sensitivity

Signaling Circuit

$$\phi \underset{k_a S}{\overset{k}{\rightleftharpoons}} I \xrightarrow{k_p} P \xrightarrow{1} \phi$$

$$\phi \underset{k_d}{\overset{k_s}{\rightleftharpoons}} S$$





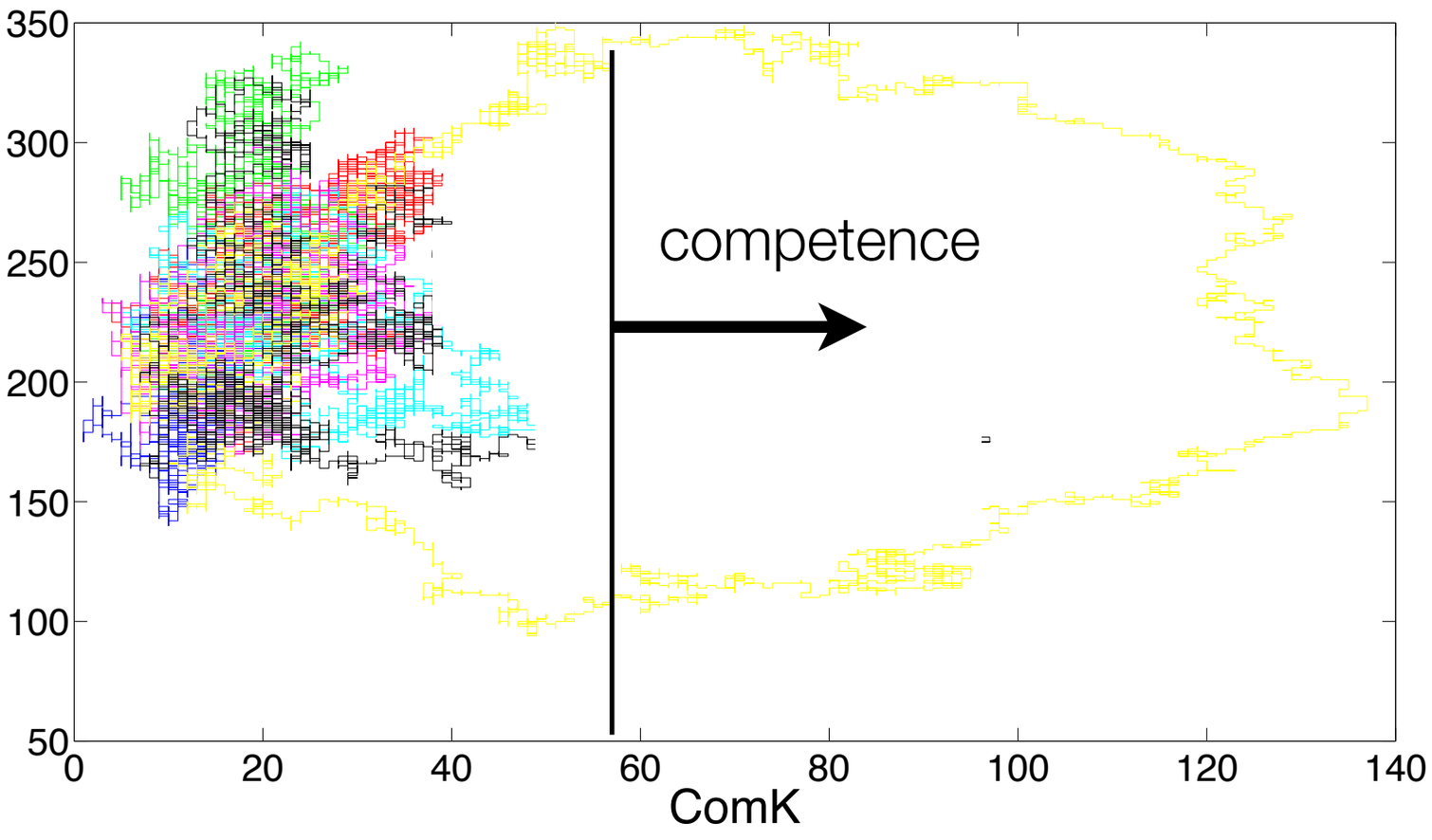
k_s reduced by 50%

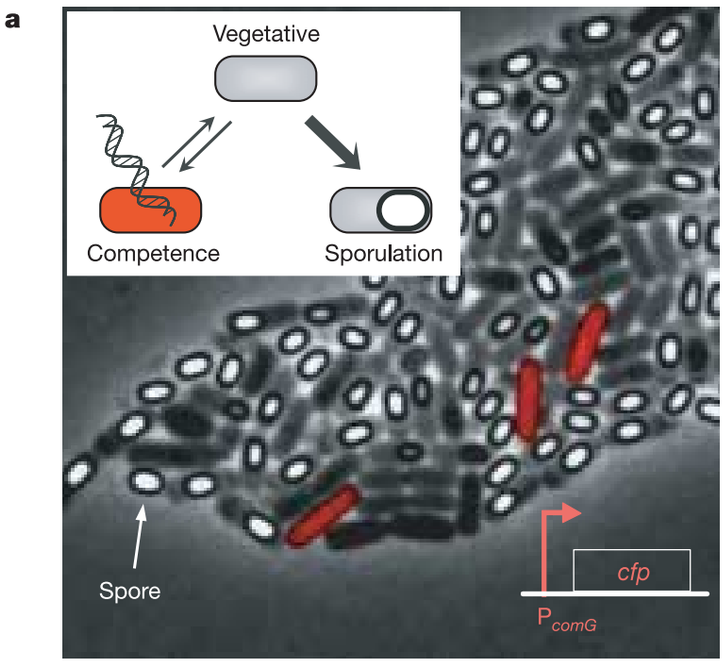
- Stochastic mean value different from deterministic steady state
- Noise *enhances* signal!

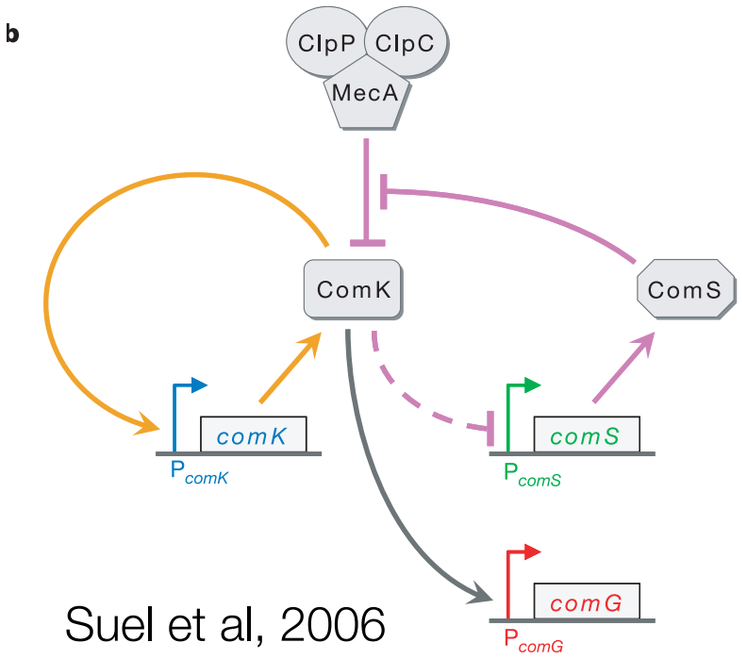
Johan Paulsson , Otto G. Berg , and Måns Ehrenberg, PNAS 2000

Bacterial Competence

- Competence is a process by which bacteria takes up foreign DNA
- Only a fraction of cells become competent



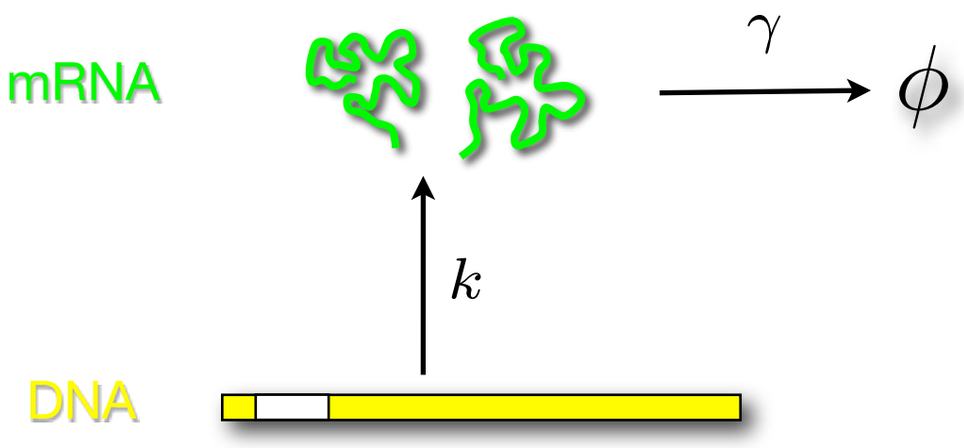




Suel et al, 2006

Stochastic Modeling Framework

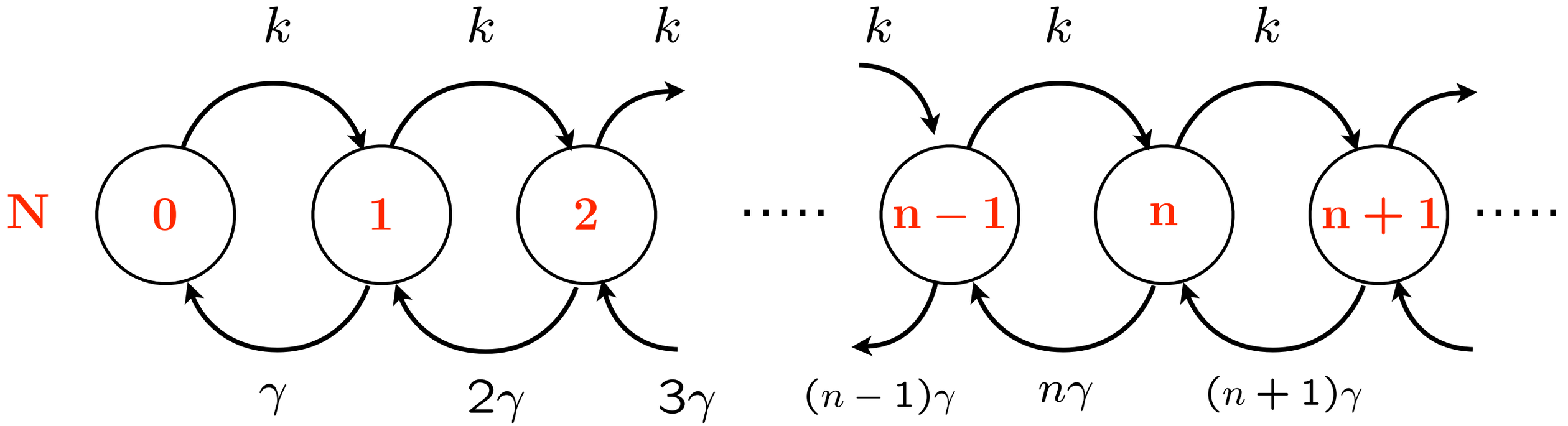
A Simple Example



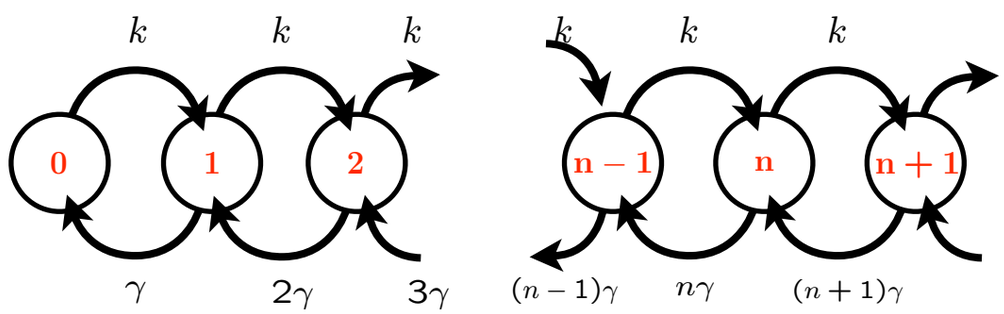
mRNA copy number $N(t)$ is a random variable

Transcription: Probability a single mRNA is transcribed in time dt is $k\ dt$

Degradation: Probability a single mRNA is degraded in time dt is $n\gamma dt$



Key Question:



Find $p(n,t)$, the probability that $N(t) = n$.

$$P(n, t+dt) = P(n-1, t) \cdot k dt$$ Prob.$\{N(t) = n-1$ and mRNA created in [t,t+dt)$\}$

$$+ \; P(n+1, t) \cdot (n+1)\gamma dt$$ Prob.$\{N(t) = n+1$ and mRNA degraded in [t,t+dt)$\}$

$$+ \; P(n, t) \cdot (1 - kdt)(1 - n\gamma dt)$$ Prob.$\{N(t) = n$ and mRNA not created nor degraded in [t,t+dt)$\}$

$$P(n, t+dt) - P(n,t) = P(n-1,t)kdt + P(n+1,t)(n+1)\gamma dt - P(n,t)(k + n\gamma)dt + O(dt^2)$$

Dividing by dt and taking the limit as $dt \to 0$

The Chemical Master Equation

$$\frac{d}{dt}P(n,t) = kP(n-1,t) + (n+1)\gamma P(n+1,t) - (k + n\gamma)P(n,t)$$

mRNA Stationary Distribution

We look for the stationary distribution $P(n,t) = p(n)\ \forall t$

The stationary solution satisfies: $\frac{d}{dt}P(n,t) = 0$

From the Master Equation ...

$$(k + n\gamma)p(n) = kp(n-1) + (n+1)\gamma p(n+1)$$

$n = 0$ $\quad kp(0) = \gamma p(1)$

$n = 1$ $\quad kp(1) = 2\gamma p(2)$

$n = 2$ $\quad kp(2) = 3\gamma p(3)$

$\vdots$

$$kp(n-1) = n\gamma\ p(n)$$

$kp(n-1) = n\gamma\ p(n)$ We can express $p(n)$ as a function of $p(0)$:

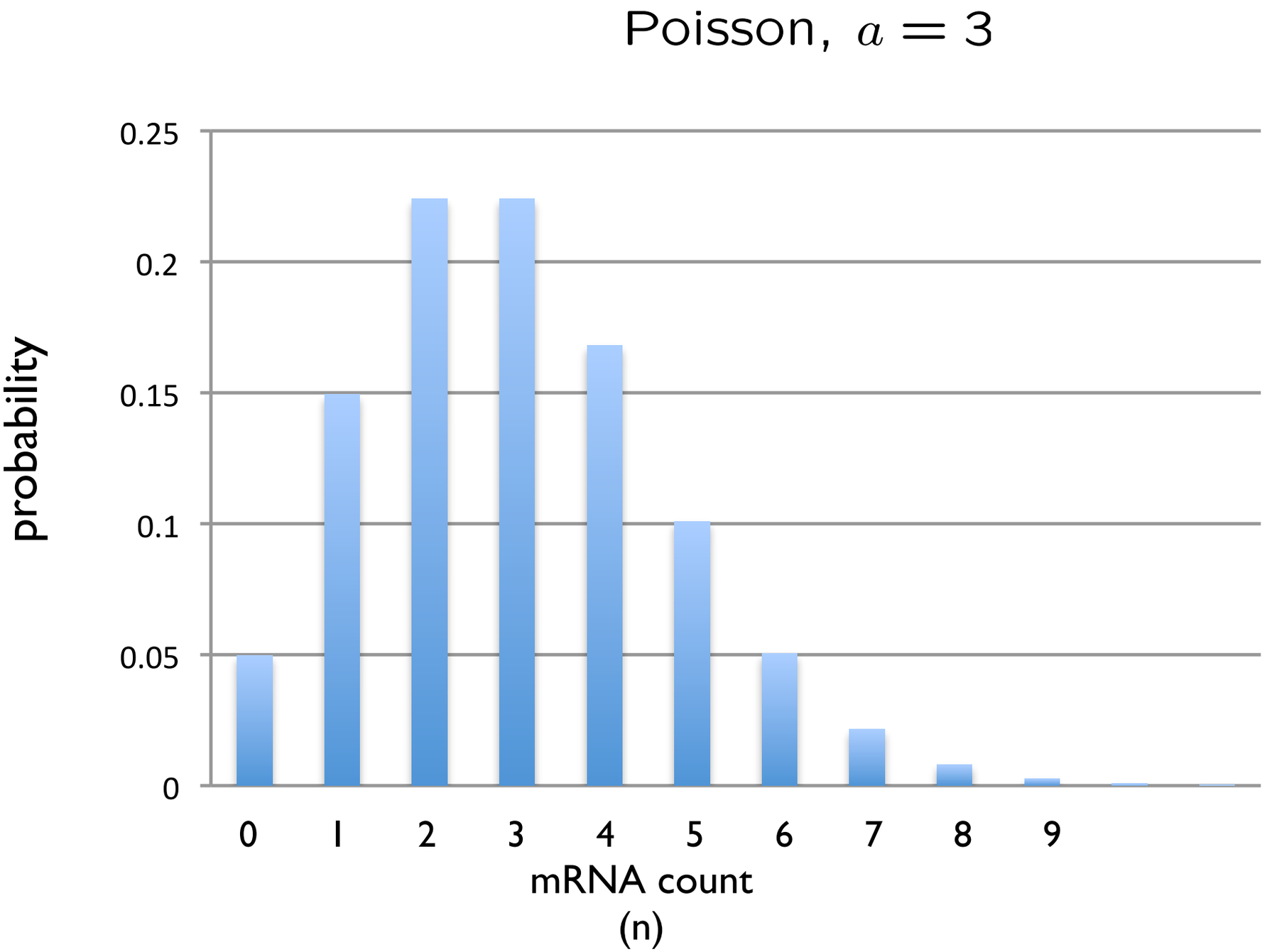
$$\begin{aligned} p(n) &= \frac{k}{\gamma}\frac{1}{n}\,p(n-1) \\ &= \left(\frac{k}{\gamma}\right)^2 \frac{1}{n}\,\frac{1}{n-1}\,p(n-2) \\ &\vdots \\ &= \left(\frac{k}{\gamma}\right)^n \frac{1}{n!}\,p(0) \end{aligned}$$

We can solve for $p(0)$ using the fact $\sum_{n=0}^{\infty} p(n) = 1$

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty}\left(\frac{k}{\gamma}\right)^n \frac{1}{n!}\,p(0) \\ &= e^{k/\gamma}\,p(0) \qquad \Longrightarrow \qquad p(0) = e^{-k/\gamma} \end{aligned}$$

$$p(n) = e^{-a}\frac{a^n}{n!} \qquad a = \frac{k}{\gamma}$$

Poisson Distribution



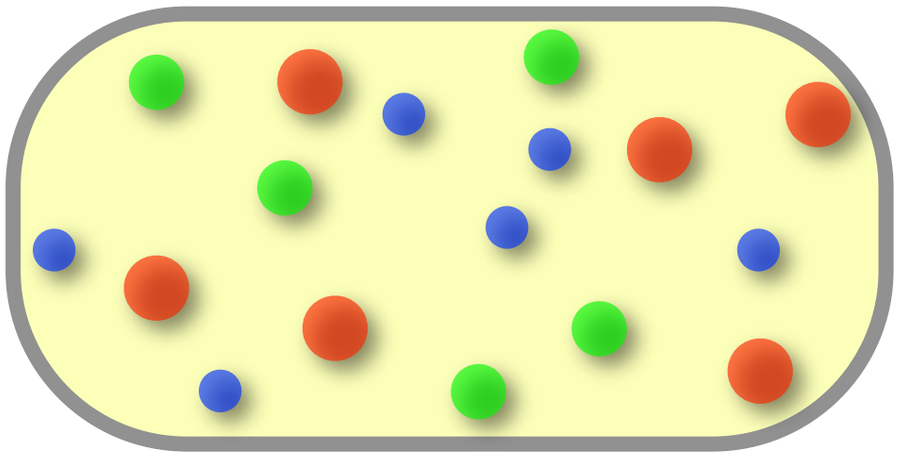
Stationary distribution:

$$P(n) = e^{-a}\frac{a^n}{n!} \qquad a = \frac{k}{\gamma}$$

Poisson Distribution

Formulation of Stochastic Chemical Kinetics

Reaction volume=Ω



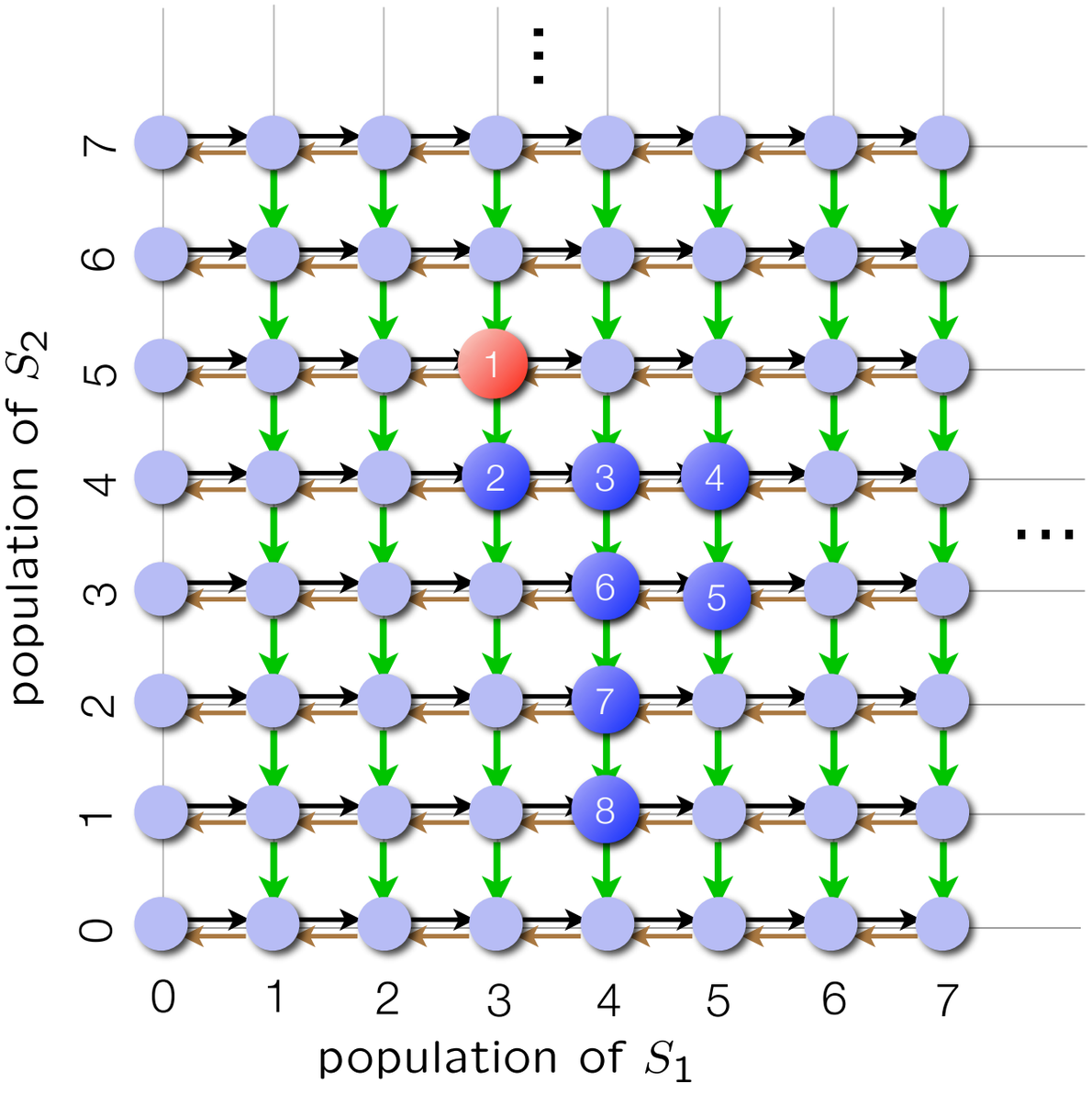
Key Assumptions

(**Well-Mixed**) The probability of finding any molecule in a region $d\Omega$ is given by $\frac{d\Omega}{\Omega}$.

(**Thermal Equilibrium**) The molecules move due to the thermal energy. The reaction volume is at a constant temperature T. The velocity of a molecule is determined according to a Boltzman distribution:

$$f_{v_x}(v) = f_{v_y}(v) = f_{v_z}(v) = \sqrt{\frac{m}{2\pi k_B T}}\, e^{-\frac{m}{2k_B T}v^2}$$

Stochastic Chemical Kinetics



- **(N-species)** $\mathcal{S}_1, \ldots, \mathcal{S}_N$. Population of each is an integer r.v.:

$$X(t) = [X_1(t), \ldots, X_N(t)]^T$$

- **(M-reactions)** The system's state can change through any one of M reaction: $R_k : k \in \{1, 2, \ldots, M\}$.

- **(State transition)** Firing of reaction R_k causes a state transition from $X(t) = x$ to $X(t^+) = x + s_k$.

Stoich. matrix: $S = \begin{bmatrix} s_1 & \cdots & s_M \end{bmatrix}$

- **(Transition Probability)** The probability that reaction R_k fires in the next dt time units is: $w_k(x)dt$.

Example: $w_1(x) = c_1$; $w_2(x) = c_2 \cdot x_1 x_2$; $w_3(x) = c_3 x_1$;

The Chemical Master Equation

$X(t)$ is Continuous-time discrete-state Markov Chain

$p(x,t) := prob(X(t) = x)$

The Chemical Master Equation (Forward Kolmogorov Equation)

$$\frac{dp(x,t)}{dt} = -p(x,t)\sum_k w_k(x) + \sum_k p(x - s_k, t) w_k(x - s_k)$$

From Stochastic to Deterministic

Define $X^{\Omega}(t) = \frac{X(t)}{\Omega}$.

Question: How does $X^{\Omega}(t)$ relate to $\Phi(t)$?

Fact: Let $\Phi(t)$ be the deterministic solution to the reaction rate equations

$$\frac{d\Phi}{dt} = Sf(\Phi), \;\; \Phi(0) = \Phi_0.$$

Let $X^{\Omega}(t)$ be the stochastic representation of the same chemical systems with $X^{\Omega}(0) = \Phi_0$. Then for every $t \geq 0$:

$$\lim_{\Omega \to \infty} \sup_{s \leq t} \left| X^{\Omega}(s) - \Phi(s) \right| = 0 \;\; a.s.$$

Simulation and Analysis Tools

1. Sample Paths Computation

Gillespie's Stochastic Simulation Algorithm:

To each of the reactions $\{R_1, \ldots, R_M\}$ we associate a RV τ_i:

τ_i is the time to the next firing of reaction R_i

Fact 0: τ_i is exponentially distributed with parameter w_i

We define two new RVs:

$\tau = \min_i\{\tau_i\}$ (Time to the next reaction)

$\mu = \arg\min_i\{\tau_i\}$ (Index of the next reaction)

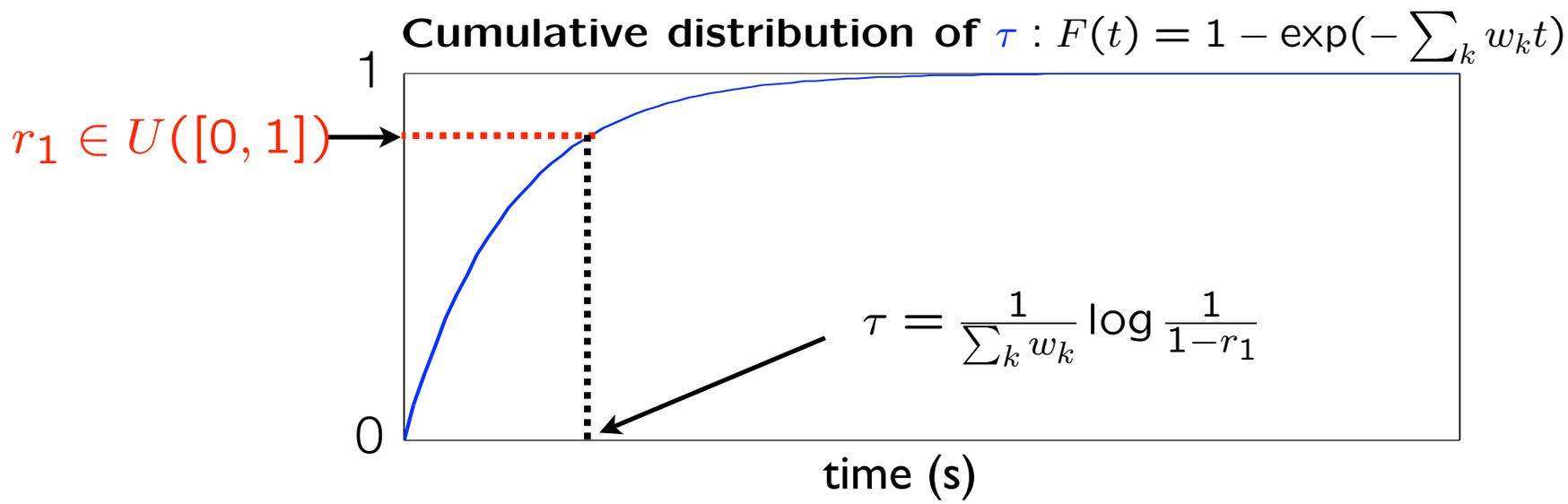
Fact 1: τ is exponentially distributed with parameter $\sum_i w_i$

Fact 2: $P(\mu = k) = \dfrac{w_k}{\sum_i w_i}$

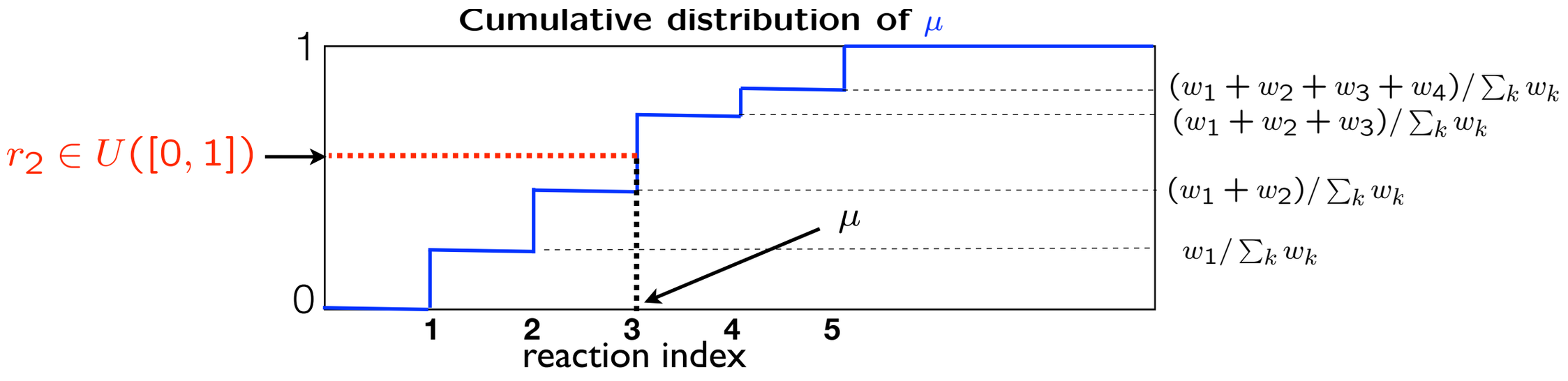
Stochastic Simulation Algorithm

- **Step 0** Initialize time t and state population x

- **Step 1** Draw a sample τ from the distribution of τ



- **Step 2** Draw a sample μ from the distribution of μ



- **Step 3** Update time: $t \leftarrow t + \tau$. Update state: $x \leftarrow x + s_\mu$.

2. Moment Computations

Let $w(x) = [w_1(x), \ldots, w_M(x)]^T$ be the vector of propensity functions

Moment Dynamics

$$\frac{dE[X]}{dt} = S\ E[w(X)]$$

$$\frac{dE[XX^T]}{dt} = SE[w(X)X^T] + E[Xw^T(X)]S^T + S\ diag(E[w(X)])\ S^T$$

- Affine propensity. Closed moment equations.
- Quadratic propensity. Not generally closed.
  - *Mass Fluctuation Kinetics* (Gomez-Uribe, Verghese)
  - *Derivative Matching* (Singh, Hespanha)

3. SDE Approximation

Let $X^\Omega(t) := \frac{X(t)}{\Omega}$

Write $X^\Omega = \Phi_0(t) + \frac{1}{\sqrt{\Omega}} V^\Omega$ where $\Phi_0(t)$ solves the deterministic RRE

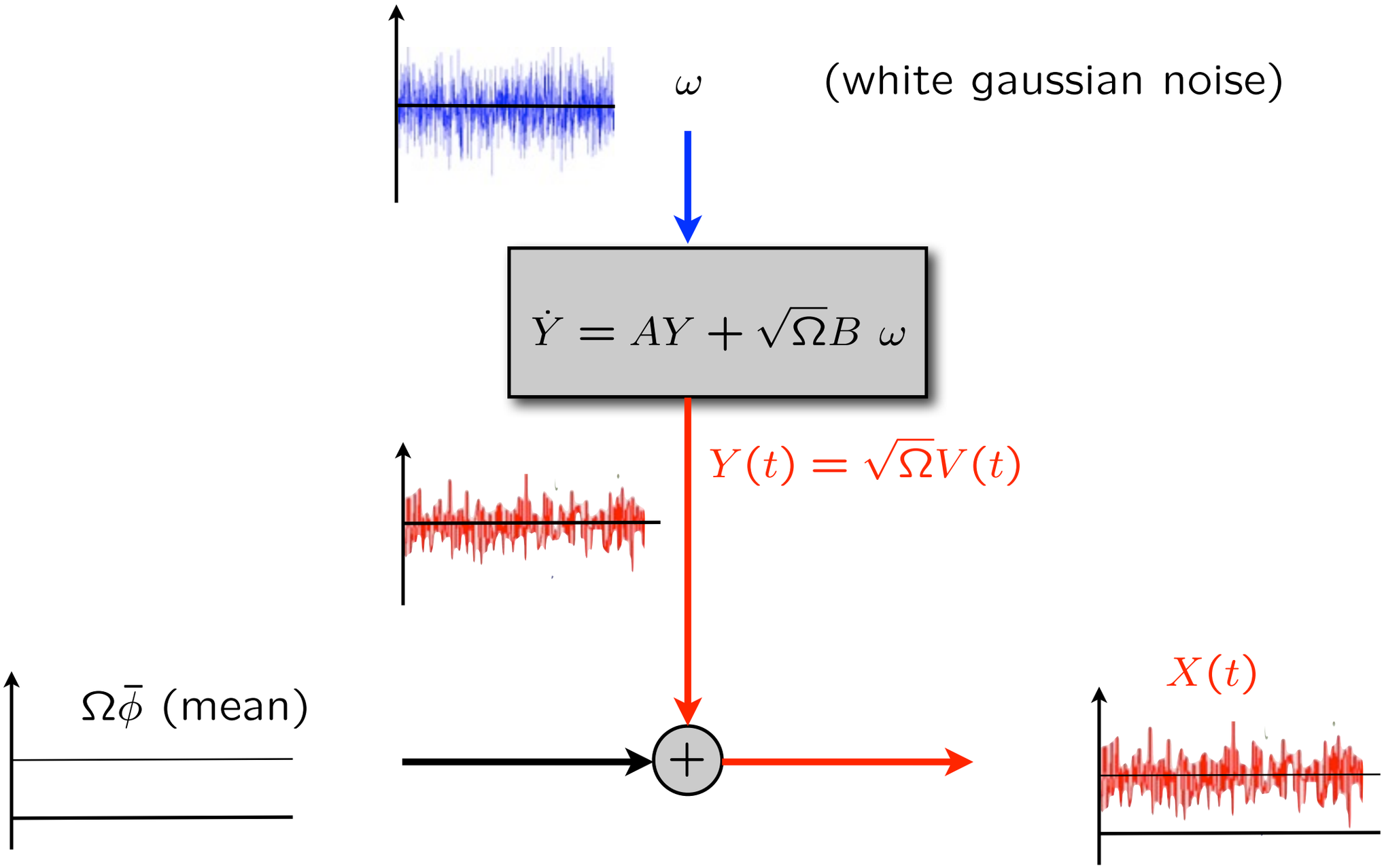
$$\frac{d\Phi}{dt} = Sf(\Phi)$$

Linear Noise Approximation

$V^\Omega(t) \to V(t)$ as $\Omega \to \infty$, where $dV(t) = A(t)V(t)dt + B(t)dW_t$

$$A(t) = \frac{d[Sf(\Phi)]}{d\Phi}(\Phi_0(t)), \qquad B(t) := S\sqrt{diag[f(\Phi_0(t))]}$$

Linear Noise Approximation: $X^\Omega(t) \approx \Phi(t) + \frac{1}{\sqrt{\Omega}} V(t)$


ω
(white gaussian noise)
$\dot{Y} = AY + \sqrt{\Omega}B\ \omega$
$Y(t) = \sqrt{\Omega}V(t)$
$\Omega\bar{\phi}$ (mean)
$X(t)$

Density Computation

Goal: Compute $p(x,t)$, the probability that $X(t) = x$.

Enumerate the state space: $\mathcal{X} = \{x_1, x_2, x_3, \ldots\}$

Form the probability density state vector $P(\mathcal{X}, \cdot) : R^+ \rightarrow \ell_1$

$$P(\mathcal{X},t) := [p(x_1,t) \quad p(x_2,t) \quad p(x_3,t) \quad \ldots \quad]^T$$

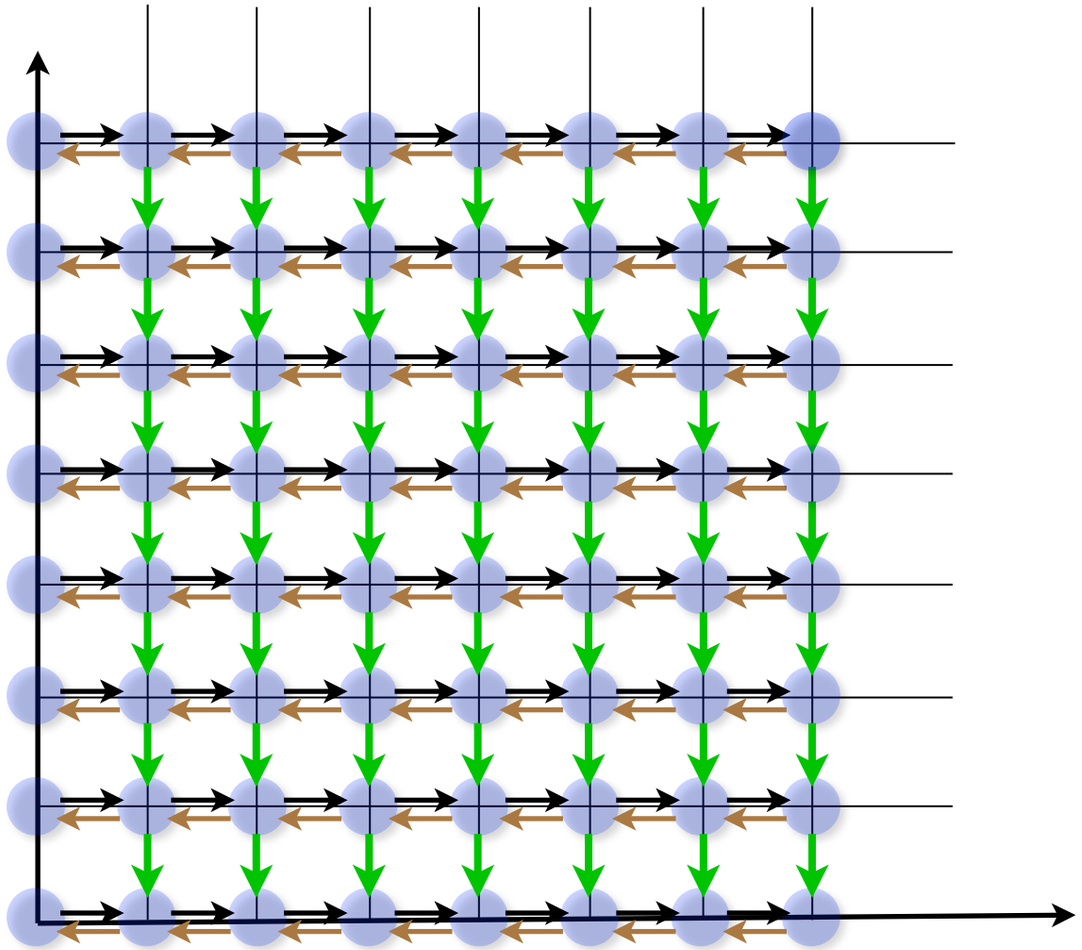
The Chemical Master Equation (CME):

$$\frac{dp(x,t)}{dt} = -p(x,t)\sum_k w_k(x) + \sum_k p(x-s_k,t)w_k(x-s_k)$$

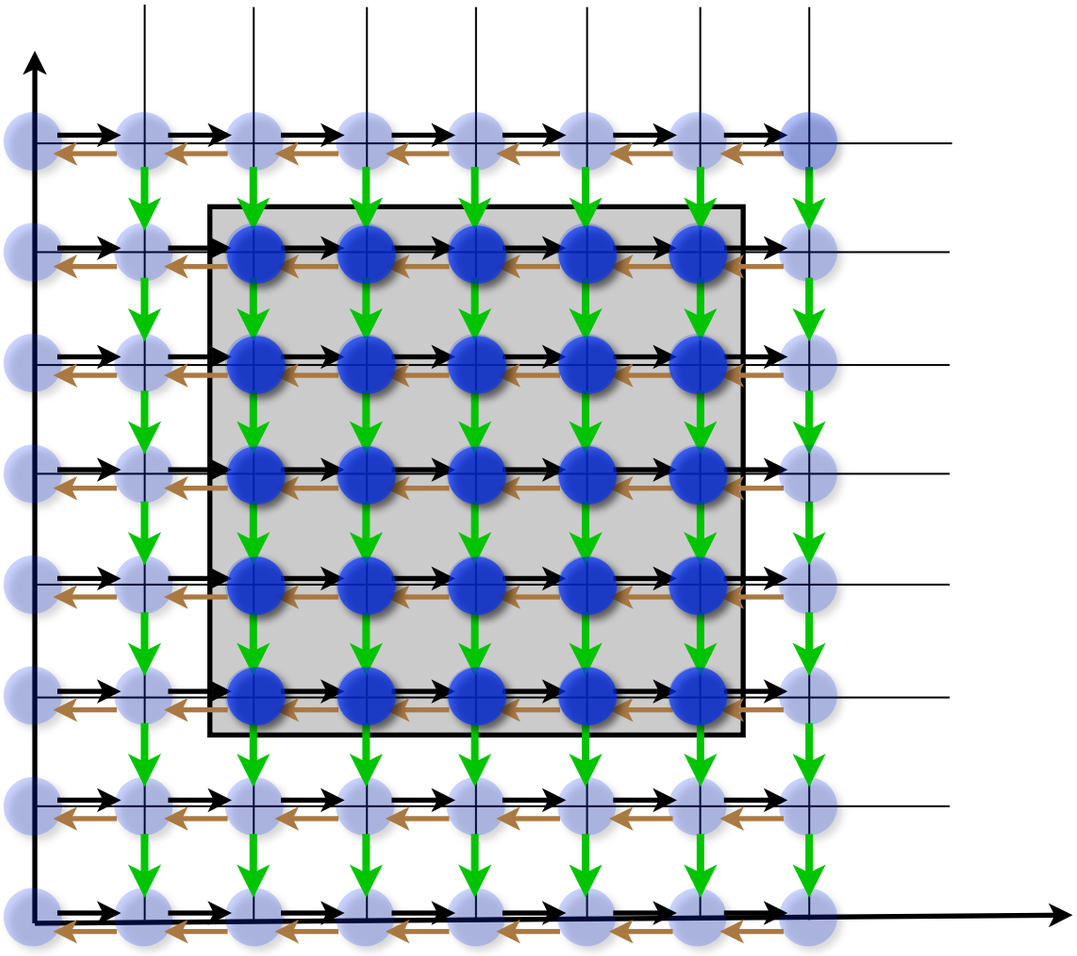
can now be written in matrix form:

$$\dot{P}(\mathcal{X},t) = \mathbf{A} \cdot P(\mathcal{X},t)$$

The Finite State Projection Approach

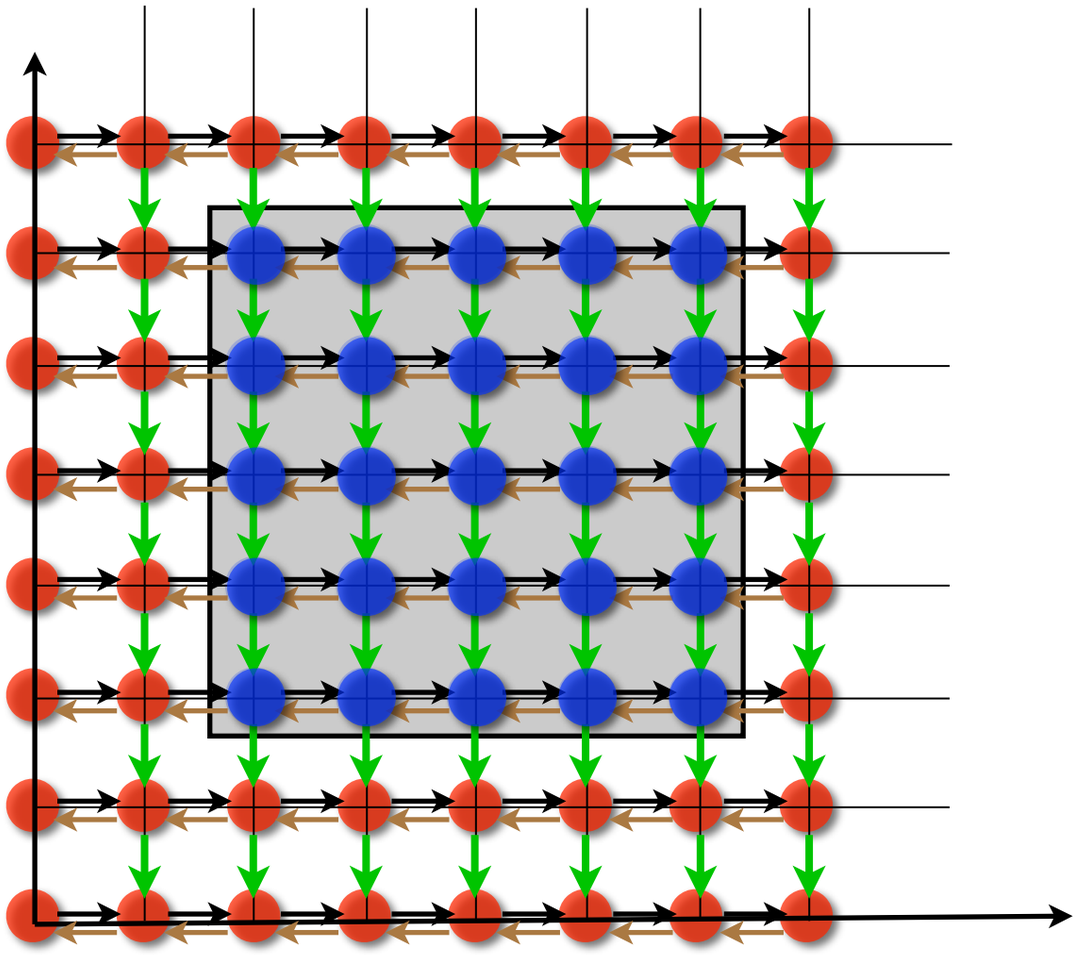


The Finite State Projection Approach



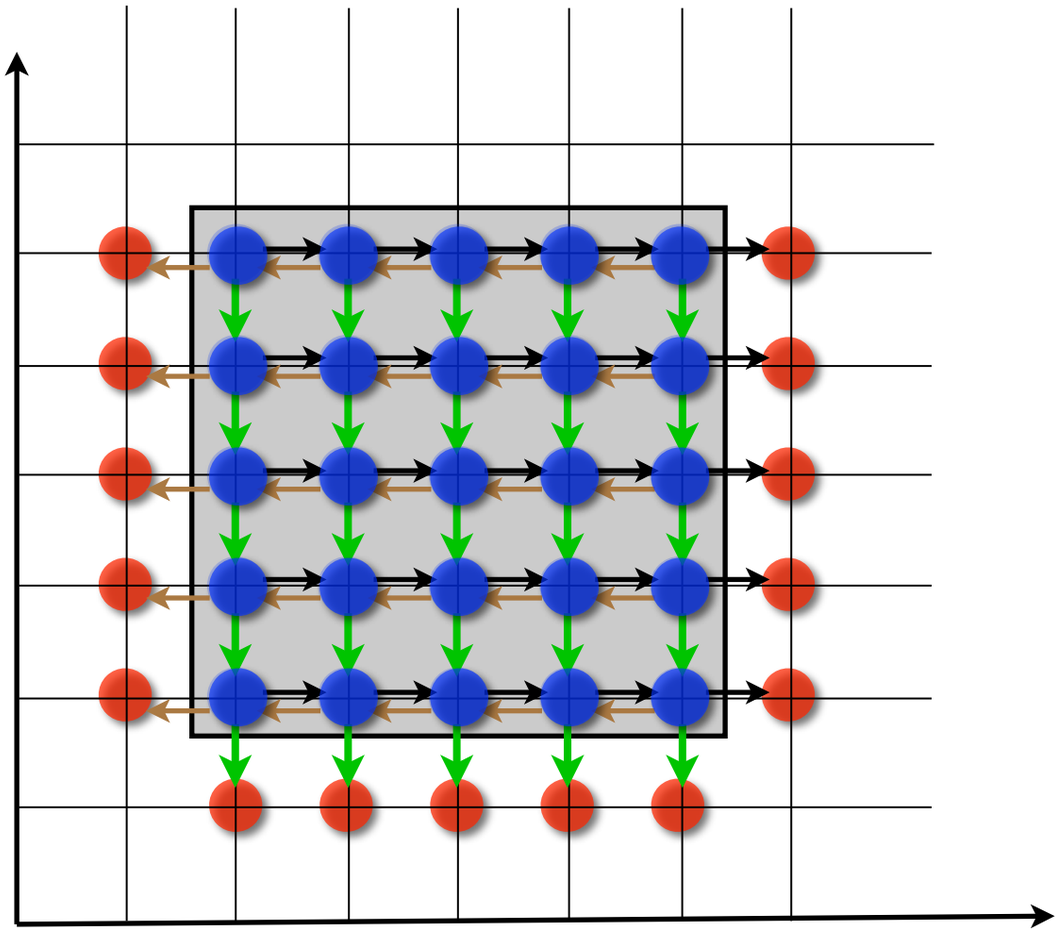
- A finite subset is appropriately chosen

The Finite State Projection Approach



- A finite subset is appropriately chosen
- The remaining (infinite) states are projected onto a single state (red)

The Finite State Projection Approach



- A finite subset is appropriately chosen
- The remaining (infinite) states are projected onto a single state (red)
- Only transitions into removed states are retained

The projected system can be solved exactly!

Finite Projection Bounds

Notation: For a matrix A, let A_J to be the principle submatrix of A indexed by J, where $J = [m_1 \dots m_n]$.

Projection Error Bounds Consider any Markov process described by the Forward Kolmogorov Equation:

$$\dot{P}(\mathcal{X}; t) = A \cdot P(\mathcal{X}; t).$$

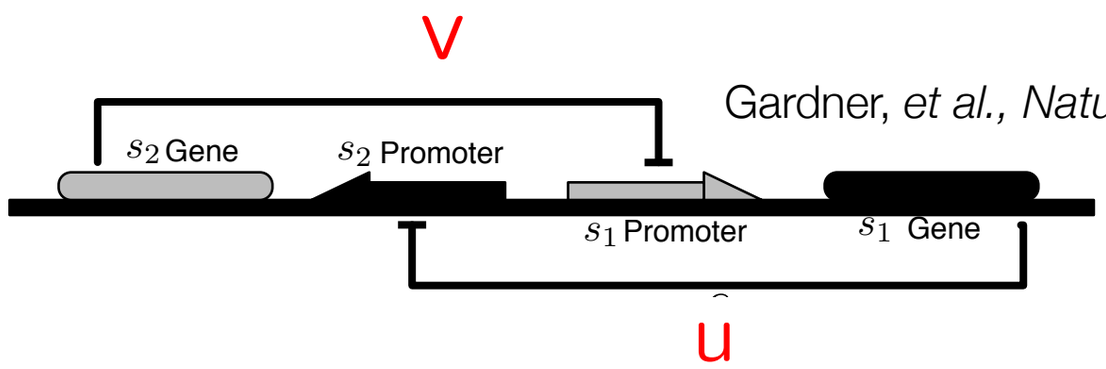
If for an indexing vector J: $\mathbf{1}^T \exp(A_J T) P(\mathcal{X}_J; 0) \geq 1 - \epsilon$, then

$$\left\| \begin{bmatrix} P(\mathcal{X}_J; t) \\ P(\mathcal{X}_{J'}; t) \end{bmatrix} - \begin{bmatrix} \exp(A_J t) P(\mathcal{X}_J; 0) \\ 0 \end{bmatrix} \right\|_1 < \epsilon \qquad t \in [0, T]$$

Munsky and Khammash, Journal of Chemical Physics, 2006

Example: Analysis of A Synthetic Stochastic Switch

Two repressors, *u* and *v*.



Gardner, *et al., Nature* 403, 339-342 (2000)

v inhibits the production of *u:*

$$a_1(u,v) = \frac{\alpha_1}{1+v^\beta} \quad \nu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

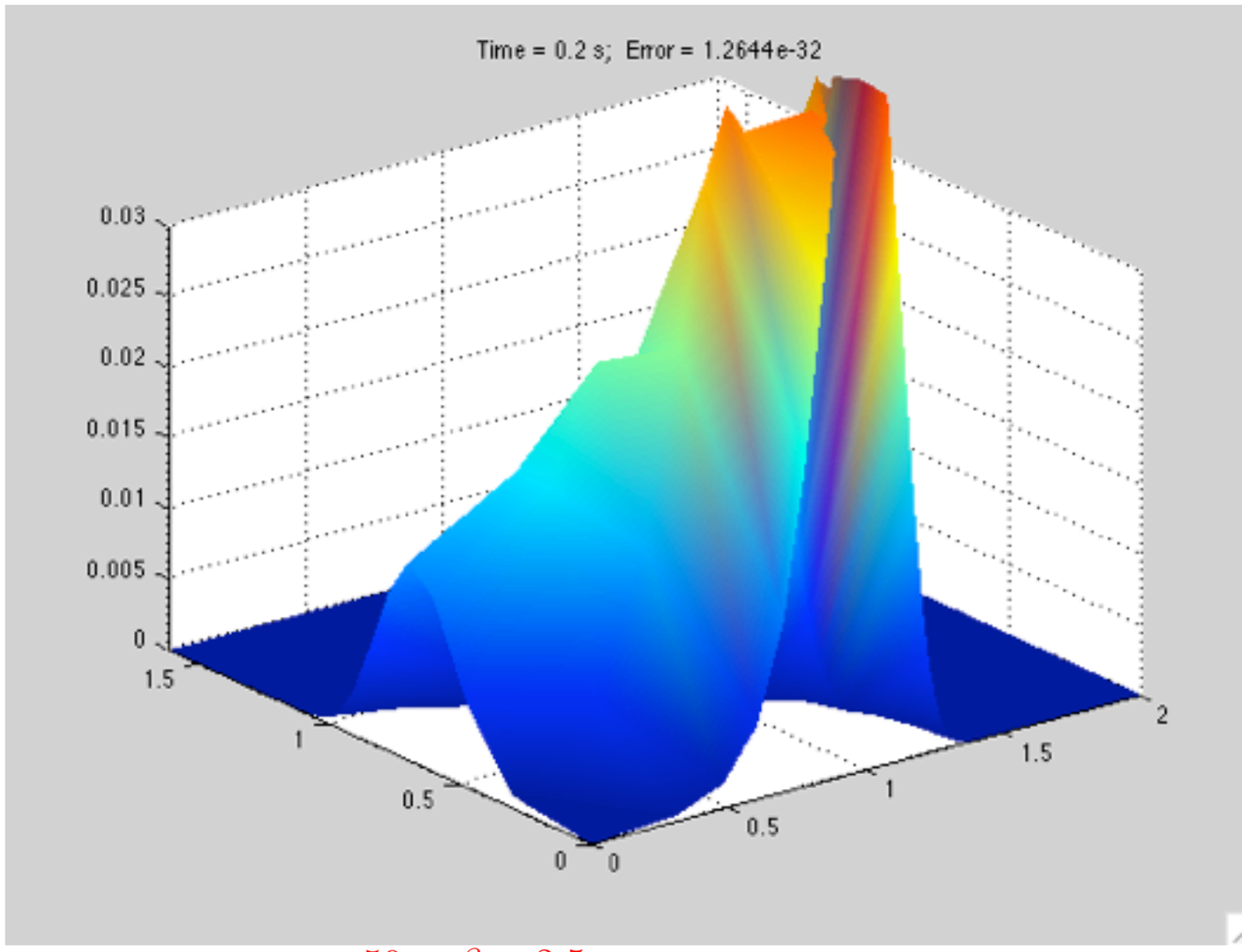
u inhibits the production of *v:*

$$a_3(u,v) = \frac{\alpha_2}{1+u^\gamma} \quad \nu_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

u and *v* degrade exponentially:

$$a_2(u,v) = u \quad \nu_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$a_4(u,v) = v \quad \nu_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$



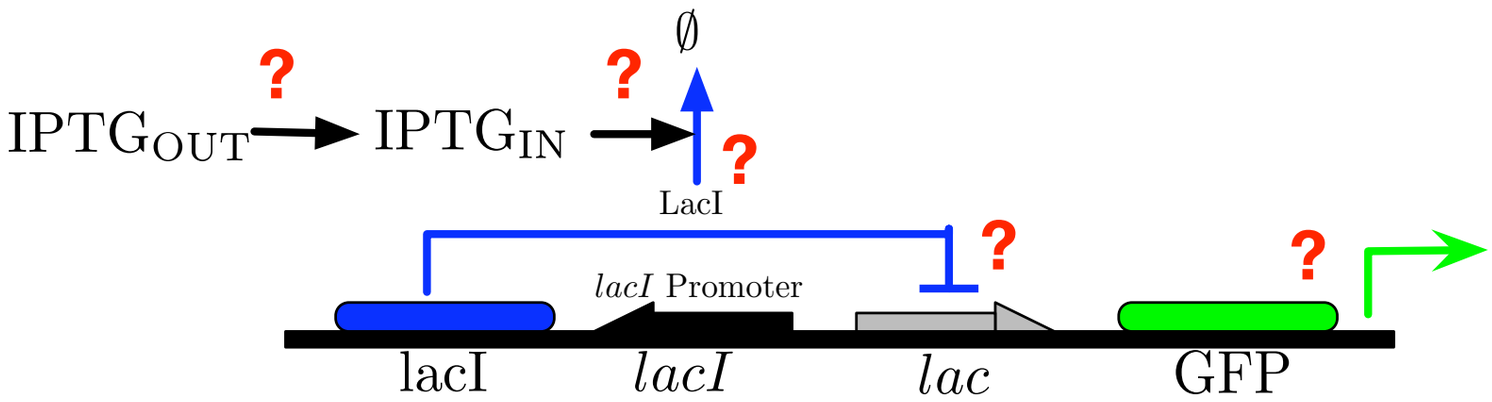
$\alpha_1 = 50 \quad \beta = 2.5$

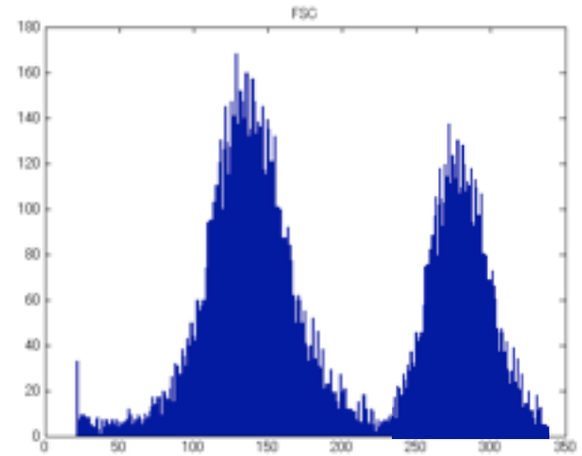
$\alpha_2 = 16 \quad \gamma = 1$

$u(0) = v(0) = 0$

Using Noise to Identify Model Parameters

Why use noise?

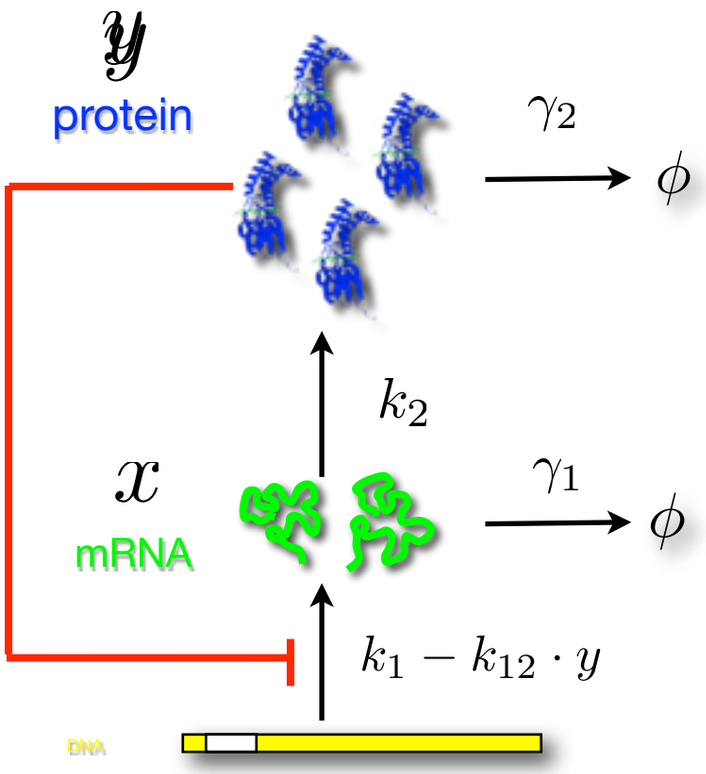




- Noise provides an excitation source for the network dynamics
- Resulting distributions of proteins can be measured
- Such distributions provide a lot of information about the dynamics
- Can they be used to identify model parameters?
- Noise has been used to discriminate among competing models

Dunlop et. al (2008). Nature Genetics. Regulatory activity revealed by dynamic correlations in gene expression noise.

Identification from Moment Information



$$\mathbf{v}(t) := \left[\begin{array}{ccccc} E\{x\} & E\{x^2\} & E\{y\} & E\{y^2\} & E\{xy\} \end{array} \right]^T$$

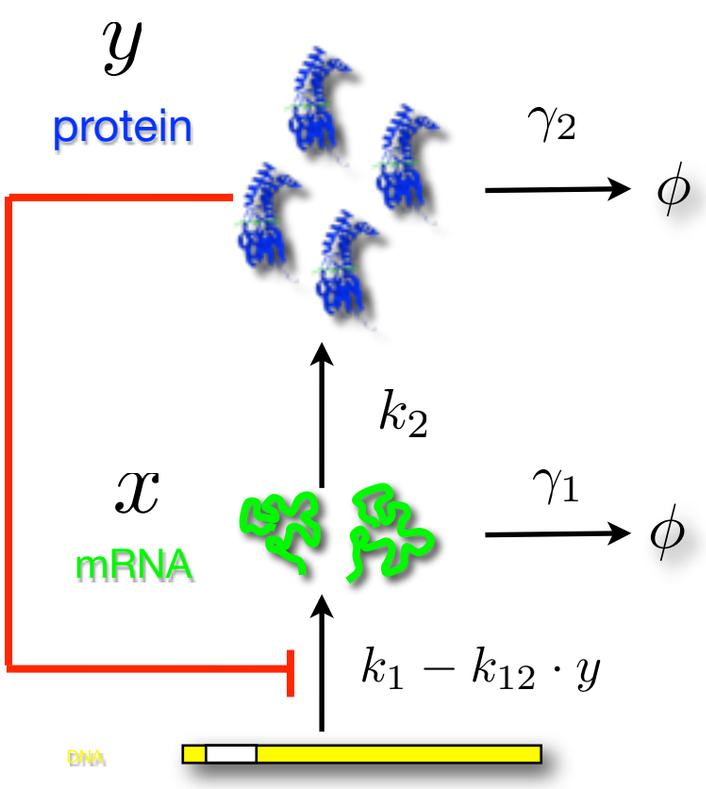
$$\dot{\mathbf{v}} = \left[\begin{array}{ccccc} -\gamma_1 & 0 & k_{21} & 0 & 0 \\ \gamma_1 + 2k_1 & -2\gamma_1 & k_{21} & 0 & 2k_{21} \\ k_2 & 0 & -\gamma_2 & 0 & 0 \\ k_2 & 0 & \gamma_2 & -2\gamma_2 & 2k_2 \\ 0 & k_2 & k_1 & k_{21} & -\gamma_1 - \gamma_2 \end{array} \right] \mathbf{v} + \left[\begin{array}{c} k_1 \\ k_1 \\ 0 \\ 0 \\ 0 \end{array} \right]$$

$$= \mathbf{A}\mathbf{v} + \mathbf{b}$$

Identifiability

Can one identify the parameters $\lambda = \{k_1, \gamma_1, k_2, \gamma_2, k_{21}\}$ from measurements of the moments $\mathbf{v}(t)$?

Identifying Using Steady-State Moments



Can the stationary distribution be used to identify all the parameters?

$$\mathbf{v}(t) := \begin{bmatrix} E\{x\} & E\{x^2\} & E\{y\} & E\{y^2\} & E\{xy\} \end{bmatrix}^T$$

$$\mathbf{v}_\infty = \lim_{t\to\infty} [v_1, v_2, v_3, v_4, v_5]^T.$$

Full Identifiability with Stationary Moments

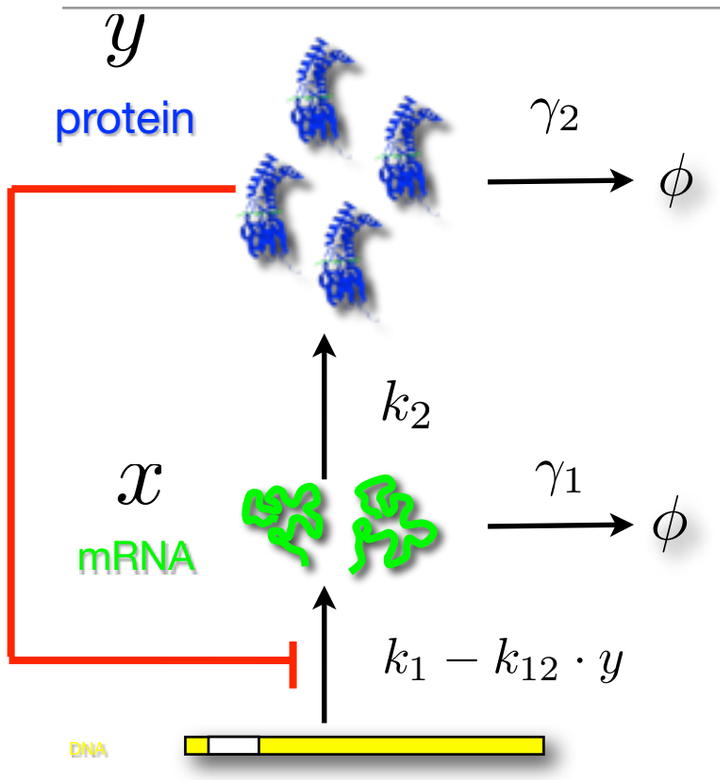
- Full identifiability is impossible using only v_∞.

  Munsky et. al, MSB, 2009

- Identifiability is possible if $\lim_{t\to\infty} E[x(t)x(t+s)]$ is available.

  Cinquemani et al, lect. notes comp. sci, 2009

Identifiability from Transient Time-Measurements



$$\mathbf{v}(t) := \left[\begin{array}{cccccc} E\{x\} & E\{x^2\} & E\{y\} & E\{y^2\} & E\{xy\} \end{array} \right]^T$$

Multiple Measurements

Suppose $\mathbf{v}_j := \mathbf{v}(t_j)$ has been measured at equally separated points in time $\{t_0, t_1, \ldots, t_m\}$

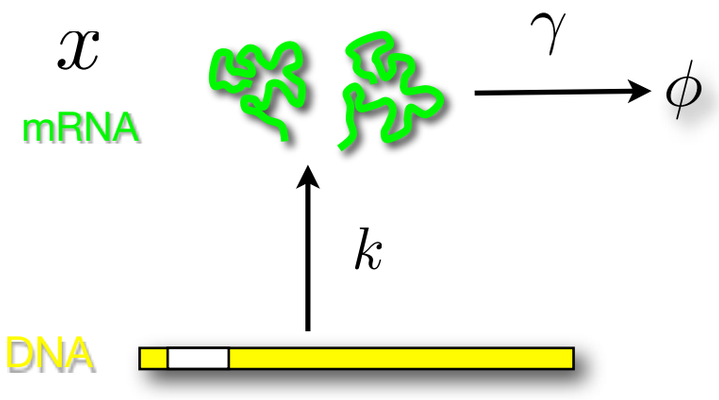
Identifiability with Multiple Moment Measurements

For $m = 6$ the model parameters are *identifiable*.

$$\mathbf{G} = \left[\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_6 \end{array} \right] \left[\begin{array}{ccc} \mathbf{v}_0 & \cdots & \mathbf{v}_5 \\ 1 & \cdots & 1 \end{array} \right]^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix}$$

$$A = \frac{1}{\tau} \log(\mathbf{G}) \qquad \mathbf{b} = -(\mathbf{I} - \mathbf{G})^{-1} \mathbf{A}\mathbf{v}$$

Identification with Two Measurements

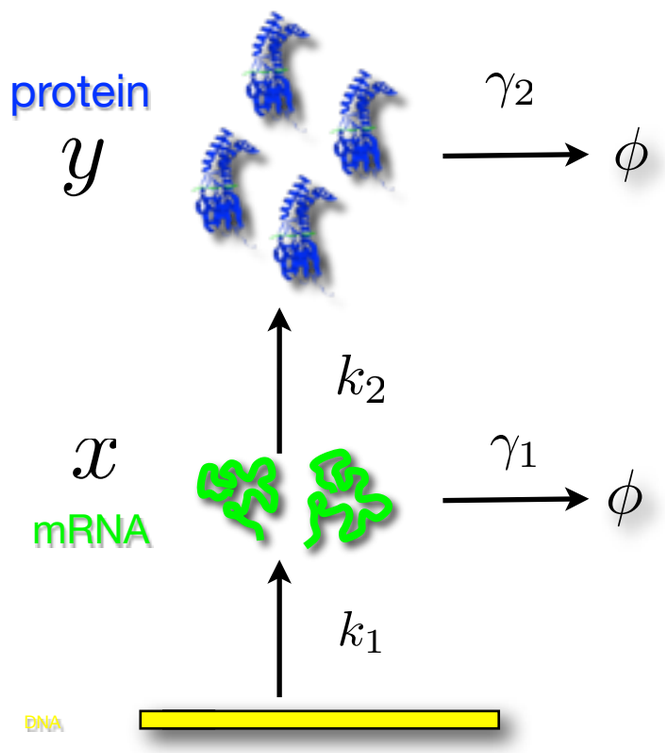


Identifiability of Transcription Parameters

Suppose the mean and variance are known at two times $t_0 < t_1 < \infty$, and define $(\mu_0, \sigma_0) := (\mu(t_0), \sigma(t_0))$ and $(\mu_1, \sigma_1) := (\mu(t_1), \sigma(t_1))$.

Then *the transcription parameters are identifiable*, and

$$\gamma = -\frac{1}{2\tau}\log\left(\frac{\sigma_1^2 - \mu_1}{\sigma_0^2 - \mu_0}\right) \qquad k = \gamma\frac{\mu_1 - \exp(-\gamma\tau)\mu_0}{1 - \exp(-\gamma\tau)}. \qquad (\tau := t_1 - t_0)$$



Identifiability of Transcription & Translation Parameters

$$\mathbf{v}(t) := \begin{bmatrix} E\{x\} & E\{x^2\} & E\{y\} & E\{y^2\} & E\{xy\} \end{bmatrix}^T$$

- Given $\mathbf{v}(t_0)$ and $\mathbf{v}(t_1)$, identifiability of all parameters $k_1, k_2, \gamma_1, \gamma_2$ is generically possible.
- An expression exists for finding the parameters.

Using Densities to Identify Network Parameters

- Moment equations can be written only in special cases.
- Densities (distributions) contain much more information than first two moments.
- Using the Chemical Master Equation, we propose to use density measurements for model identification.

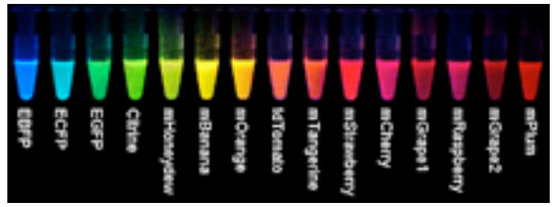
Using Density:

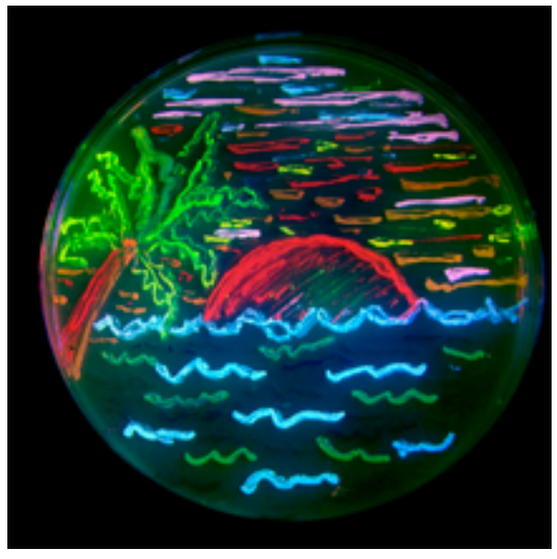
Suppose we measure $\mathbf{P}$ at different times: $\mathbf{P}(t_0), \mathbf{P}(t_1), \ldots, \mathbf{P}(t_{N-1})$

We can use these to identify unknown network parameters λ:

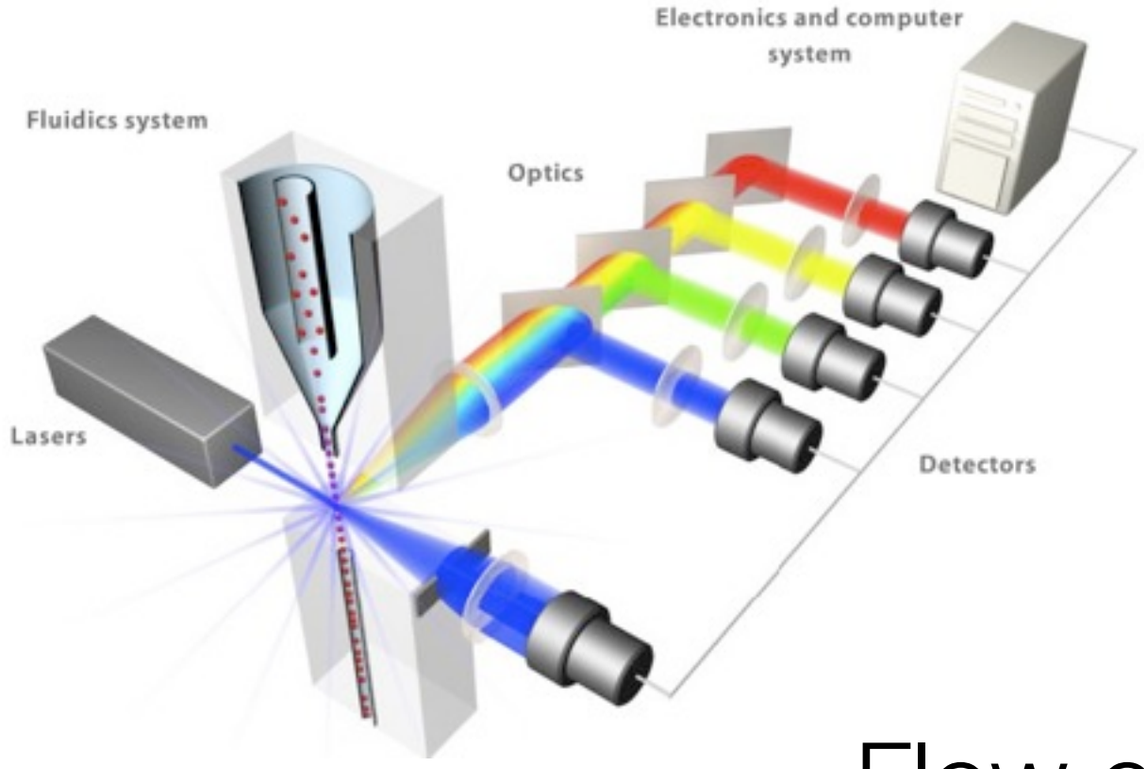
$$\begin{aligned} &\text{Find } \lambda \quad \text{subject to} \\ \dot{\mathbf{P}}^{FSP} &= A(\lambda)\mathbf{P}^{FSP} \\ \mathbf{P}^{FSP}(t_0) &= \mathbf{P}(t_0) \\ \mathbf{P}^{FSP}(t_1) &= \mathbf{P}(t_1) \\ &\vdots \\ \mathbf{P}^{FSP}(t_{N-1}) &= \mathbf{P}(t_{N-1}) \end{aligned}$$

Tsien Lab, UCSD



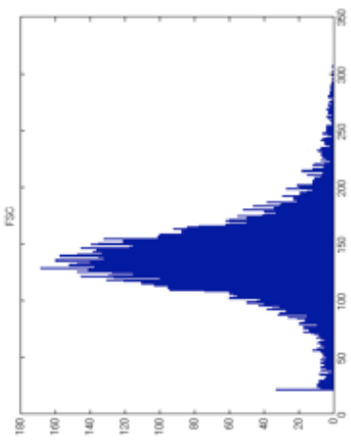


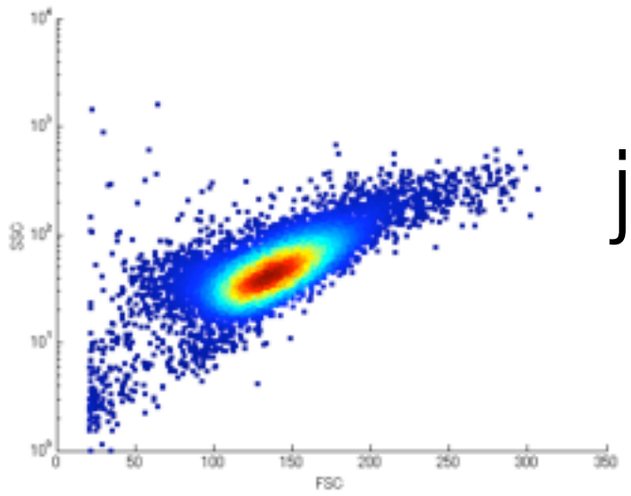
Agar Plate of Fluorescent Bacteria Colonies



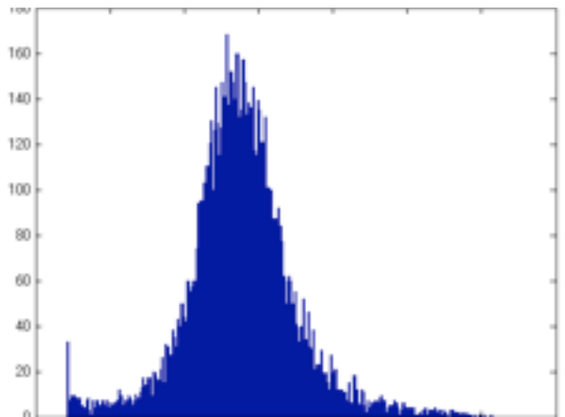
Flow cytometry

protein A histogram



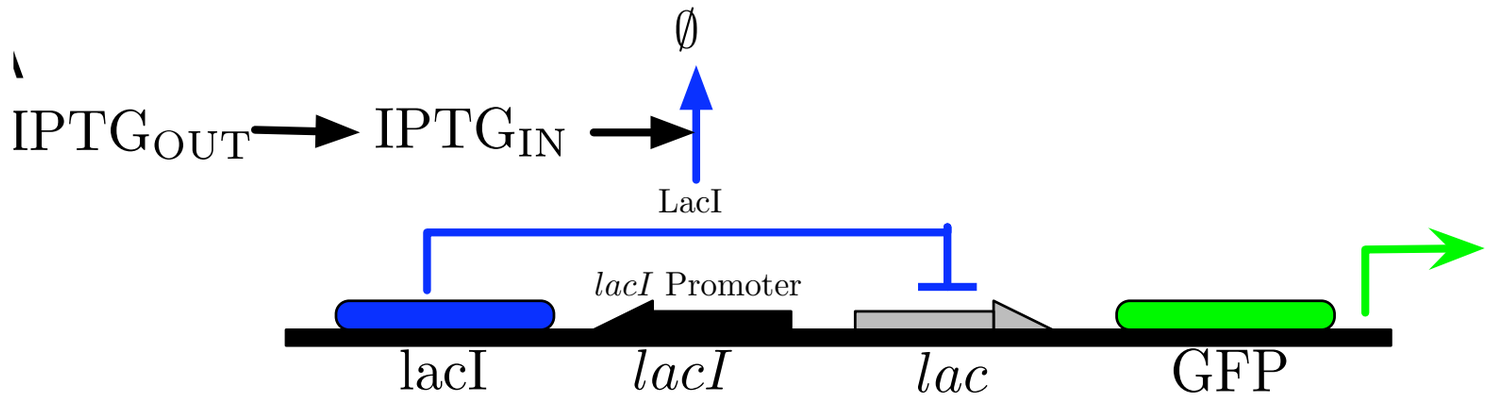


joint pdf



protein B histogram

Identification of *lac* Induction



Model

$$\phi \xrightarrow{k_L} \mathsf{LacI}$$

$$\mathsf{LacI} \xrightarrow{\delta_L} \phi \qquad \delta_L = \delta_L^{(0)} + \delta_L^{(1)}[\mathsf{IPTG}]_{\mathsf{IN}}$$

$$\phi \xrightarrow{w_G} \mathsf{GFP} \qquad w_G = \frac{k_G}{1 + \alpha[\mathsf{LacI}]^{\eta}},$$

$$\mathsf{GFP} \xrightarrow{\delta_G} \phi$$

$$\mathsf{IPTG}_{\mathsf{IN}} = \mathsf{IPTG}_{\mathsf{OUT}}(1 - e^{-rt})$$

9 unknown parameters!

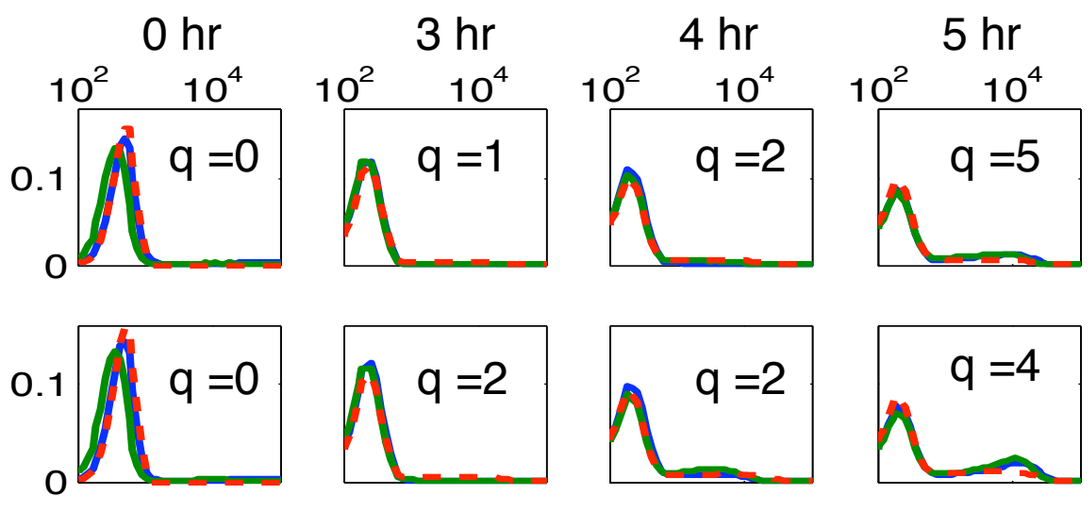
Experiment

- *E. coli* strain DL5905
- Induced with different IPTG concentrations: 5,10, 20, 40, 100 uM
- Induction times: 0, 1, 2, 3, 4, 5 hours before flow cytometry

Identified Parameters

$$\left\{ \begin{array}{lll} k_L = 1.7 \times 10^{-3}\ \mathrm{s}^{-1} & k_G = 1.0 \times 10^{-1}\ \mathrm{s}^{-1} & \eta = 2.1 \\ \delta_L^{(0)} = 3.1 \times 10^{-4}\ \mathrm{N}^{-1}\mathrm{s}^{-1} & \delta_L^{(1)} = 5.0 \times 10^{-2}\ (\mu\mathrm{M}\cdot\mathrm{N})^{-1}\mathrm{s}^{-1} & \alpha = 1.3 \times 10^{4}\ \mathrm{N}^{-\eta} \\ r = 2.8 \times 10^{-5}\ \mathrm{s}^{-1} & \mu_{\mathrm{GFP}} = 220\ \mathrm{AU} & \sigma_{\mathrm{GFP}} = 390\ \mathrm{AU} \end{array} \right\}$$

Identified Model vs. Experiment



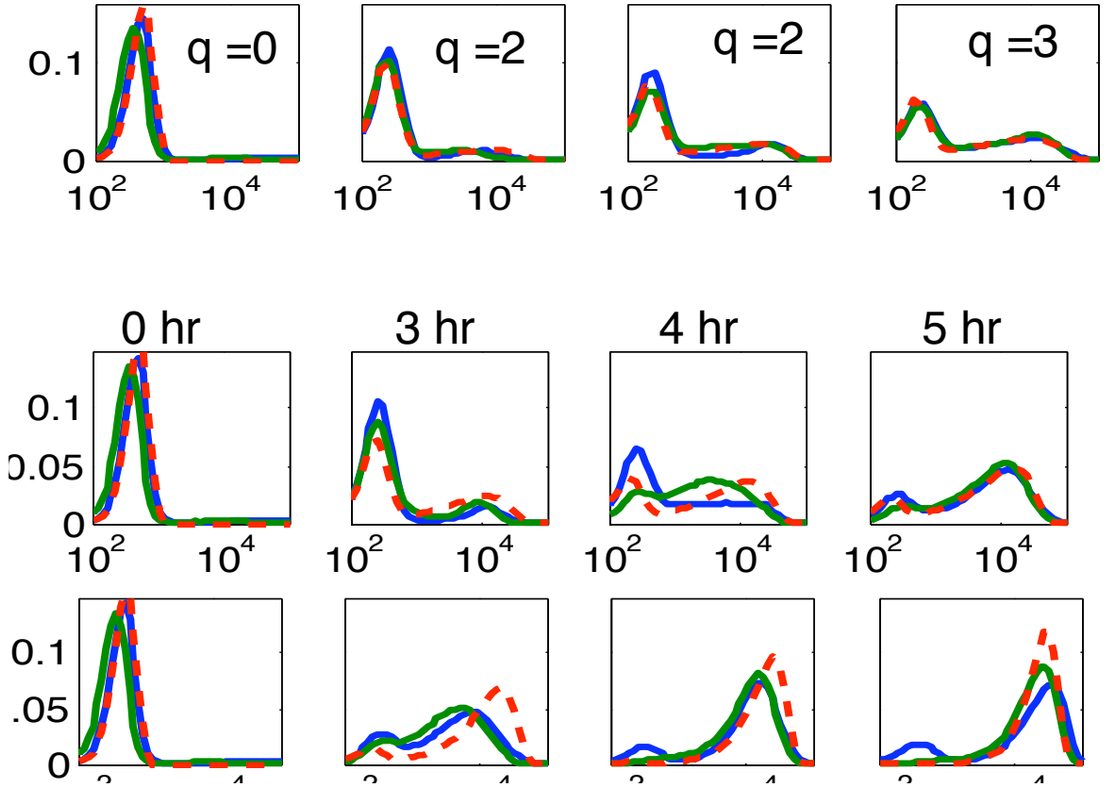
$5\ \mu\mathrm{M}$

$10\ \mu\mathrm{M}$

Model

Experiment

Model Predictions



$20\ \mu\mathrm{M}$

$40\ \mu\mathrm{M}$

$100\ \mu\mathrm{M}$

B. Munsky, B. Trinh, M. Khammash, *Nature Molecular Systems Biology*, in press.

Conclusions

- Randomness “noise” leads to cell-cell variability
  - Stochastic models are necessary
- Some stochastic analysis tools available (more needed)
  - Kinetic Monte Carlo
  - Moment approximation
  - Linear noise approximation (van Kampen)
  - Density computation (FSP)
- Noise reveals network parameters
  - Enabling technologies: flow cytometry and FISH/microscopy
  - A small number of transient measurements suffices
  - FSP exploits full pdf measurements
  - Cellular noise (process noise) vs. measurement noise (output noise)

Acknowledgement

- Brian Munsky, UCSB, LANL

- Brooke Trinh UCSB (lac induction)