Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Probabilistic Methods for Biochemical Reaction Networks

Mustafa Khammash
Department of Biosystems Science and Engineering
ETH Zürich

LECTURE I

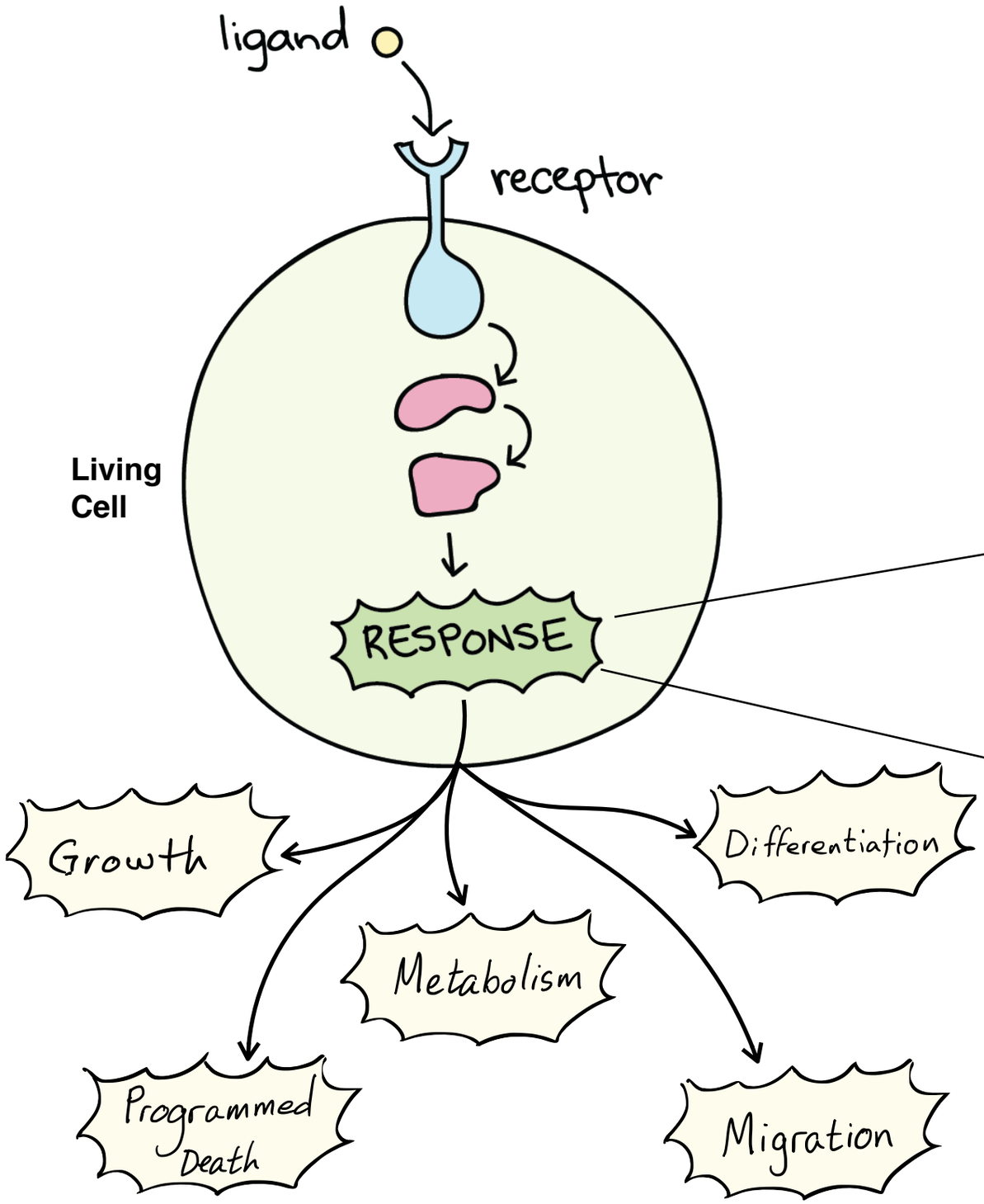
Outline

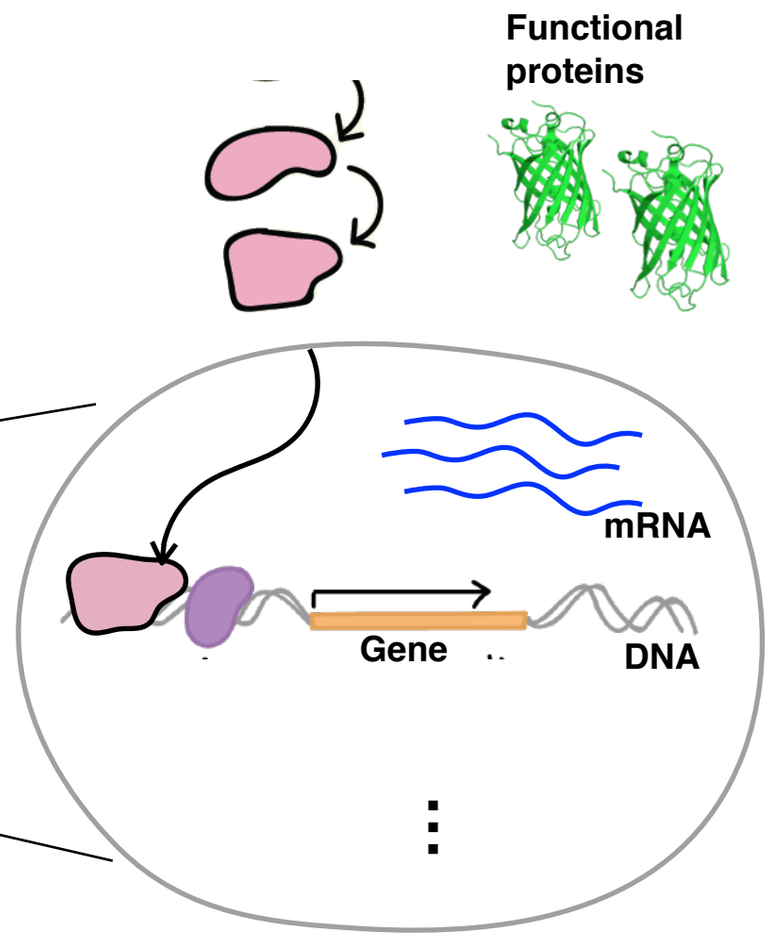
- A gentle introduction to molecular biology
  - The biology of gene expression
  - Measuring gene expression
  - Variability in gene expression and its consequences
  - Motivation for using probabilistic models

- Introduction to stochastic modeling and analysis
  - The Chemical Master Equation
  - Using biological data for model inference

- Controlling gene expression mean and variance

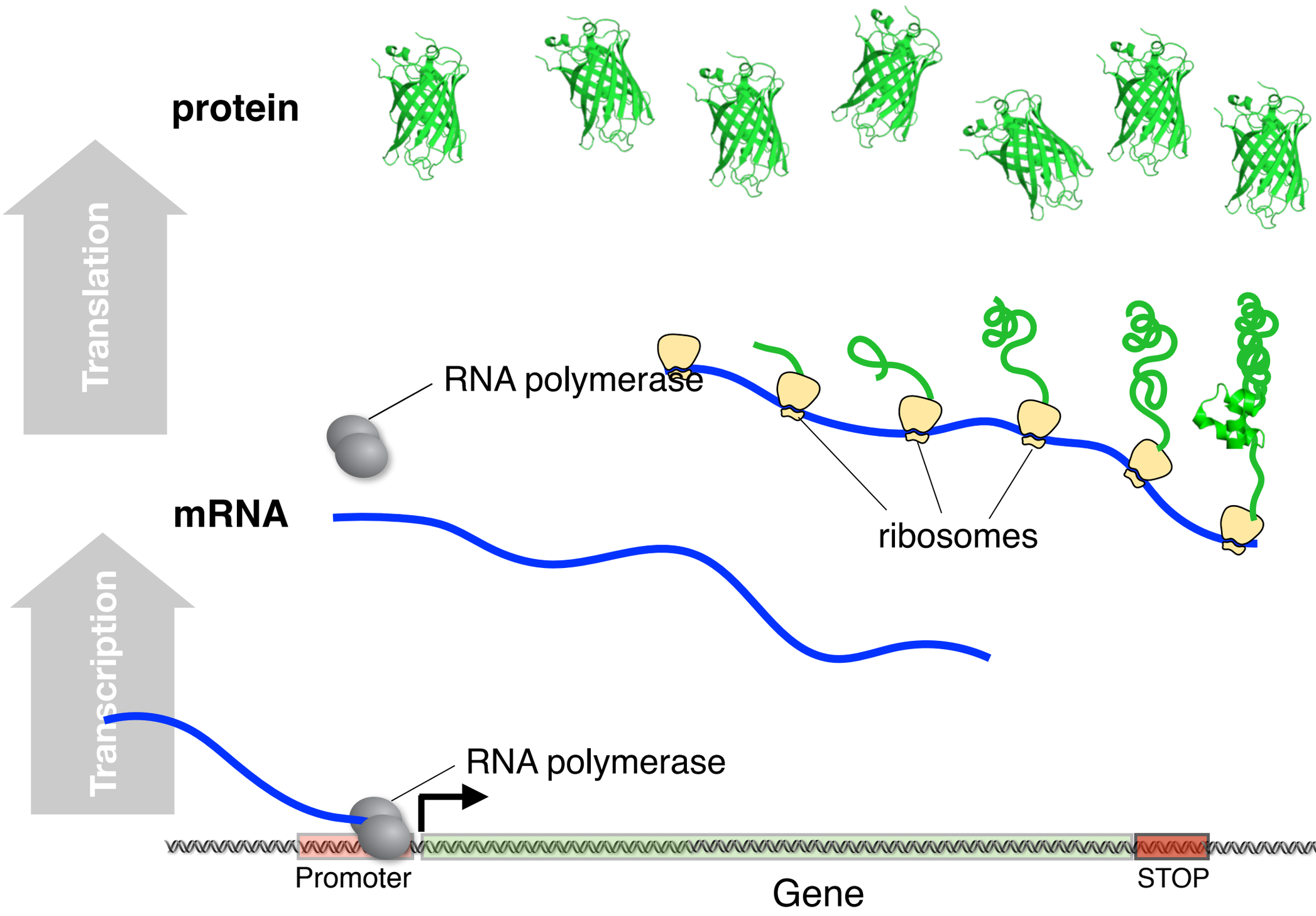
From Stimulus to Response





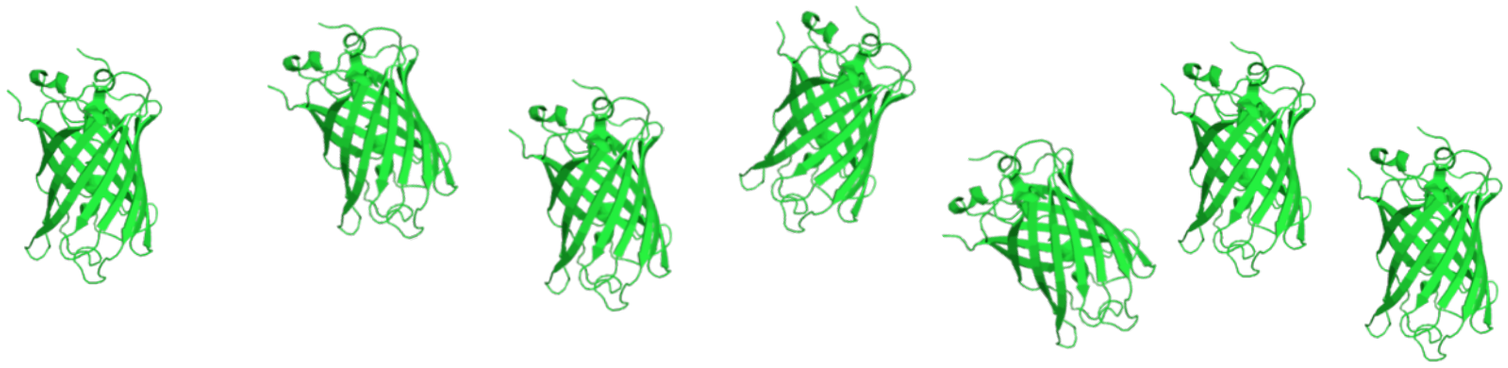
Gene expression uses information in DNA to synthesize cellular machinery

Gene Expression: From DNA to Protein



Regulation of Gene Expression: Activation

protein



RNA polymerase

High Expression

Activator

Promoter

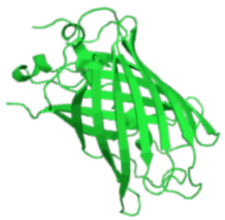
Gene

STOP

DNA

Regulation of Gene Expression: Repression

protein



RNA polymerase

Repressor

Low Expression

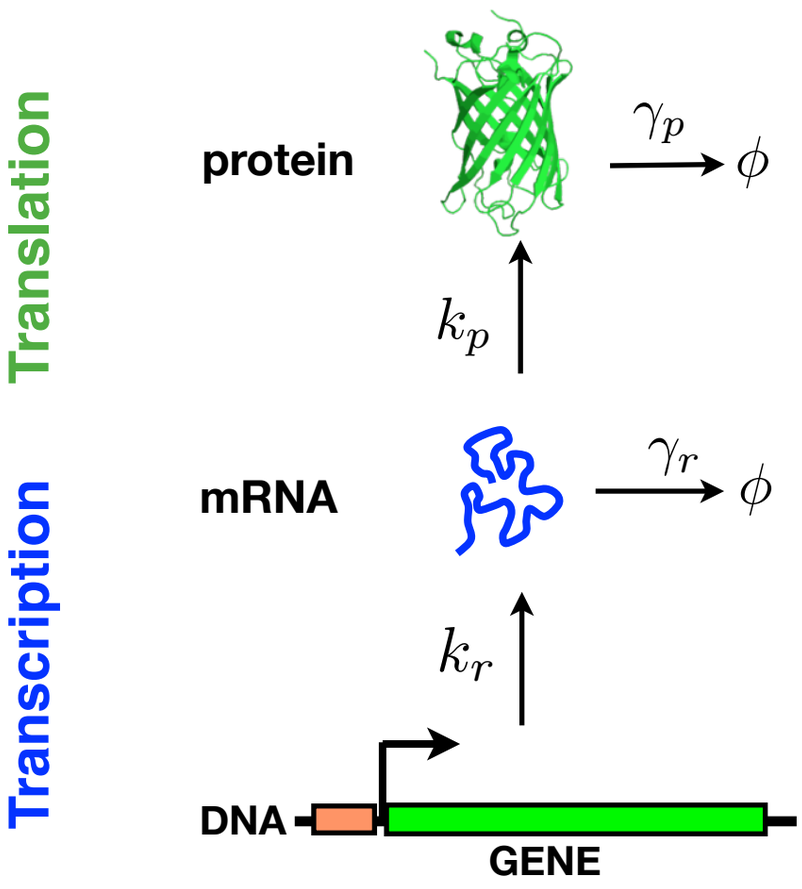
Promoter

Gene

STOP

DNA

Modeling Gene Expression



protein reactions

$$\begin{cases} \text{mRNA} \xrightarrow{k_p} \text{mRNA} + \text{protein} \\ \text{protein} \xrightarrow{\gamma_p} \phi \end{cases}$$

mRNA reactions

$$\begin{cases} \phi \xrightarrow{k_r} \text{mRNA} \\ \text{mRNA} \xrightarrow{\gamma_r} \phi \end{cases}$$

Gene Expression Dynamics

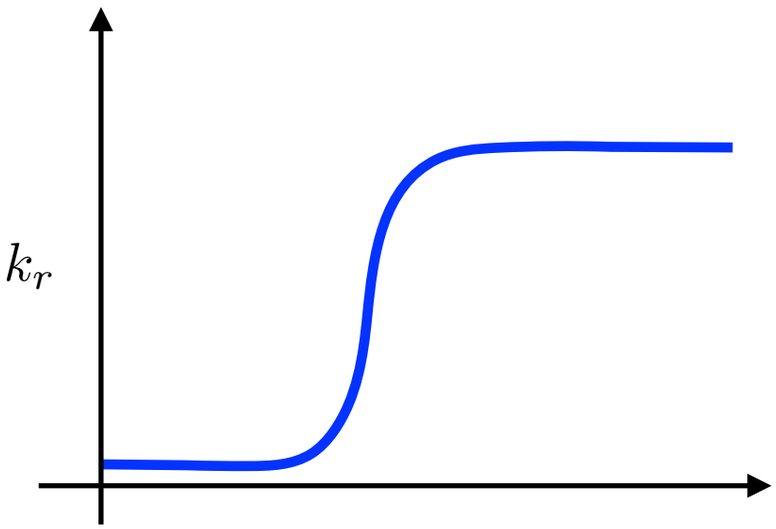
$$\dot{r} = k_r - \gamma_r r$$

$$\dot{p} = k_p r - \gamma_p p$$

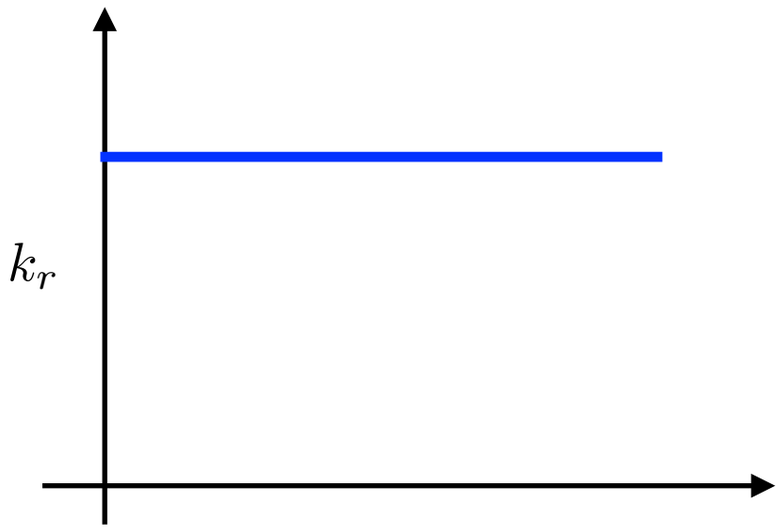
$r(t)$ — RNA concentration at time *t*

$p(t)$ — protein concentration at time *t*

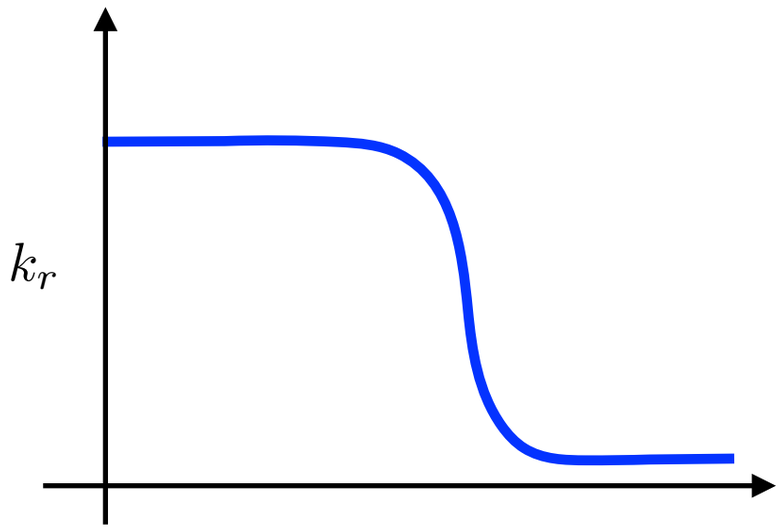
Transcription rate depends on transcription factor concentration



positively regulated gene



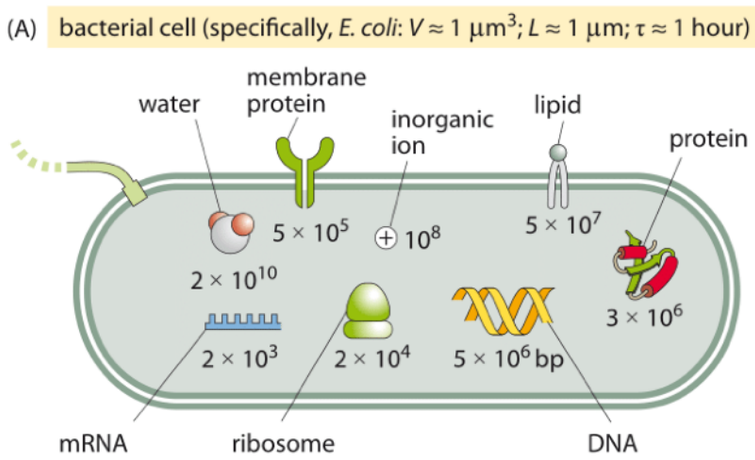
constitutively regulated gene



negatively regulated gene

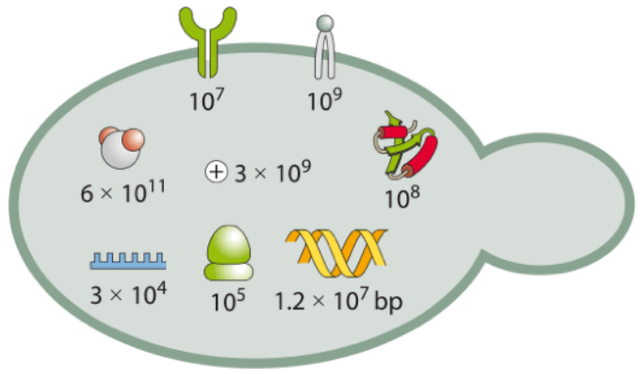
Common Cell Types Studied in Molecular Biology

E. coli



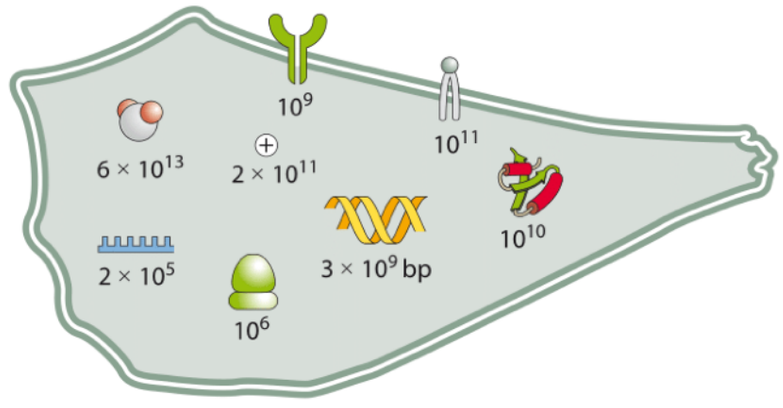
~ 4300 genes

Yeast



~ 5700 genes

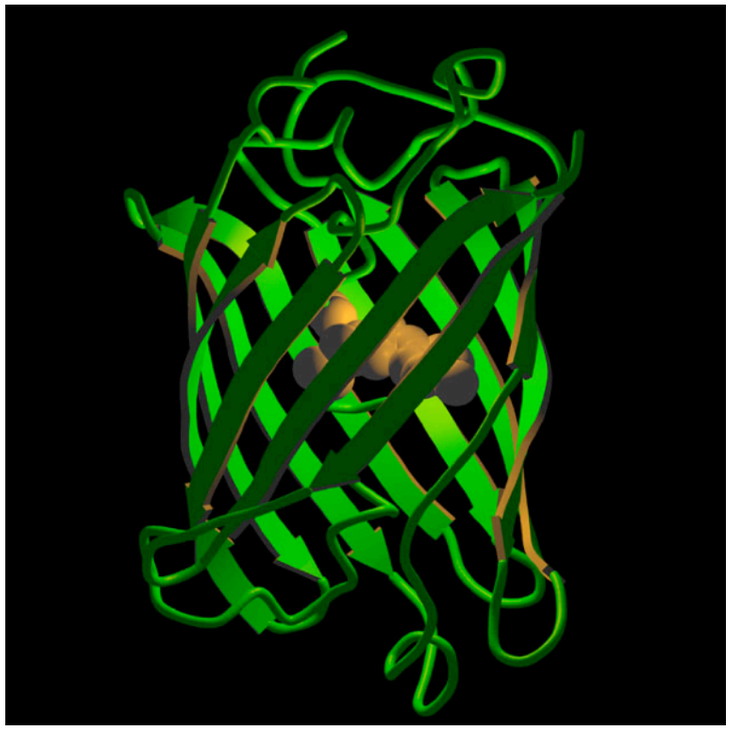
Mammalian

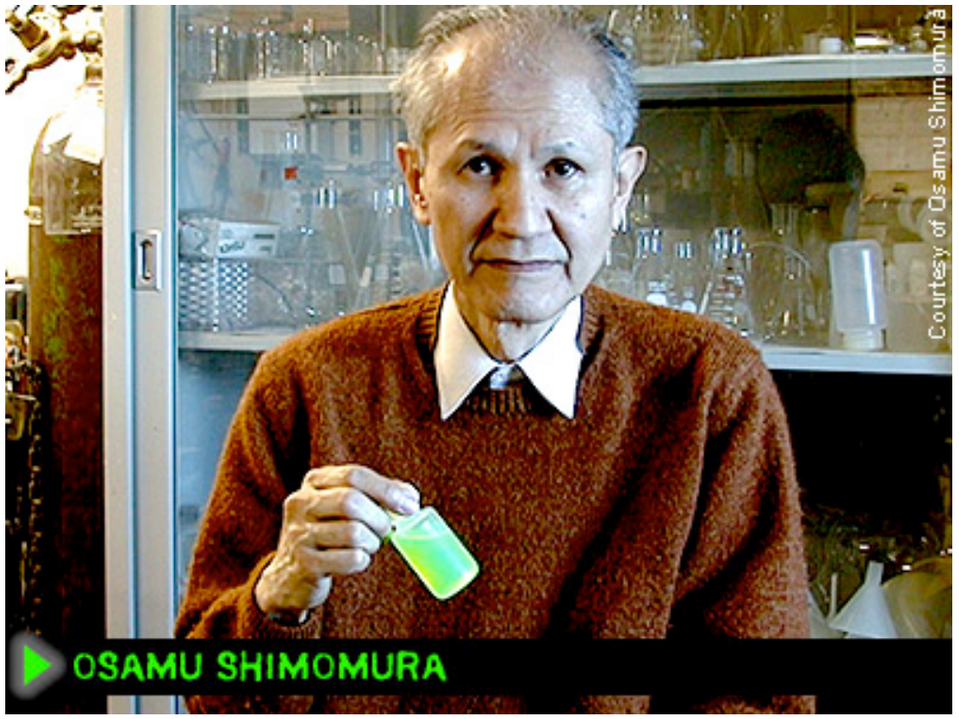


~ 20,000 genes

Cell Biology by the Numbers, 2017

How do we measure cellular proteins?



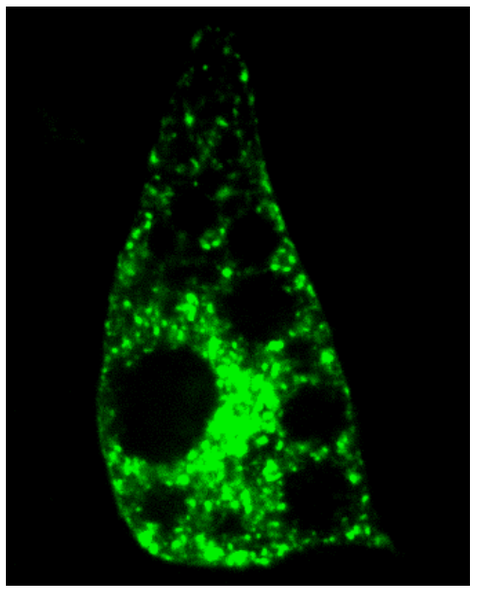


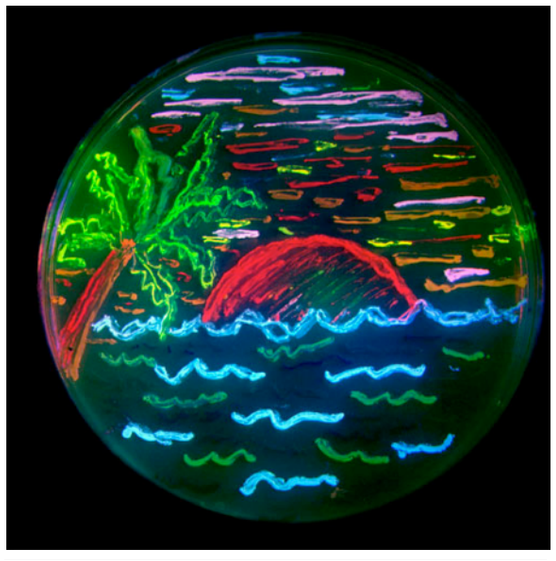
Jellyfish *Aequorea victoria*

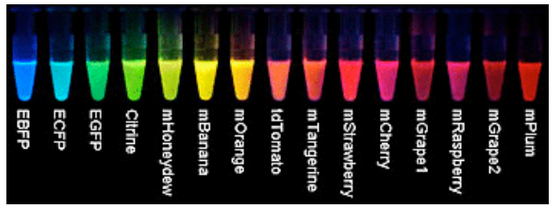
Nobel Prize in Chemistry, 2008
Osamu Shimomura, Martin Chalfie and Roger Tsien

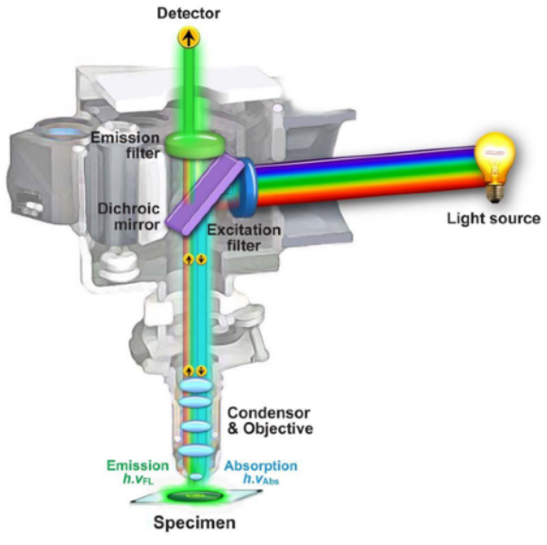
“for the discovery and development of the green fluorescent protein, GFP”

Measuring Cellular Proteins

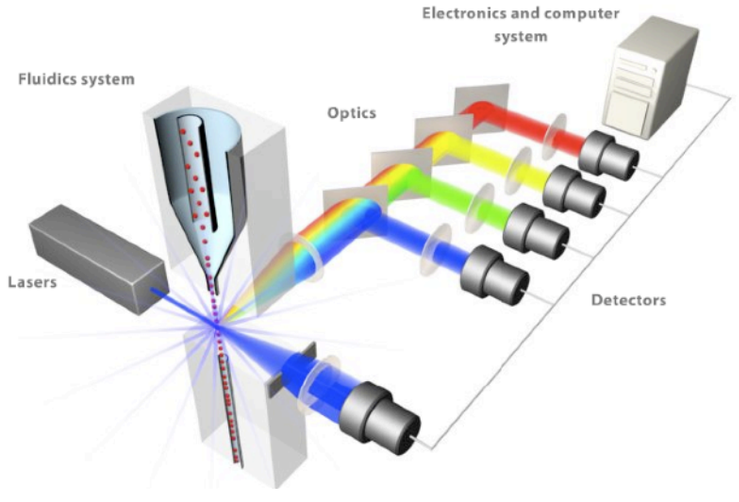








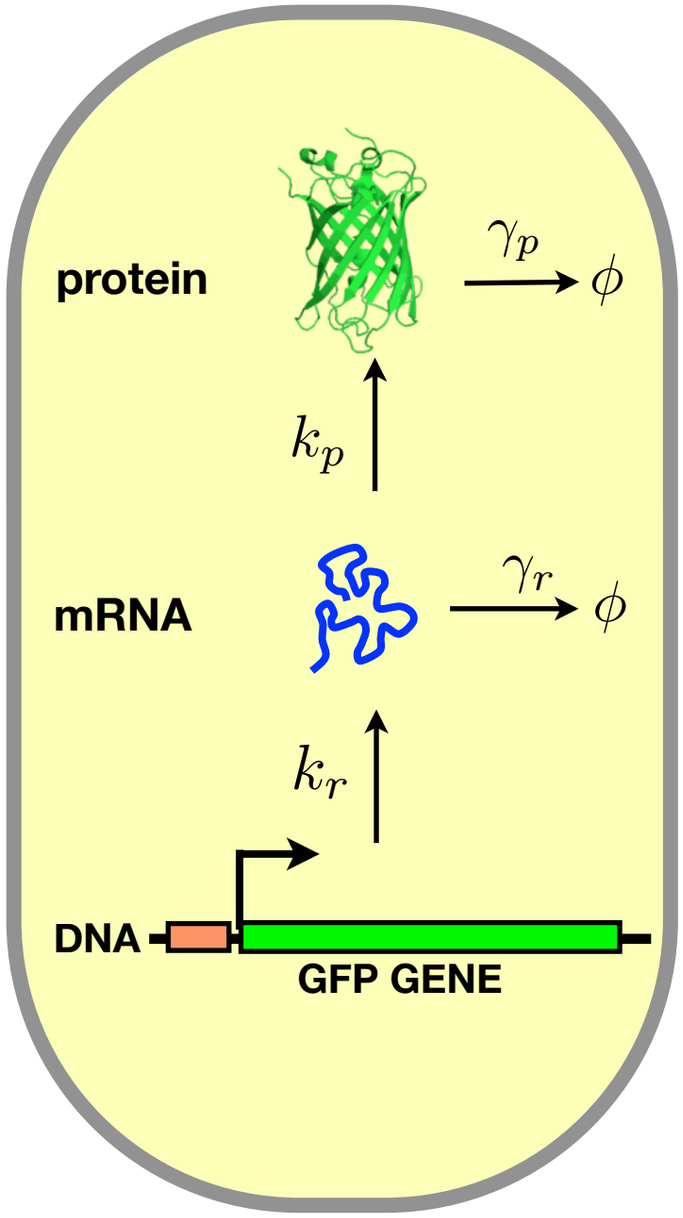
Microscopy



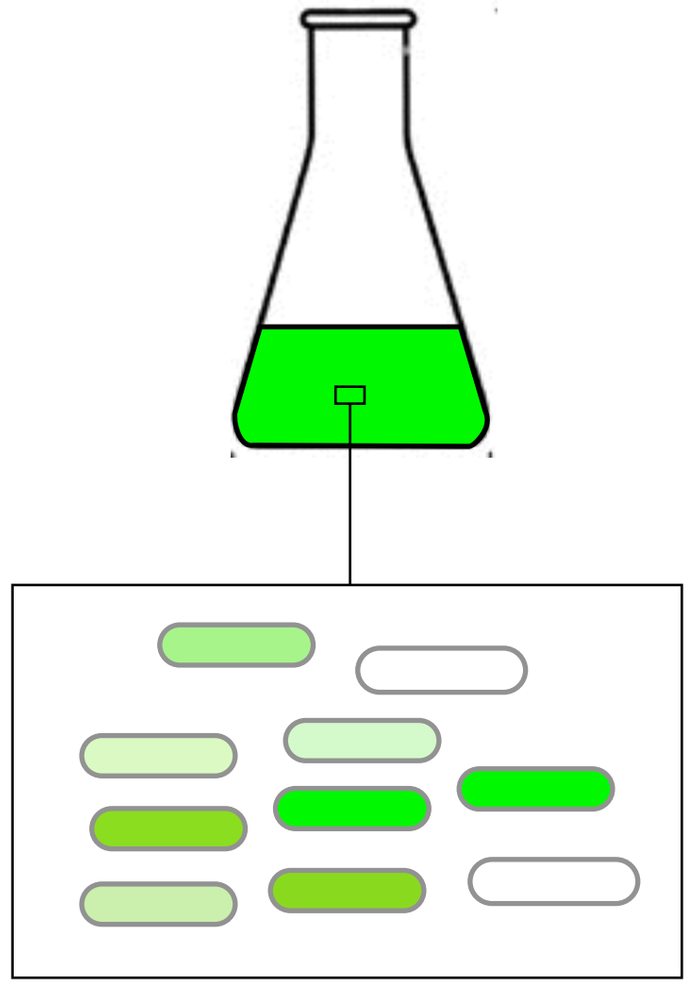
Flow Cytometry

Experimental Evidence of Random Variability in Gene Expression

Bacterial Cell

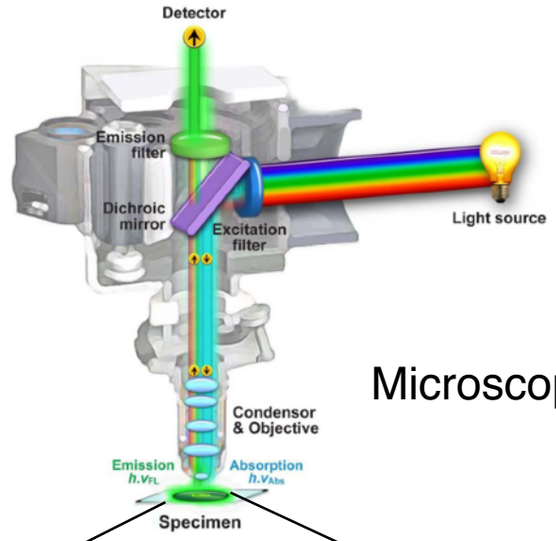


Cell Population

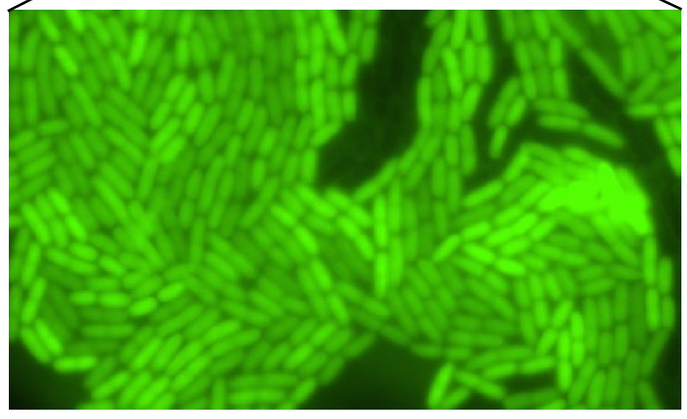


Fluorescence intensity proportional to protein level

Single Cells

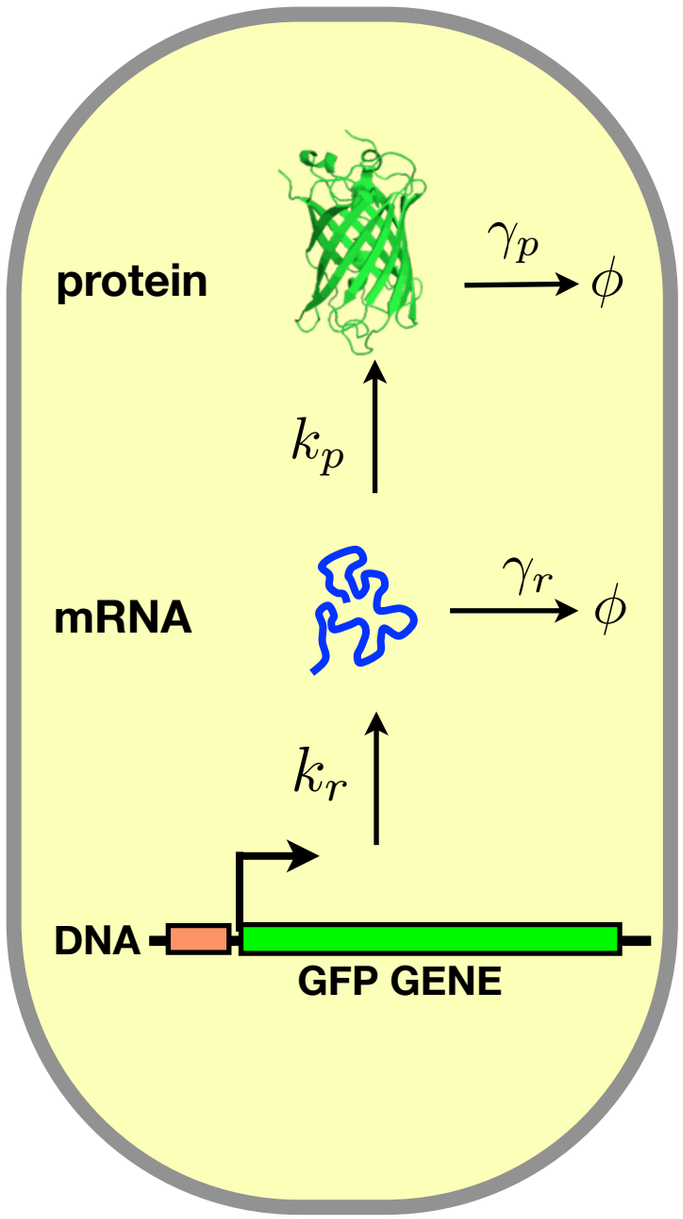


Microscopy

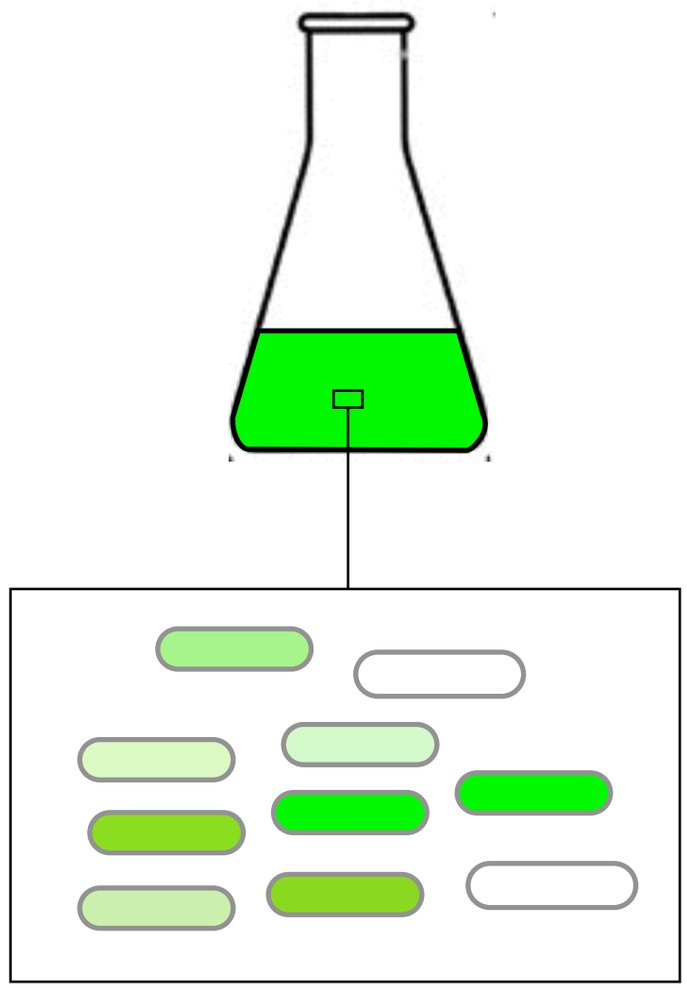


Quantifying Variability in Gene Expression

Single Cell

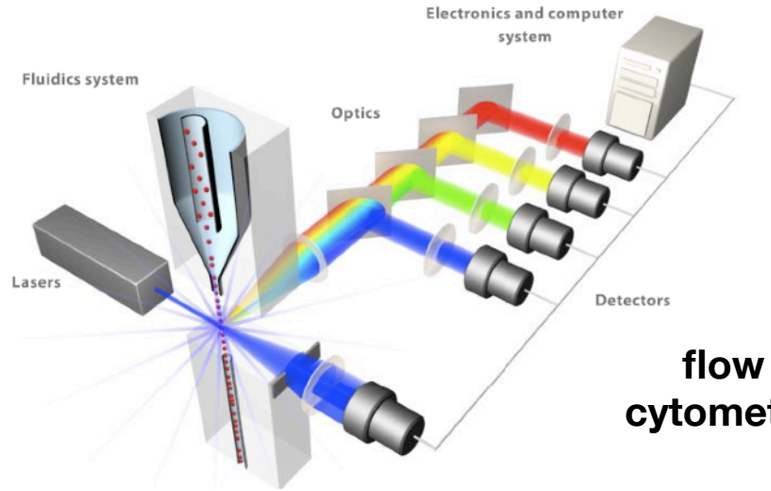


Cell Population

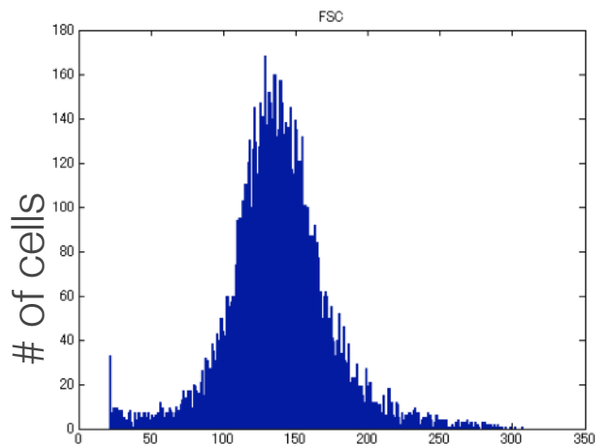


Fluorescence intensity
proportional to protein level

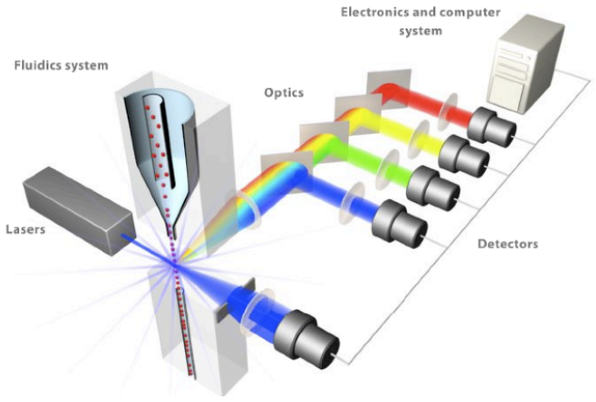
Quantifying Variability



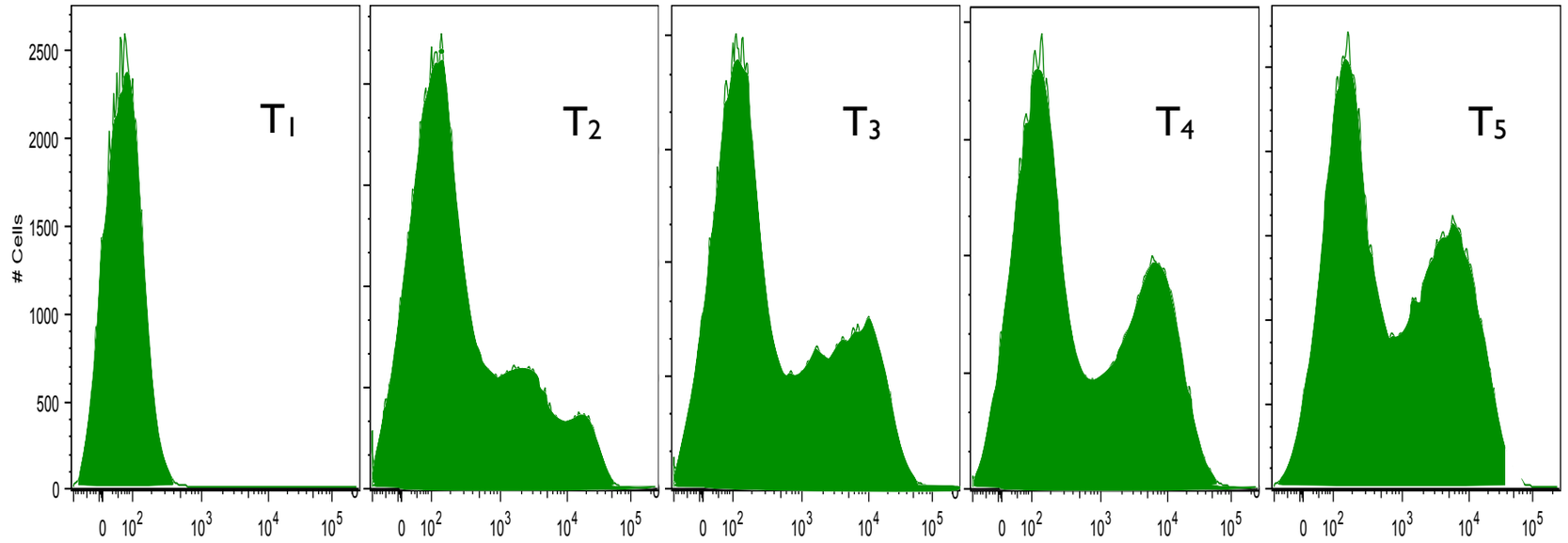
flow cytometry

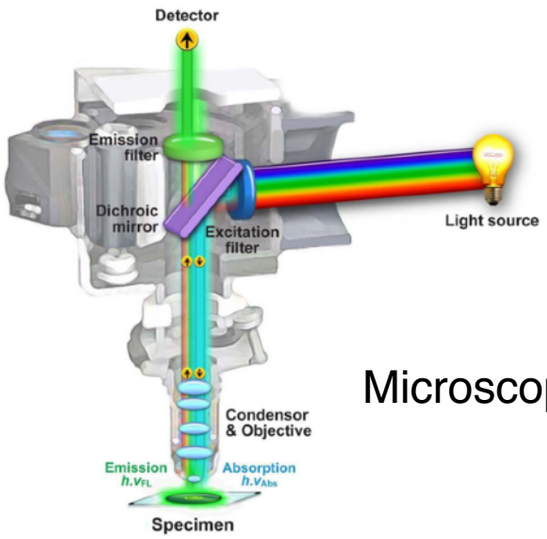


Two Types of Time-Resolved Data

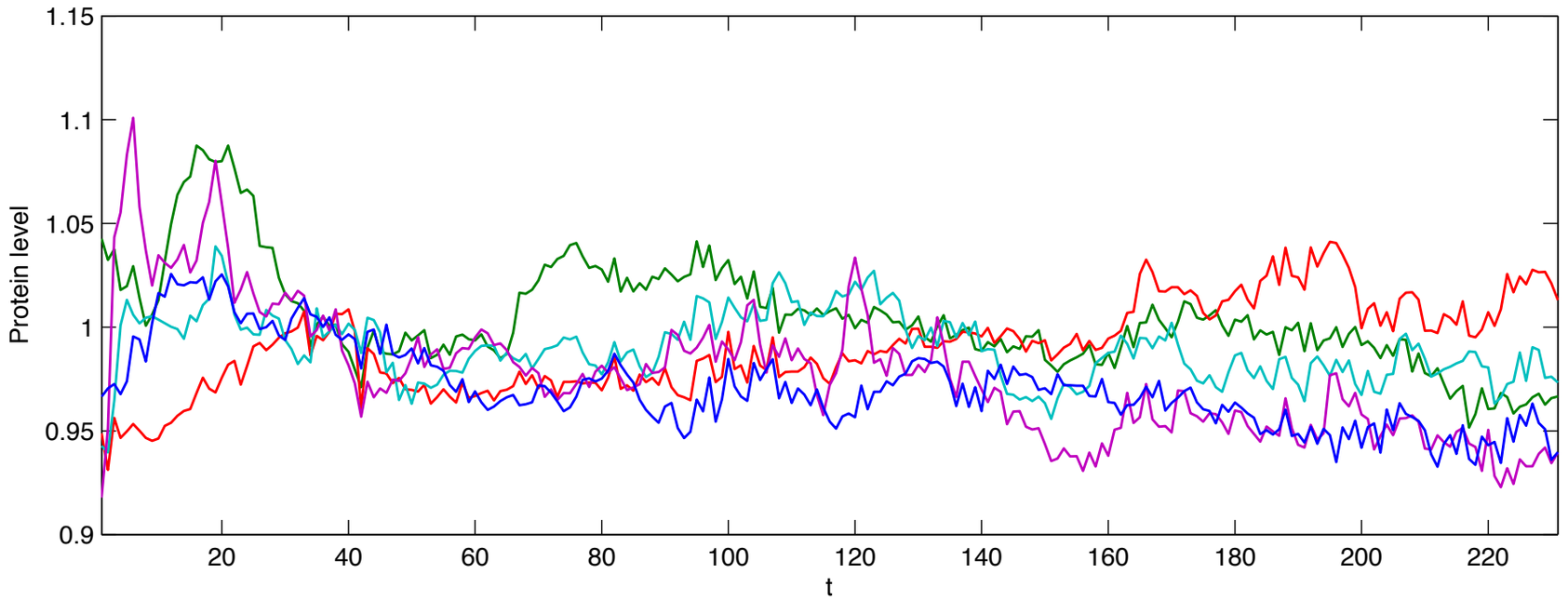


Flow cytometry



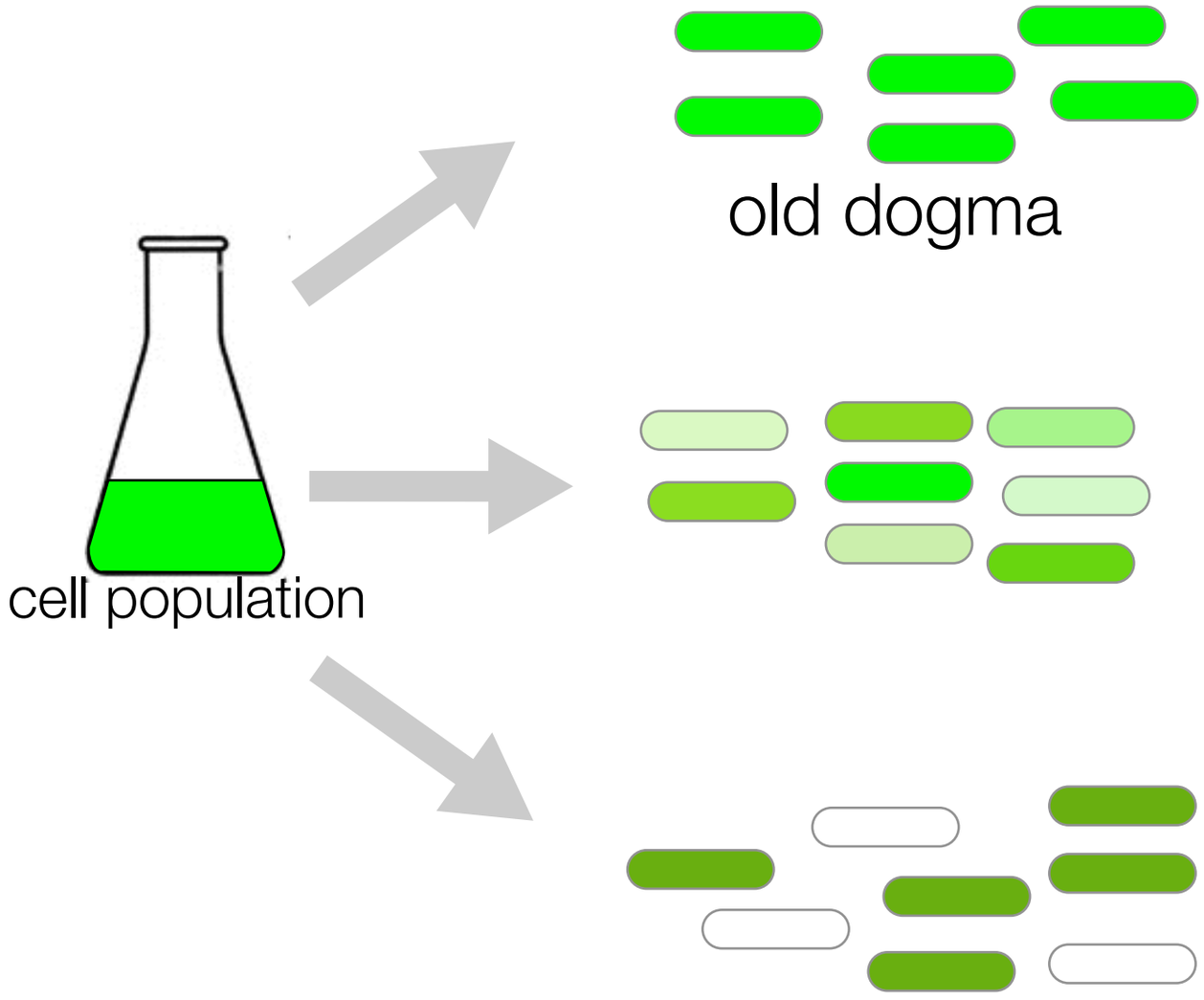


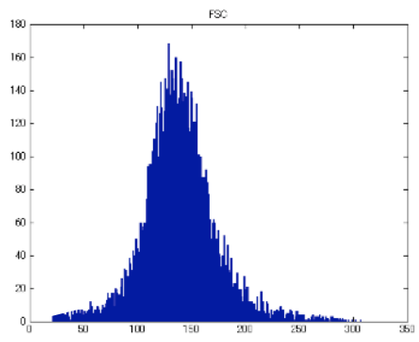
Microscopy

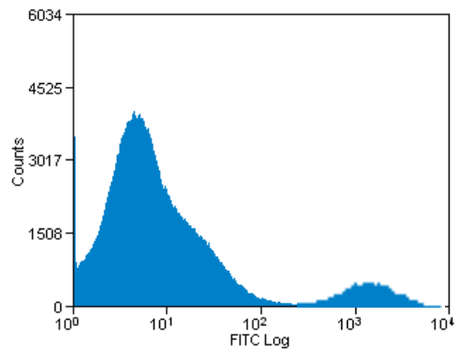


Do Individual Differences within a Population Matter?

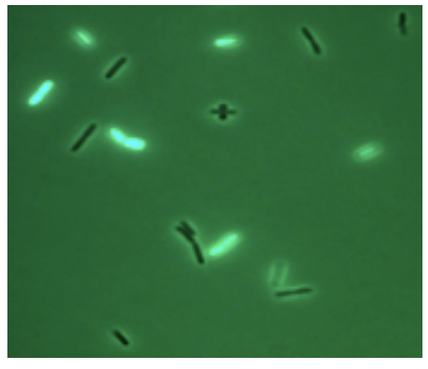
Averages hide important information





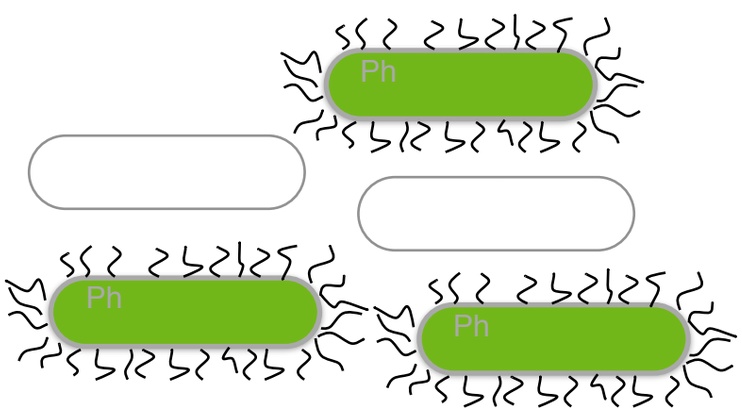


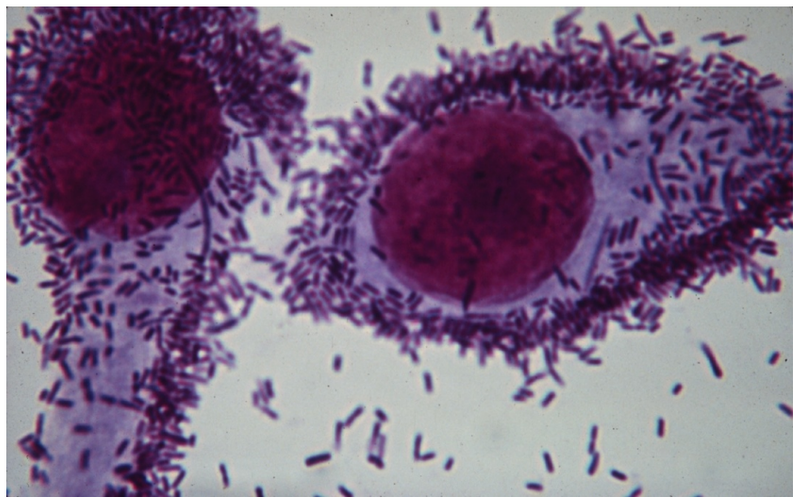
Elowitz et al, Science 2002



Biological Influences of Random Gene Expression

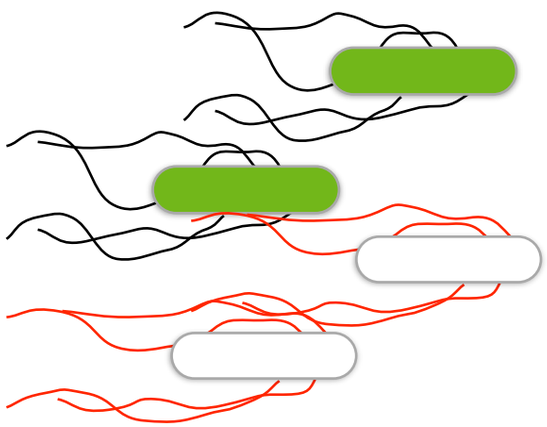
E. coli

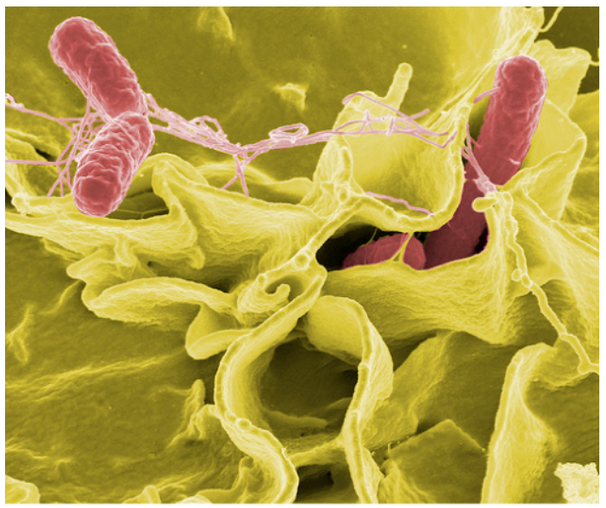




Kaper et al., Nature Rev. Microbiol. (2004)

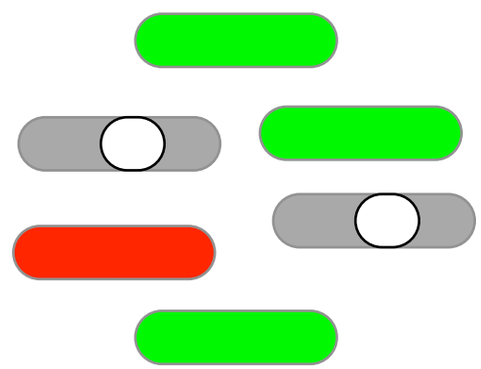
Salmonella

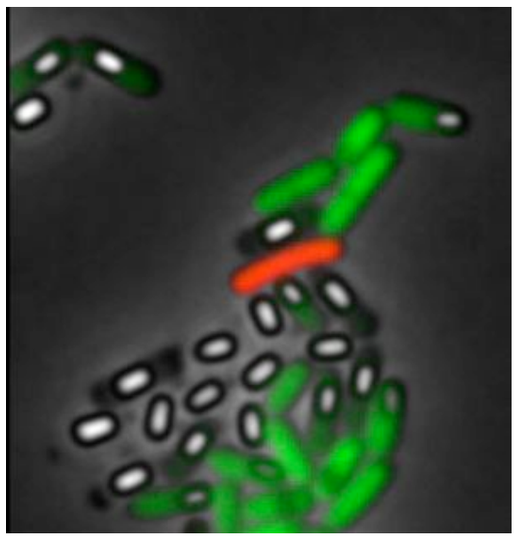




Credit: Rocky Mountain Laboratories

Bacillus subtilis





Credit: Michael Ellowitz

Biological Influences of Random Gene Expression

Fingerprints of identical twins

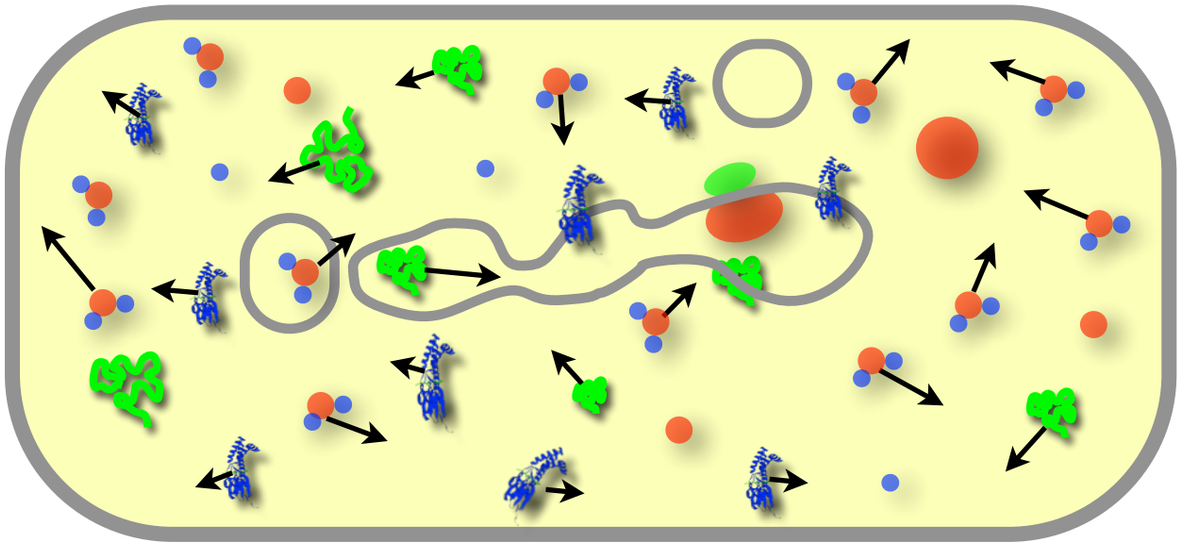
Cc, the first cloned cat and Rainbow, her genetic mother

J. Raser and E. O'Shea, Science, 2005

Origin of Randomness in Gene Expression

The Picture inside a Cell

- Reactants are **discrete** in nature; some are **scarce**
- Chemical reactions are **random**

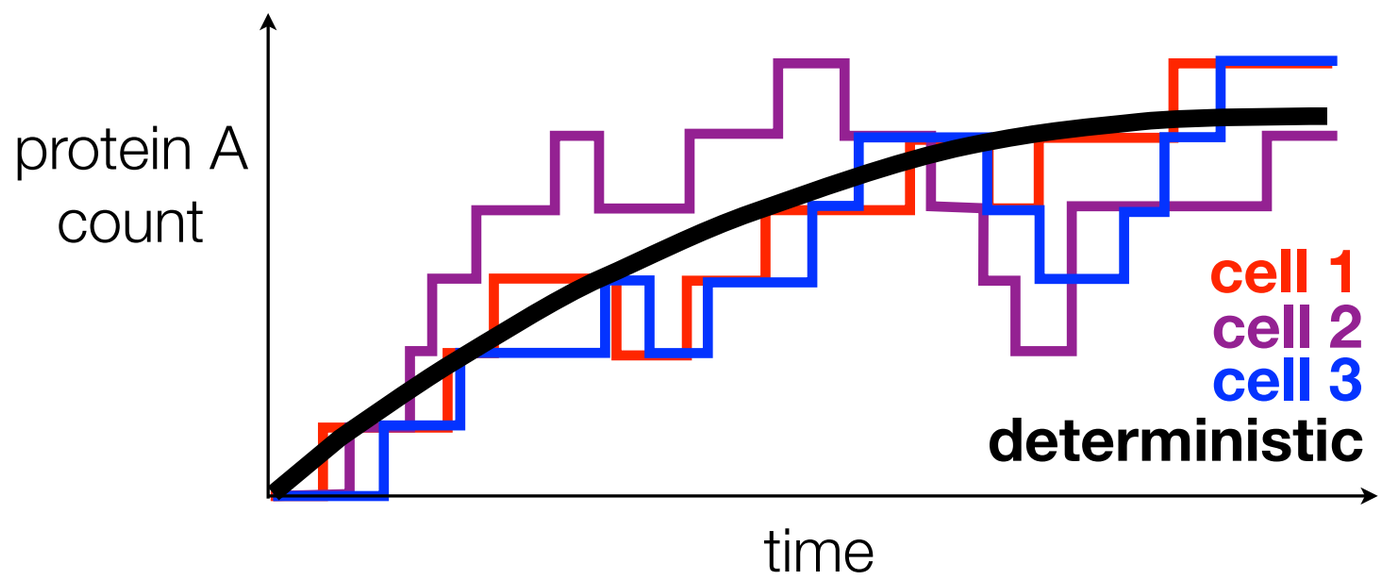


Biological Consequences

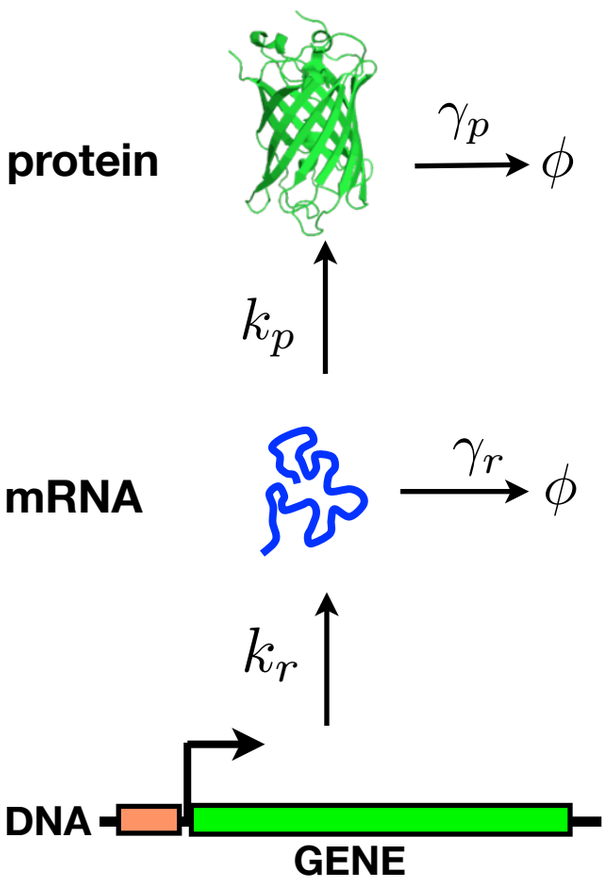
- Random fluctuations in a cell
- Cell-cell variability

Modeling Consequences

- A probabilistic approach is needed



Modeling Gene Expression



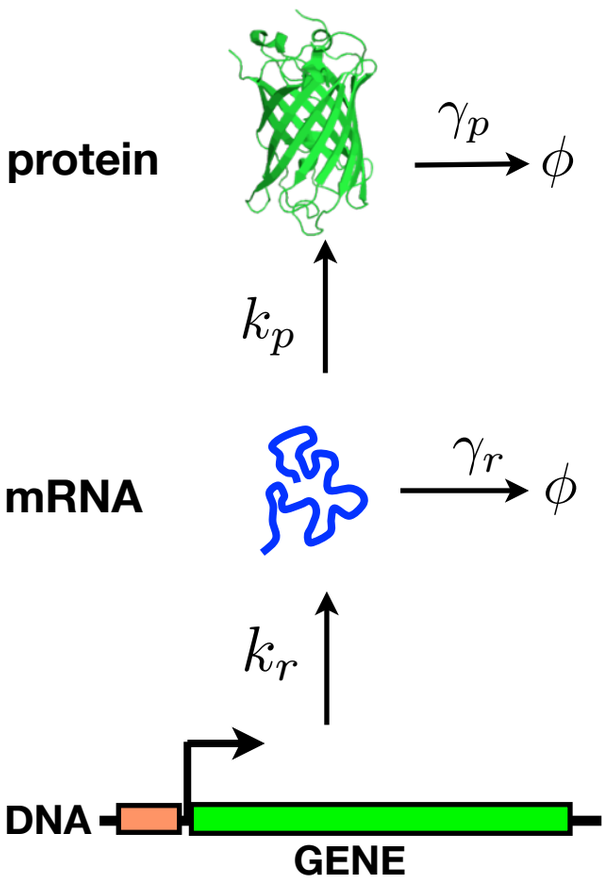
Stochastic model

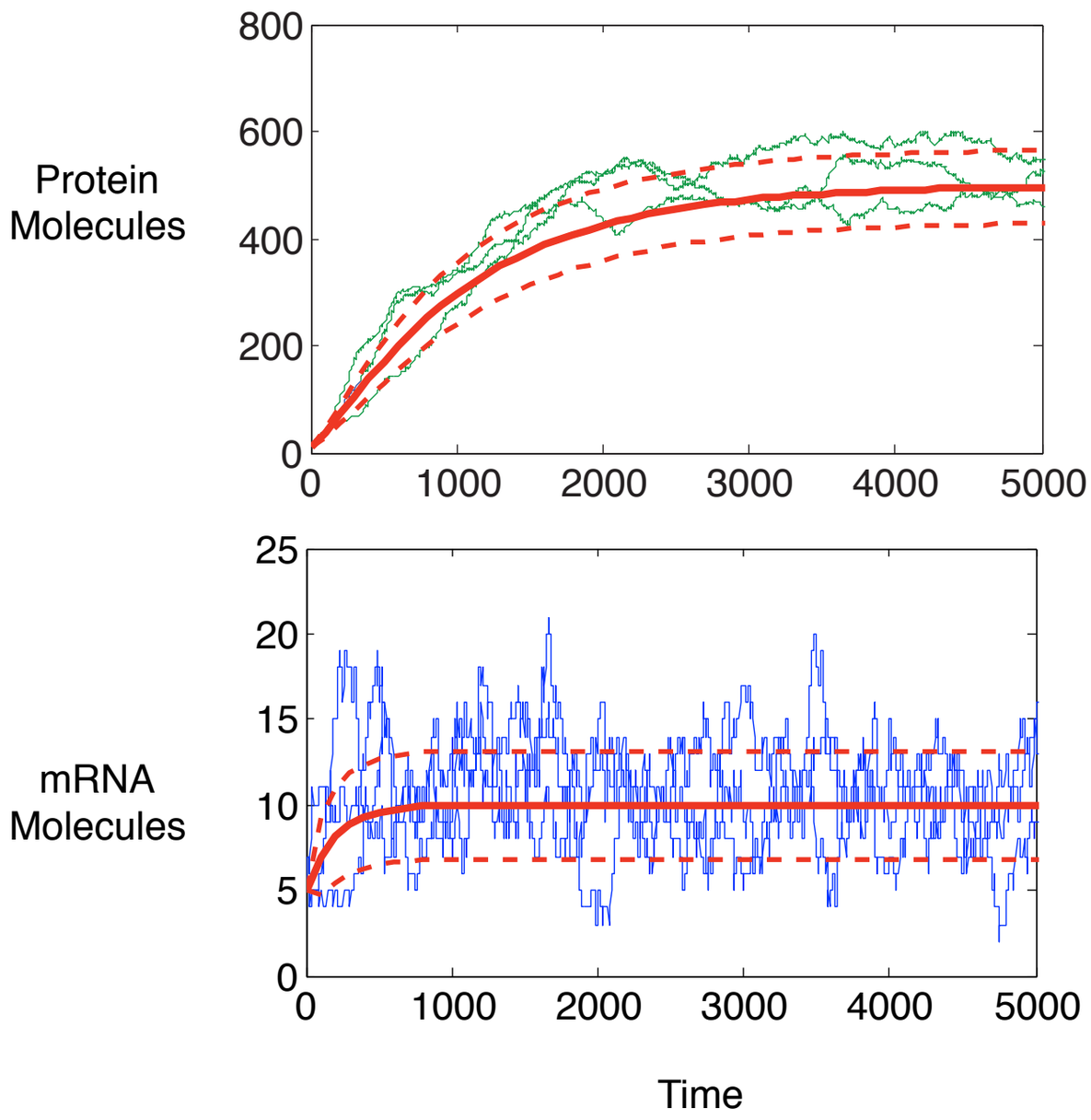
- The number of mRNAs and proteins in a cell are discrete random variables: $X_r(t)$ and $X_p(t)$
- The probability that a single mRNA is transrcibed in time h is $k_r h + o(h)$
- The probability that a single mRNA is degraded in time h is $X_r(t)\gamma_r h + o(h)$

$o(h)$ notation: $\frac{o(h)}{h} \to 0$ as $h \to 0$

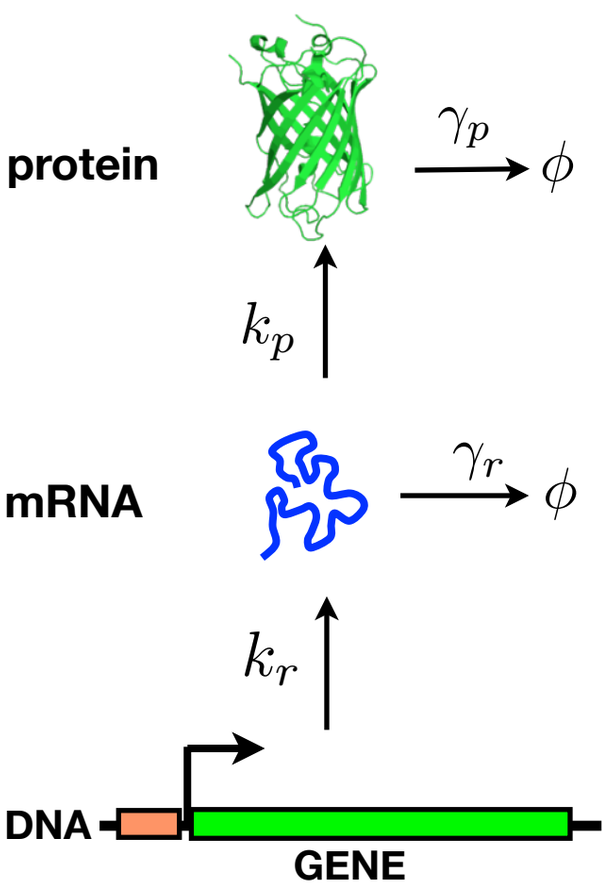
$$X(t) = \begin{bmatrix} X_r(t) \\ X_p(t) \end{bmatrix} \text{ is a } \textbf{continuous-time discrete-state Markov process}$$

Modeling Gene Expression





Modeling Gene Expression



At stationarity

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

$$C_v(p) = \frac{1}{\sqrt{\mathbb{E}(p)}}\left(1 + \frac{k_p}{\gamma_p + \gamma_r}\right)^{1/2}$$

(protein)

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

$$C_v(r) = \frac{1}{\sqrt{\mathbb{E}(r)}}$$

(mRNA)

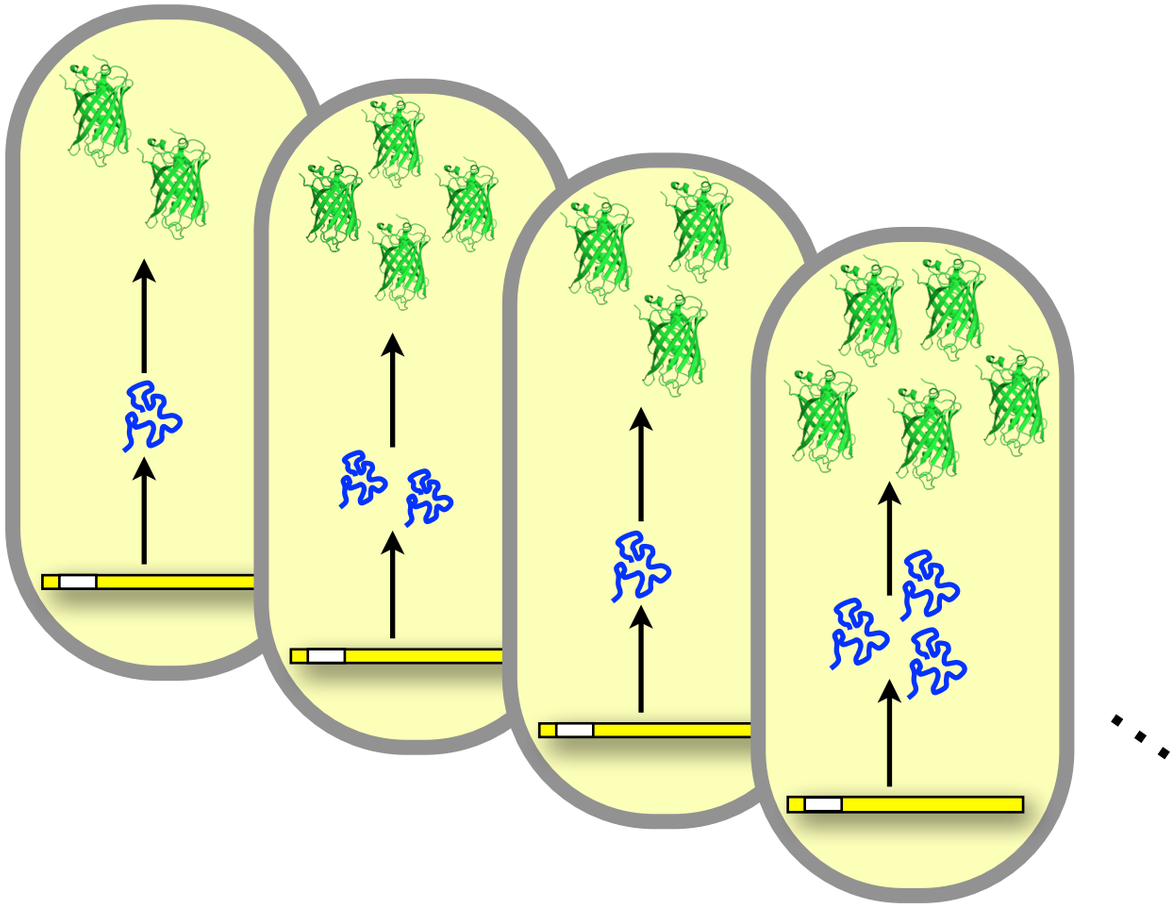
C_v = coefficient of variation = $\frac{\text{standard deviation}}{\text{mean}}$

Mean Dynamics

$$\frac{d}{dt}\mathbb{E}(X_r) = k_r - \gamma_r \mathbb{E}(X_r)$$

$$\frac{d}{dt}\mathbb{E}(X_p) = k_p \mathbb{E}(X_r) - \gamma_p \mathbb{E}(X_p)$$

Model Allows Heterogeneity in Genetically Identical Cells



Questions we can ask:

What is the probability of finding N mRNA molecules in a give cell at time t?

What is the stationary mean and variance of the protein in a population?

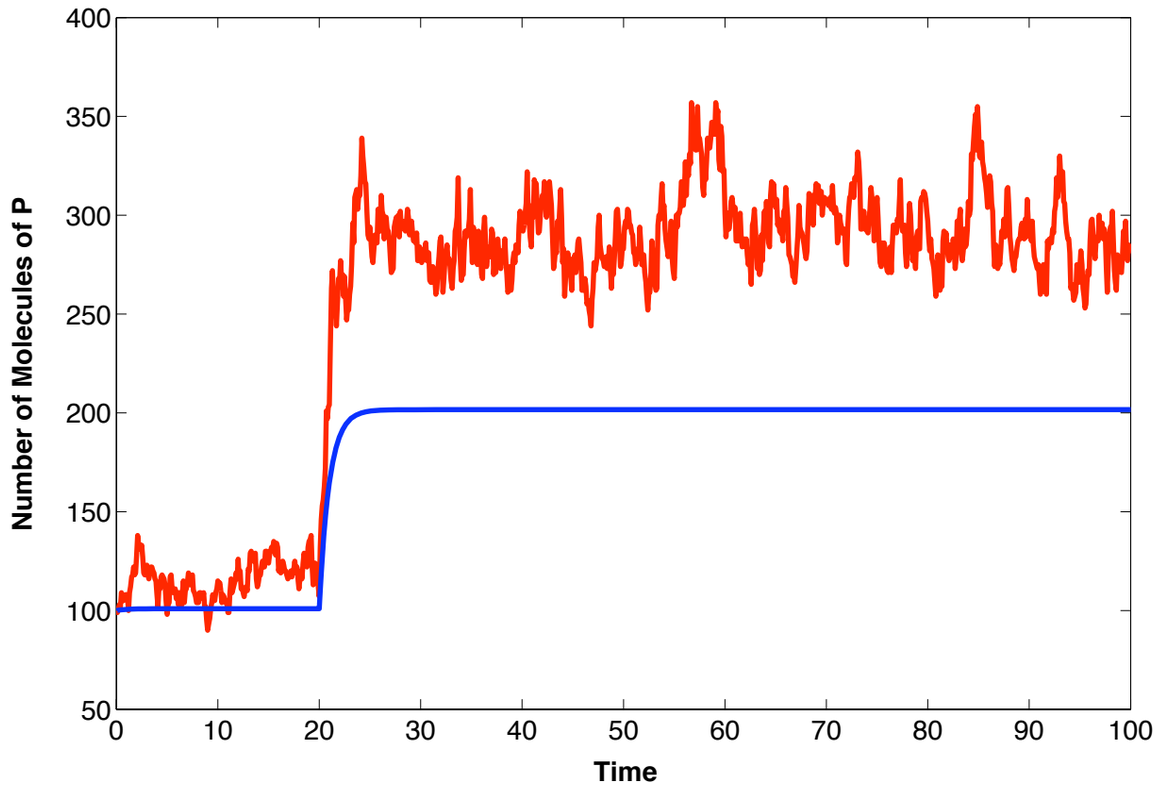
Given measurements of the joint distribution of protein and mRNA at times $T_1,\ldots,T_n$, can we infer the gene expression parameters?

⋮

Deterministic Model Fails to Capture Mean

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

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

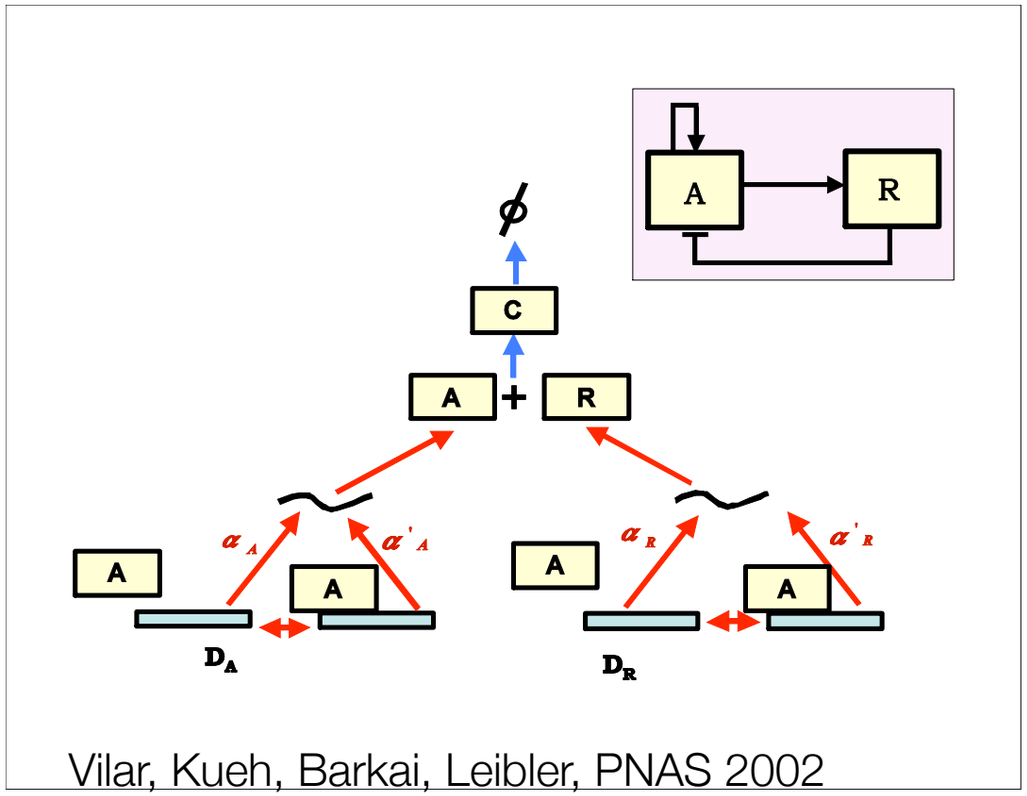


Johan Paulsson , Otto G. Berg , and Måns Ehrenberg, PNAS 2000

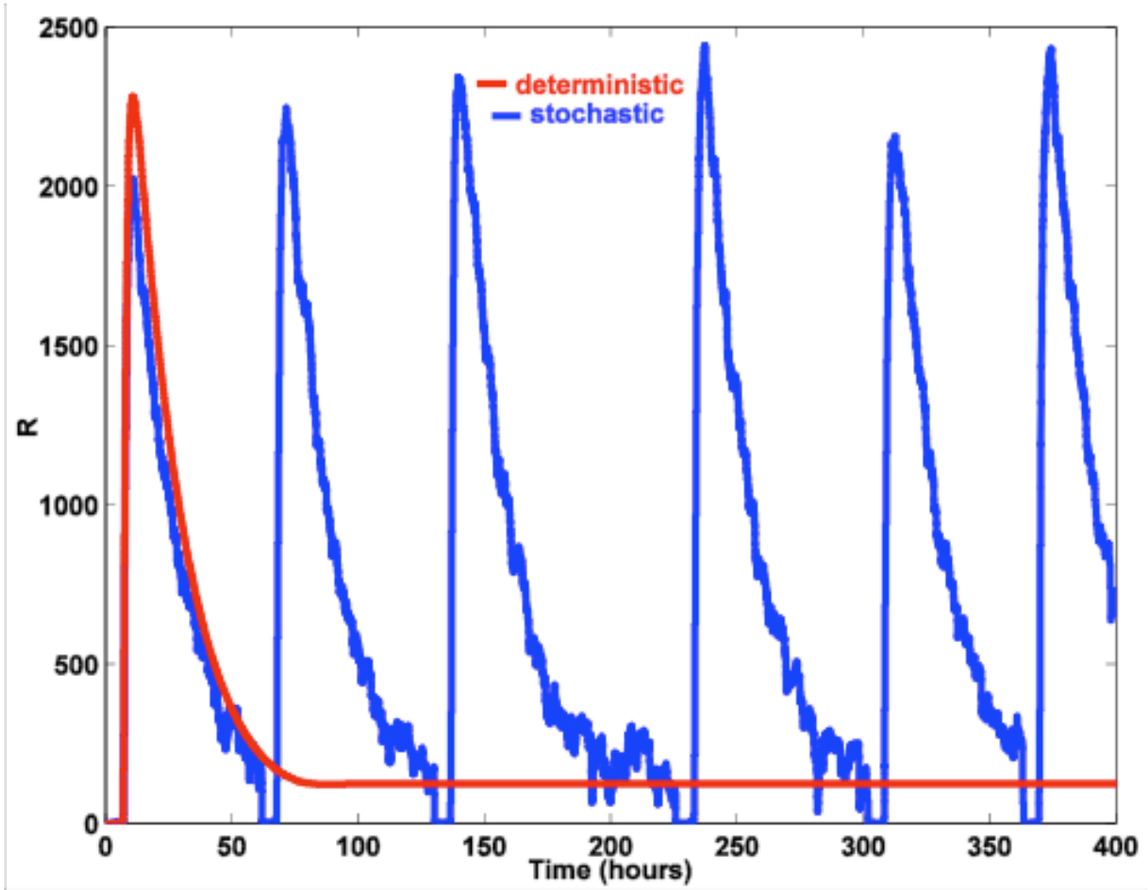
- Stochastic mean value different from deterministic steady state
- Noise *enhances* signal!

Noise Induced Oscillations

Circadian rhythm

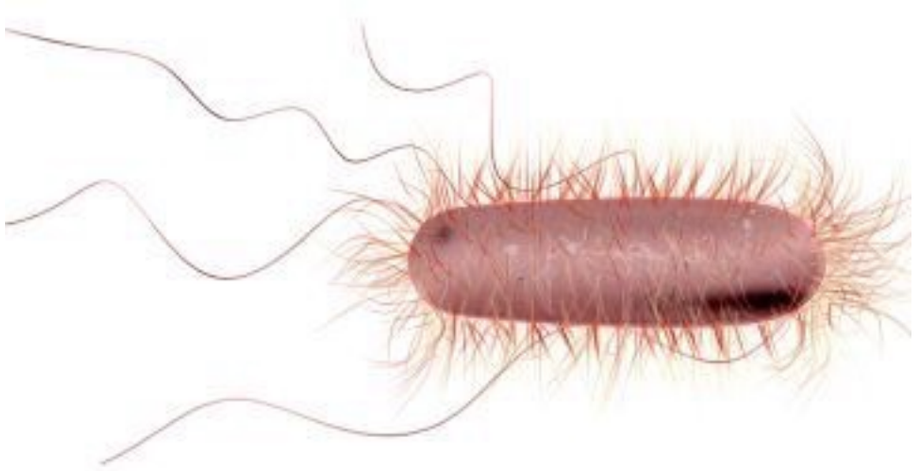


Vilar, Kueh, Barkai, Leibler, PNAS 2002

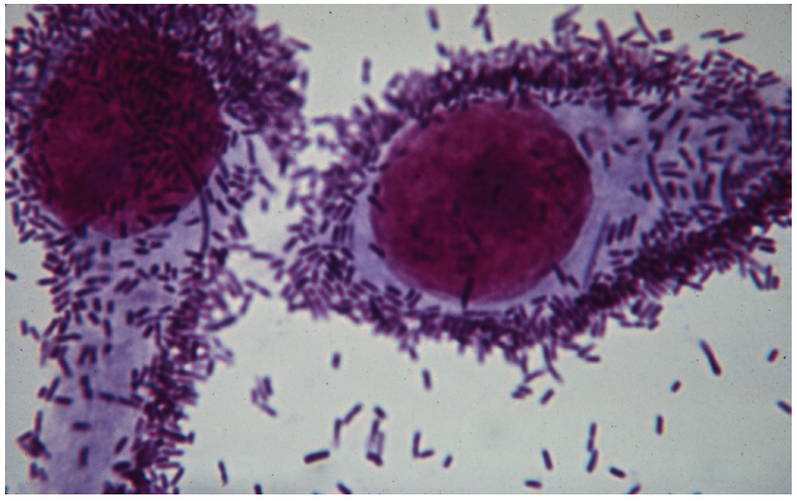


- Oscillations disappear from deterministic model after a small reduction in deg. of repressor
- (Coherence resonance) Regularity of noise induced oscillations can be manipulated by tuning the level of noise [*El-Samad, Khammash*]

The Pap Pili Stochastic Switch

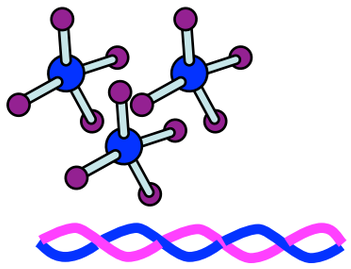


Uropathogenic
E. coli



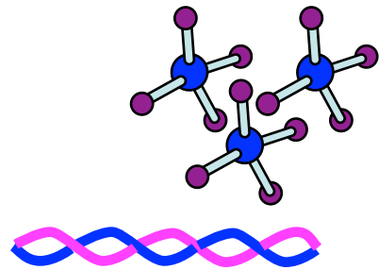
- Pili enable uropathogenic E. coli to attach to epithelial cell receptors
  - Plays an essential role in the pathogenesis of urinary tract infections
- E. coli expresses two states ON (piliated) or OFF (unpiliated)
- Piliation is controlled by a stochastic switch that involves random molecular events

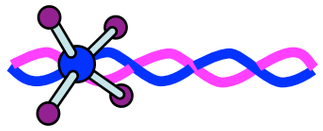
Stochastic Switching: Identical Genotype Produces Different Phenotype



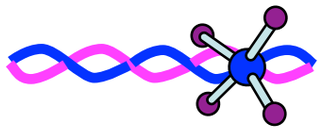
Same chemical environment.

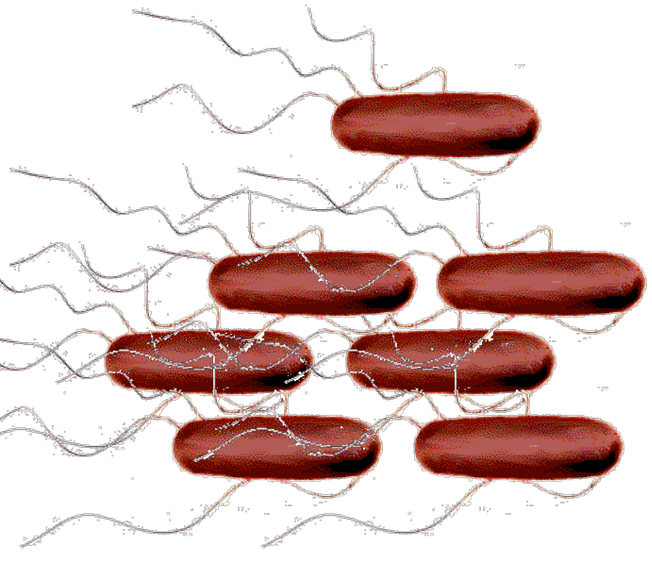
Same genetic code.



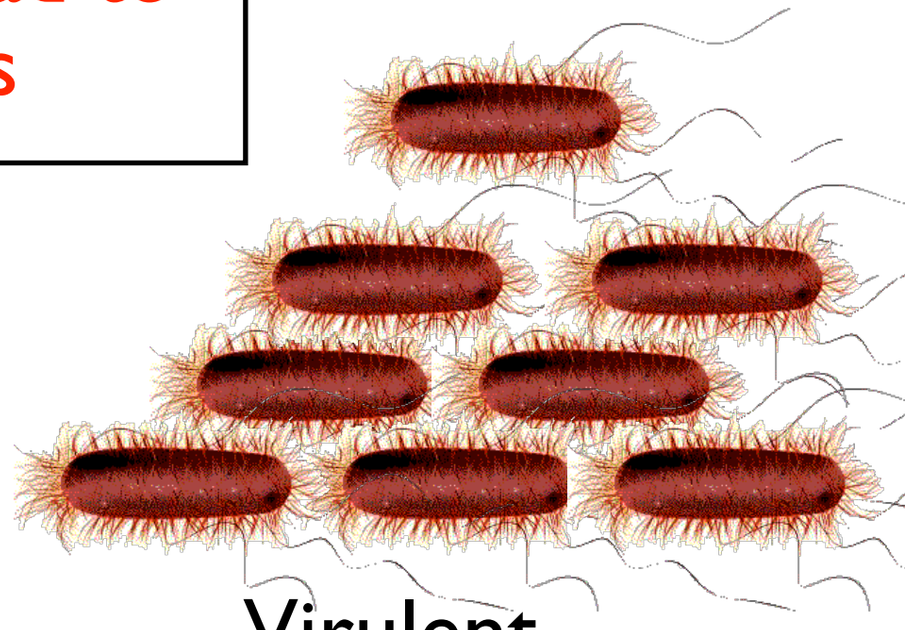


Random Reactions can lead to vastly different results





Harmless phenotype.

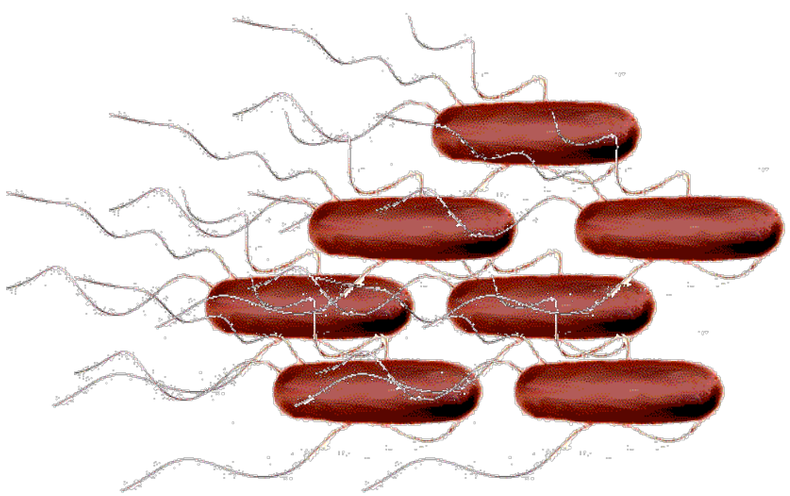


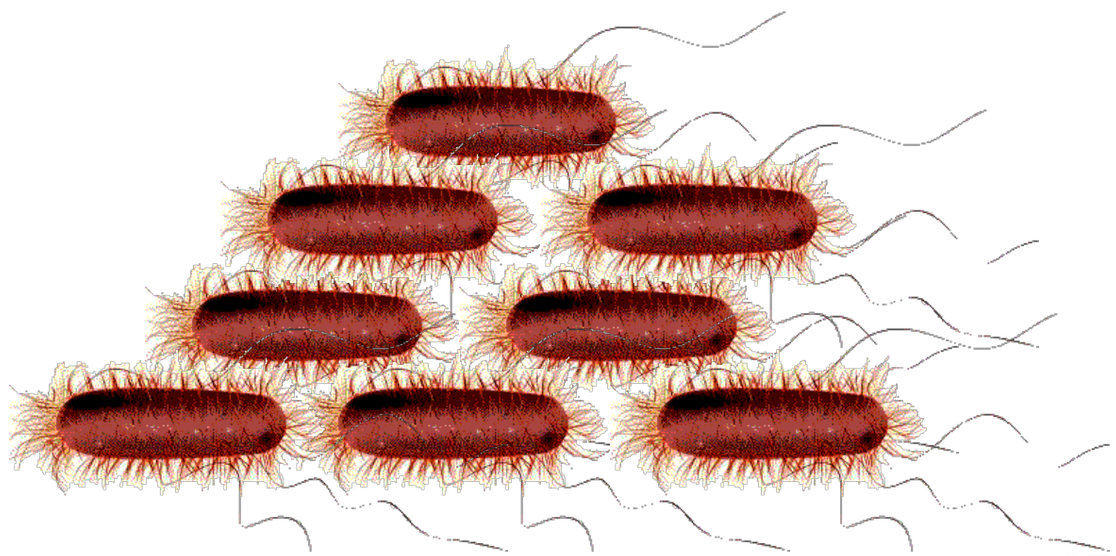
Virulent phenotype.

Stochastic Switching: Identical Genotype Produces Different Phenotype

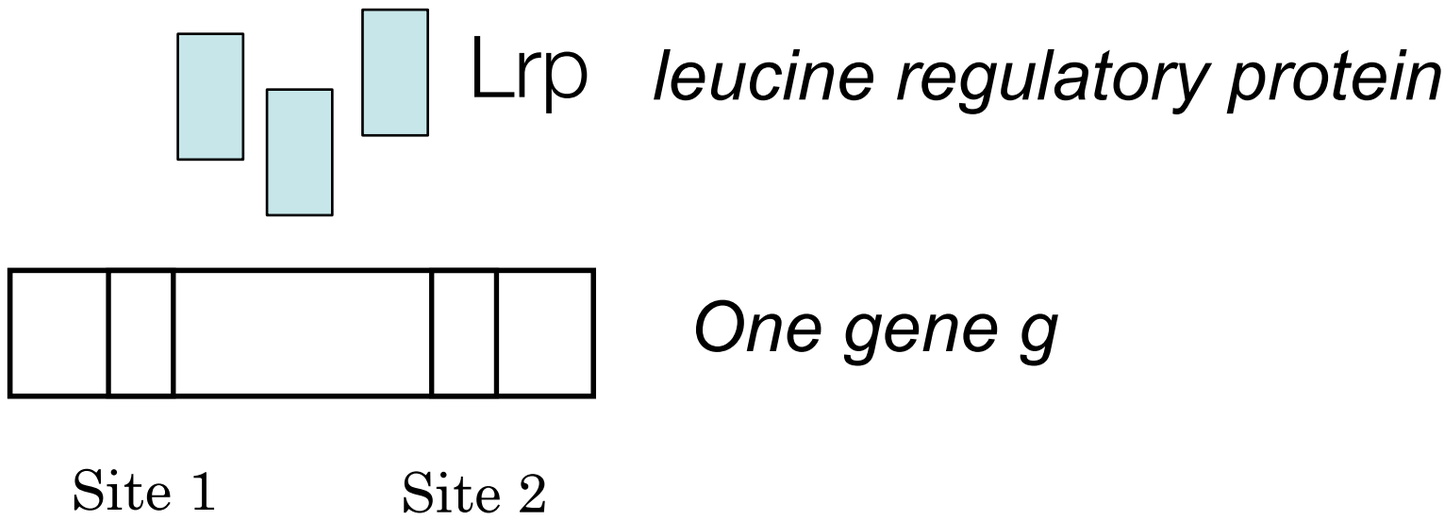
For these systems, we need analytical models to answer:

- What will happen?
- How frequently?
- Why does it happen?
- Under what conditions?
- What advantages does it provide?
- How can we prevent it?
- How can we cause it?

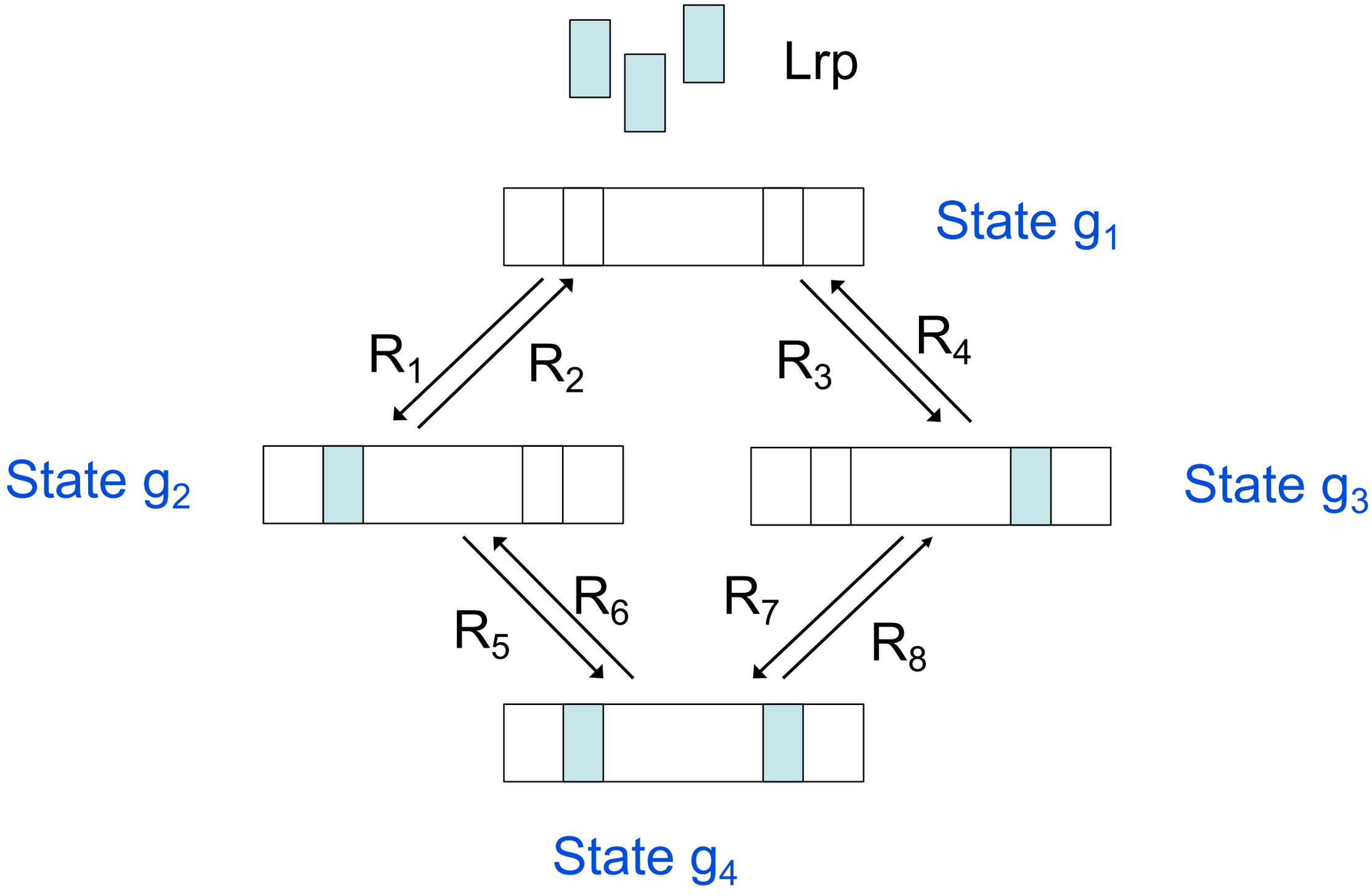


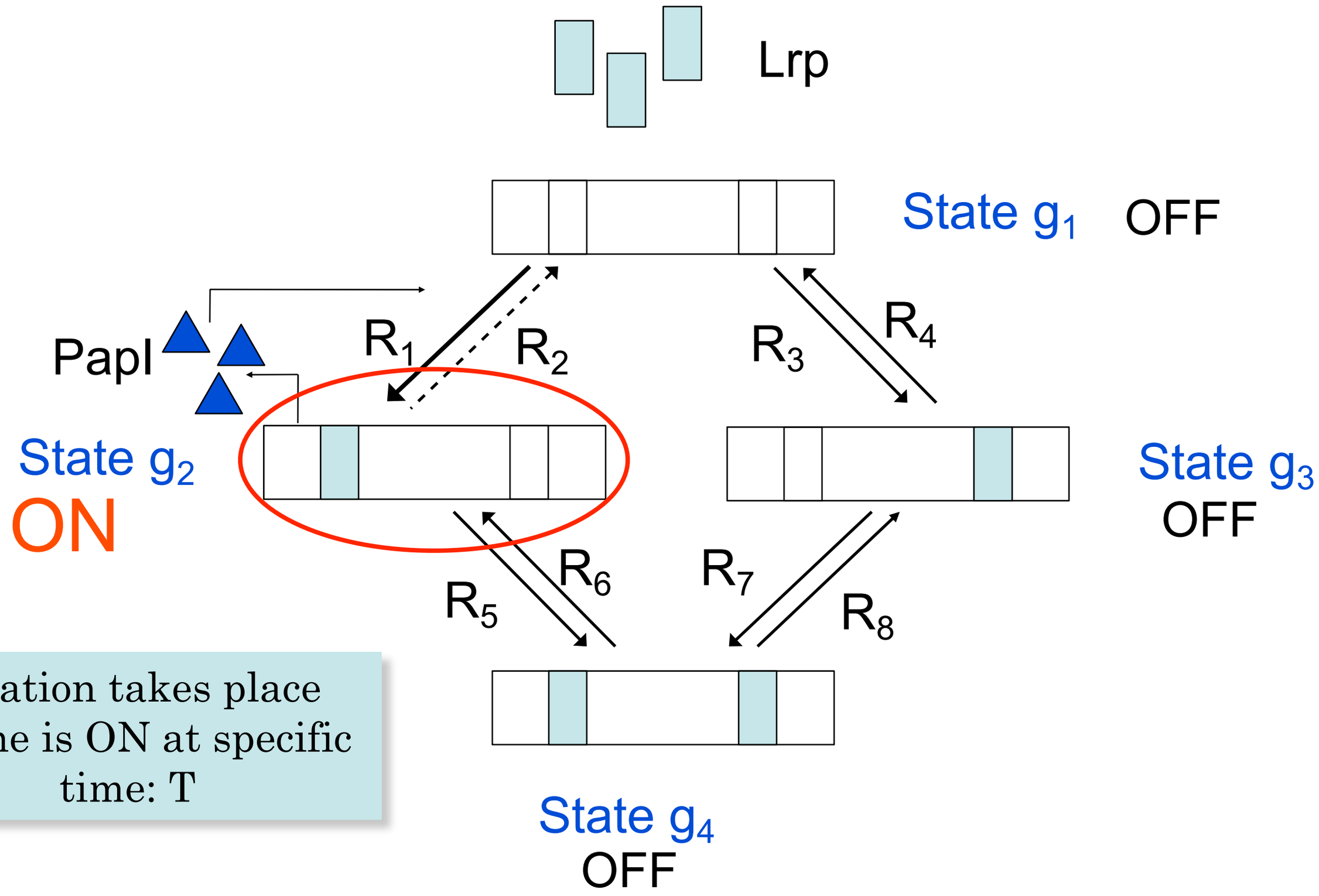


A Simplified Pap Switch Model



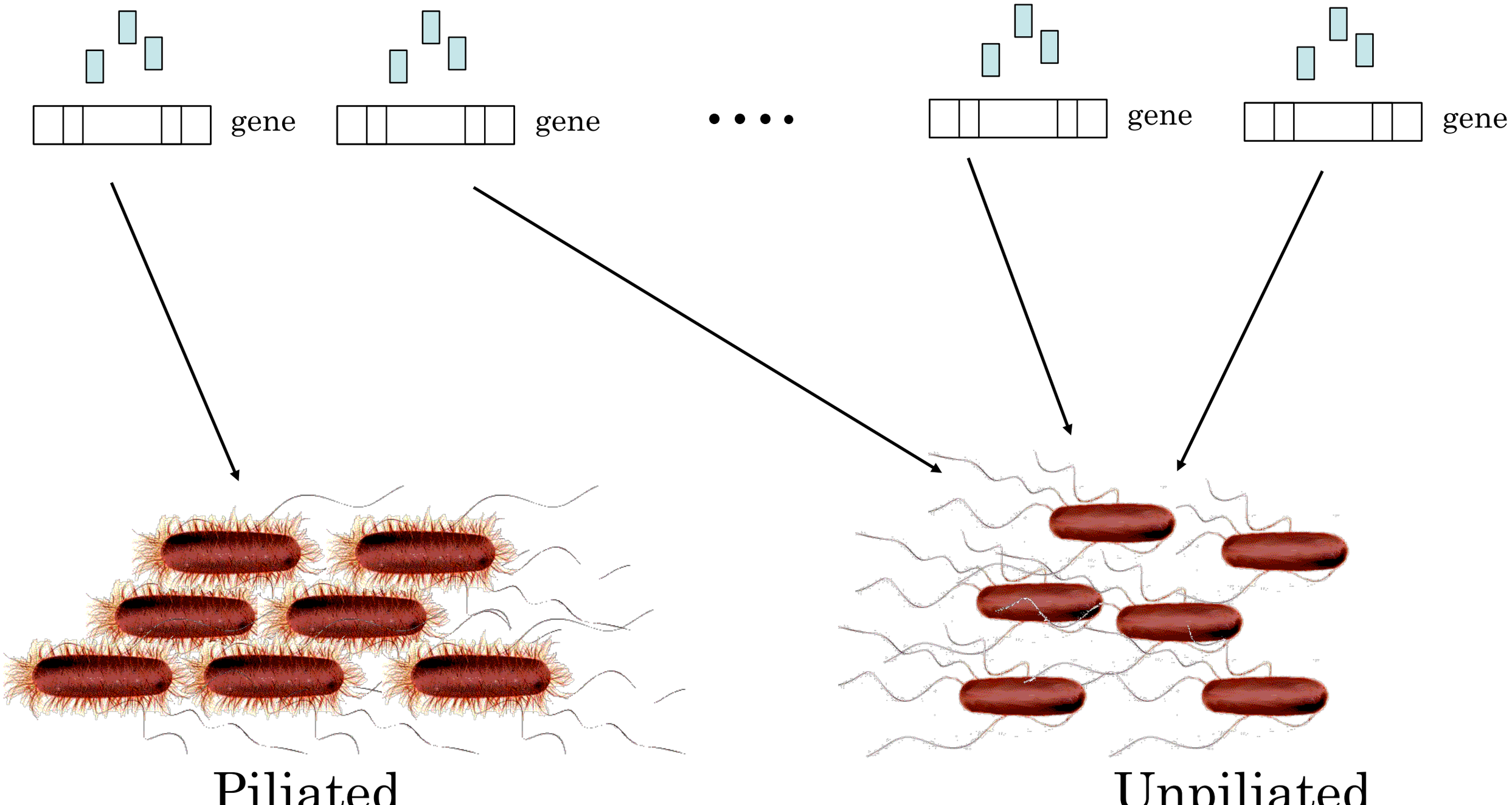
- Lrp can (un)bind either or both of two binding sites
- A (un)binding reaction is a random event


Lrp
State g_1
R_1
R_2
R_3
R_4
State g_2
State g_3
R_5
R_6
R_7
R_8
State g_4



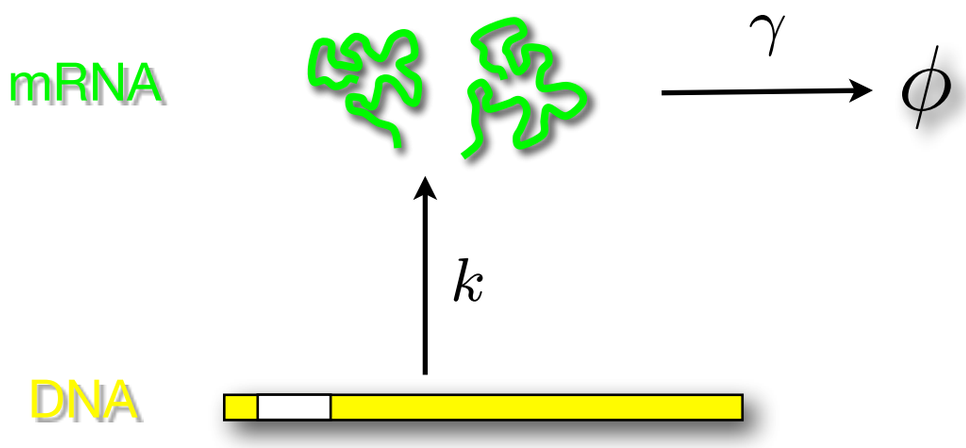
Piliation takes place if gene is ON at specific time: T

Identical Genotype Leads to Different Phenotype



An Introduction to Stochastic Modeling:
Gene Transcription

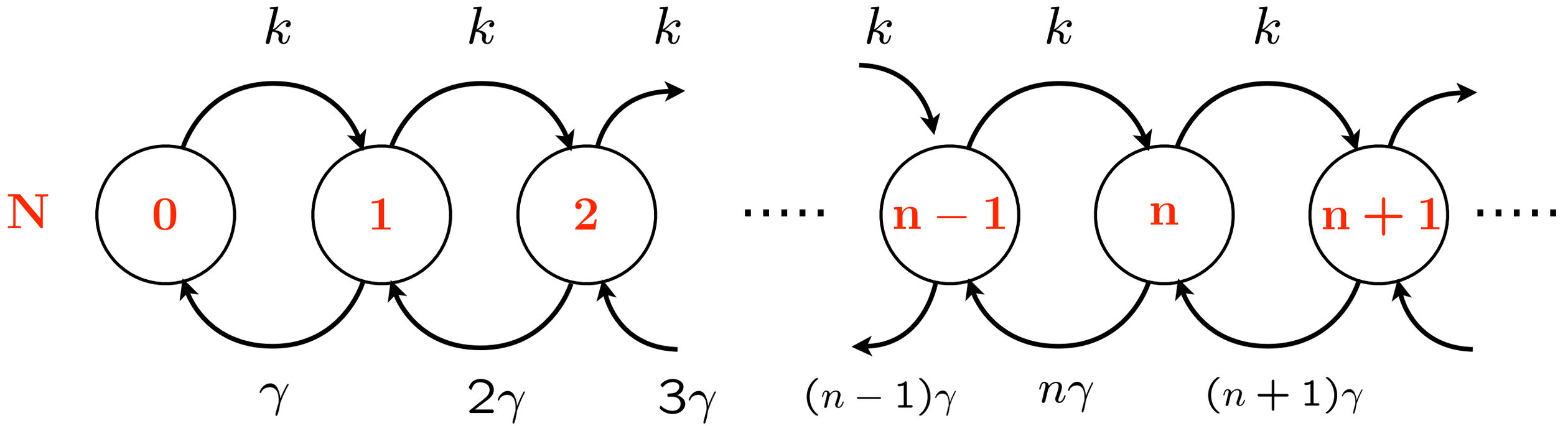
A Simple Example



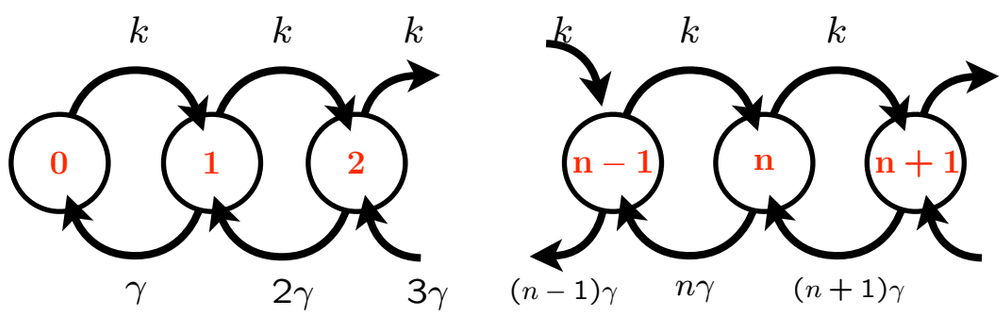
mRNA copy number $N(t)$ is a random variable

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 $\sum_{n=0}^{\infty} p(n) = 1$

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

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

Poisson Distribution

We can compute the mean and variance of the Poisson RV $\bar{N}$ with density $p(n) = e^{-a}\frac{a^n}{n!}$:

$$\mu = E[\bar{N}] = \sum_{n=0}^{\infty} np(n) = e^{-a} \sum_{n=0}^{\infty} n\frac{a^n}{n!} = a$$

The second moment

$$E[\bar{N}^2] = \sum_{n=0}^{\infty} n^2 p(n) = a^2 + a$$

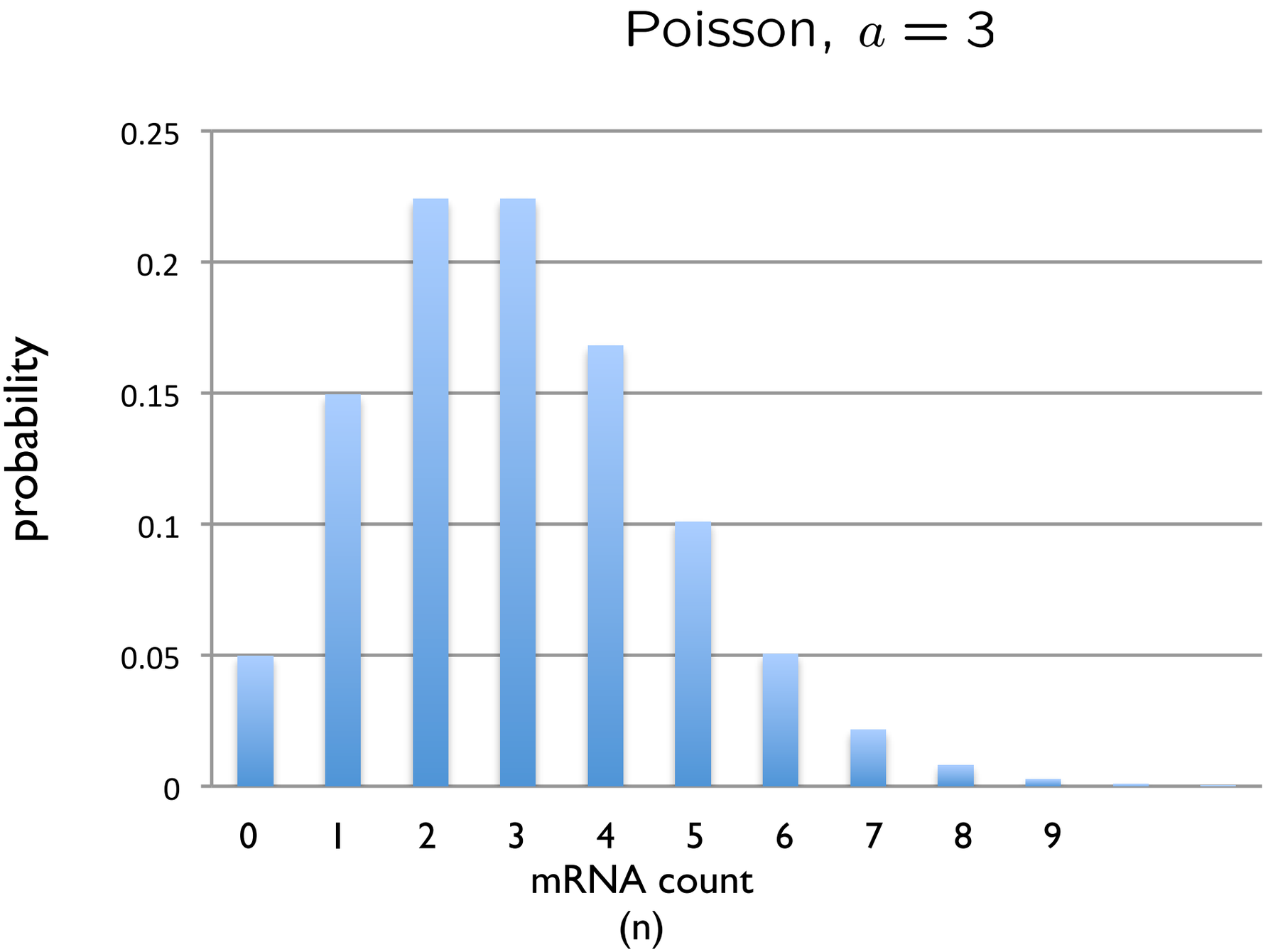
Therefore,

$$\sigma^2 = E[\bar{N}^2] - E[\bar{N}]^2 = a$$

mean = variance = a

The coefficient of variation $C_v = \sigma/\mu$ is

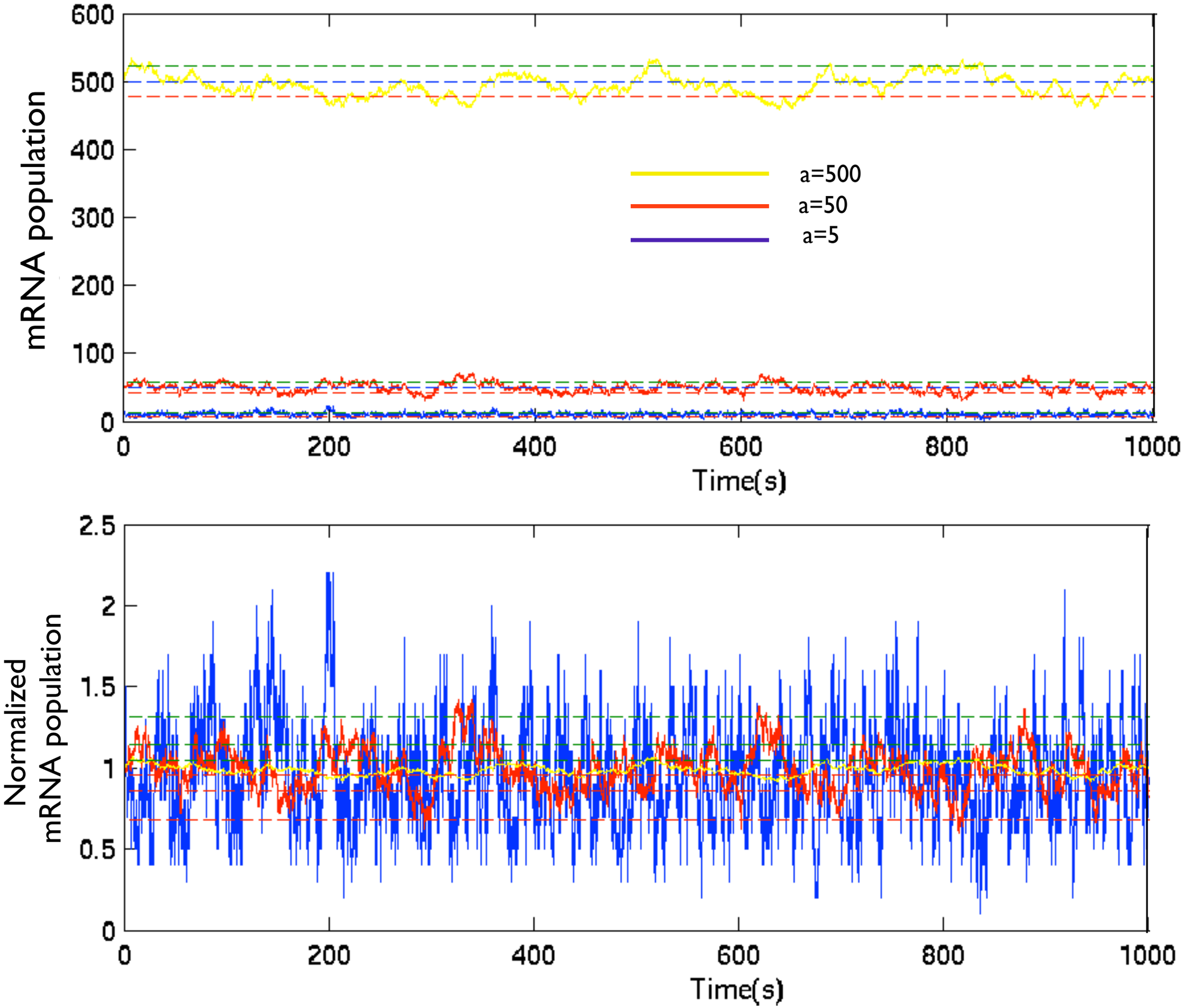
$$C_v = \frac{1}{\sqrt{a}} = \frac{1}{\sqrt{\mu}}$$



Stationary distribution:

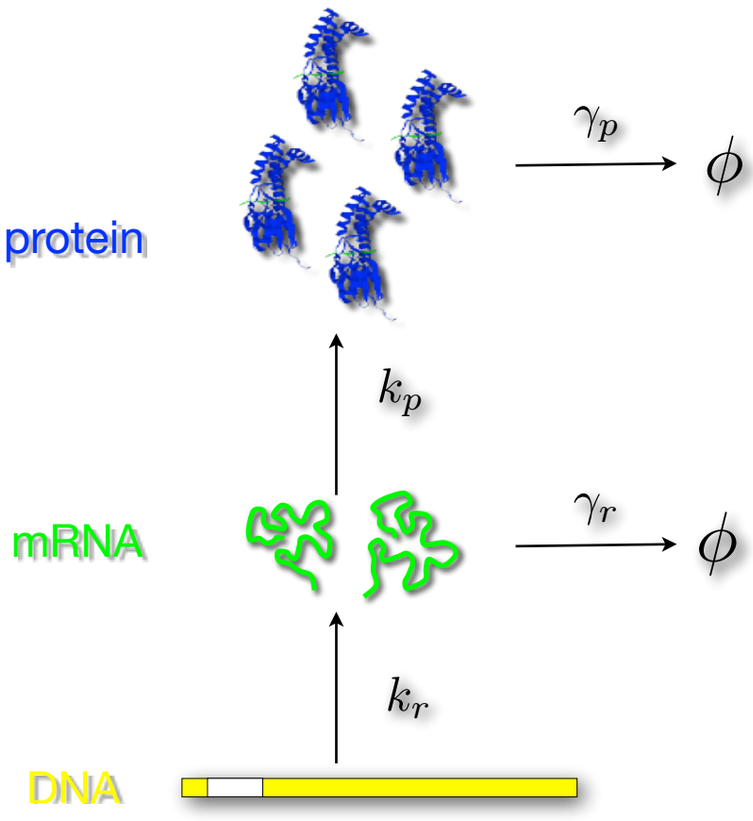
$$P(n) = e^{-a}\frac{a^n}{n!} \qquad a = \frac{k}{\gamma}$$

Poisson Distribution


mRNA population
600
500
400
300
200
100
0
0
200
400
600
800
1000
Time(s)
a=500
a=50
a=5
Normalized
mRNA population
2.5
2
1.5
1
0.5
0
0
200
400
600
800
1000
Time(s)

Using Data for Parameter Inference

Parameter Inference from Population Statistics



Munsky, Trinh, and Khammash (2009)
Nature/EMBO Molecular Systems Biology

Lack of identifiability from average protein measurements

It is *impossible* to identify all model parameters using average proteins measurements: $E[p(t_1)], E[p(t_2)], \ldots$

Protein Variability Measurements Enables Identifiability

If measurements of $E[p]$ and $E[p^2]$ are used, then *identifiability is possible* with five time measurements.

Explicit formulae in the case of mRNA measurements

Given mean and standard deviation at two times instances:

$$(\mu_0, \sigma_0) := (\mu(t_0), \sigma(t_0)) \text{ and } (\mu_1, \sigma_1) := (\mu(t_1), \sigma(t_1))$$

$$\gamma_r = -\frac{1}{2\tau} \log\left(\frac{\sigma_1^2 - \mu_1}{\sigma_0^2 - \mu_0}\right) \quad \text{and} \quad k_r = \gamma_r \frac{\mu_1 - \exp(-\gamma_r \tau)\mu_0}{1 - \exp(-\gamma_r \tau)}.$$

$$(\tau := t_1 - t_0)$$

Identifiability of All Model Parameters

$$\mathbf{v}(t) := \begin{bmatrix} E[r] & E[r^2] & E[p] & E[p^2] & E[pr] \end{bmatrix}^T$$

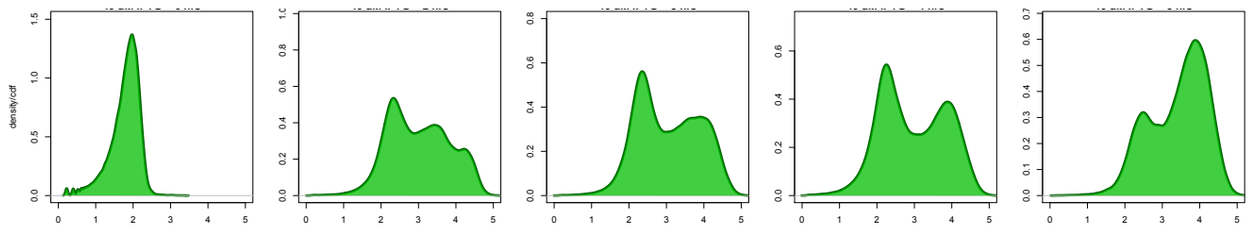
Suppose the vector of moments $\mathbf{v}(t)$ is known at two times $t_0 < t_1 < \infty$.

Then *all four model parameters are identifiable* using only $\mathbf{v}(t_1)$ and $\mathbf{v}(t_2)$.

Using pdf Estimates to Identify Parameters

Using Density Measurements:

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



$\mathbf{P}(t_0)$ $\mathbf{P}(t_1)$ ... $\mathbf{P}(t_{N-1})$

more informative than mean and variance alone

We can use these to identify unknown network parameters λ:

Minimum mismatch:

$$\min_{\lambda} \sum_i |\mathbf{P}_\lambda(t_i) - \mathbf{P}(t_i)| \quad \text{subject to} \quad \underset{\text{(Chemical Master Equation)}}{\dot{\mathbf{P}}_\lambda = A(\lambda)\mathbf{P}_\lambda}$$

Maximum likelihood:

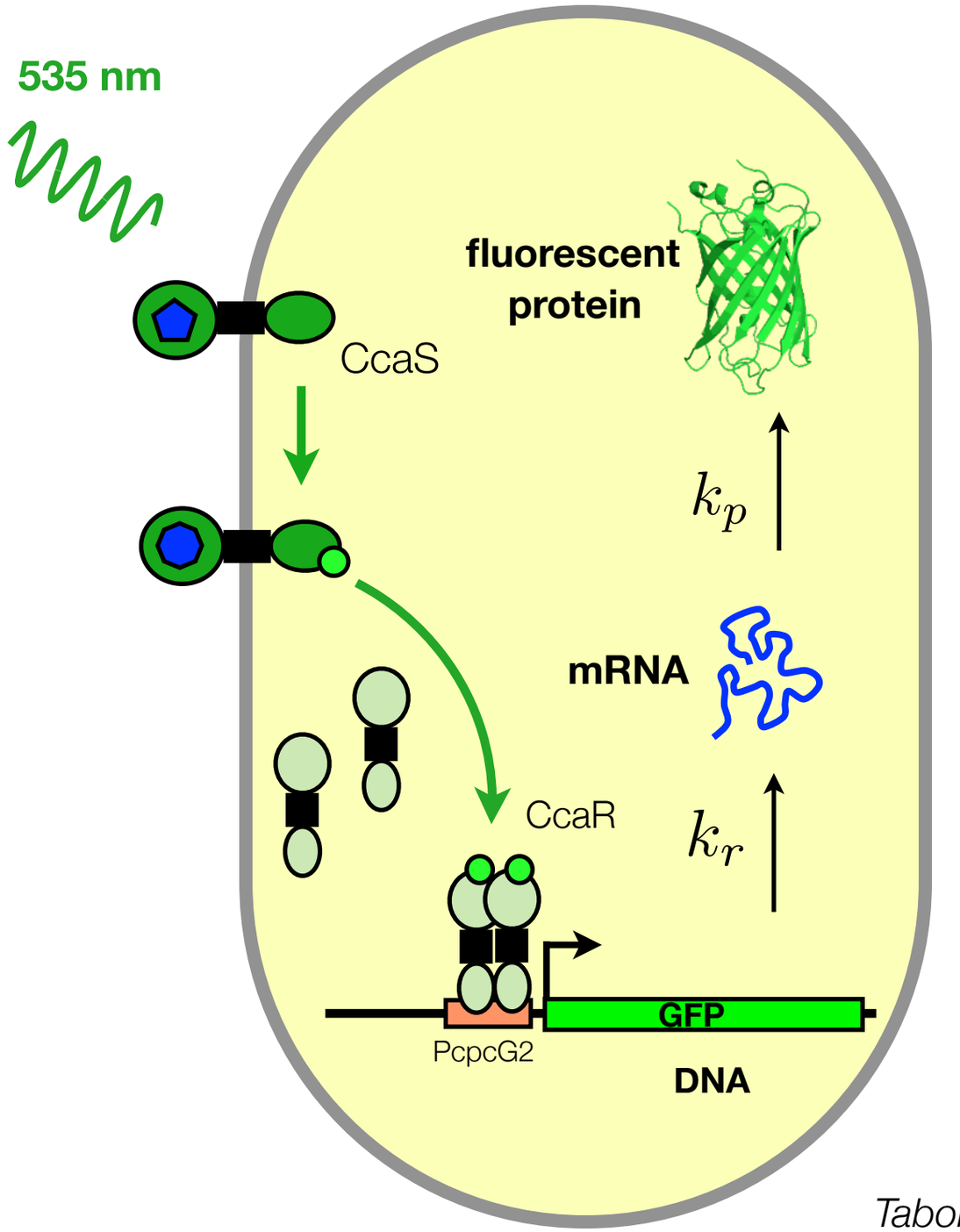
$$\max_{\lambda} \log \mathcal{L}(\lambda|\{\mathbf{P}(t_i)\}) = \max_{\lambda} \sum_i \langle \mathbf{P}(t_i), \log \mathbf{P}_\lambda(t_i) \rangle \quad \text{subject to} \quad \dot{\mathbf{P}}_\lambda = A(\lambda)\mathbf{P}_\lambda$$

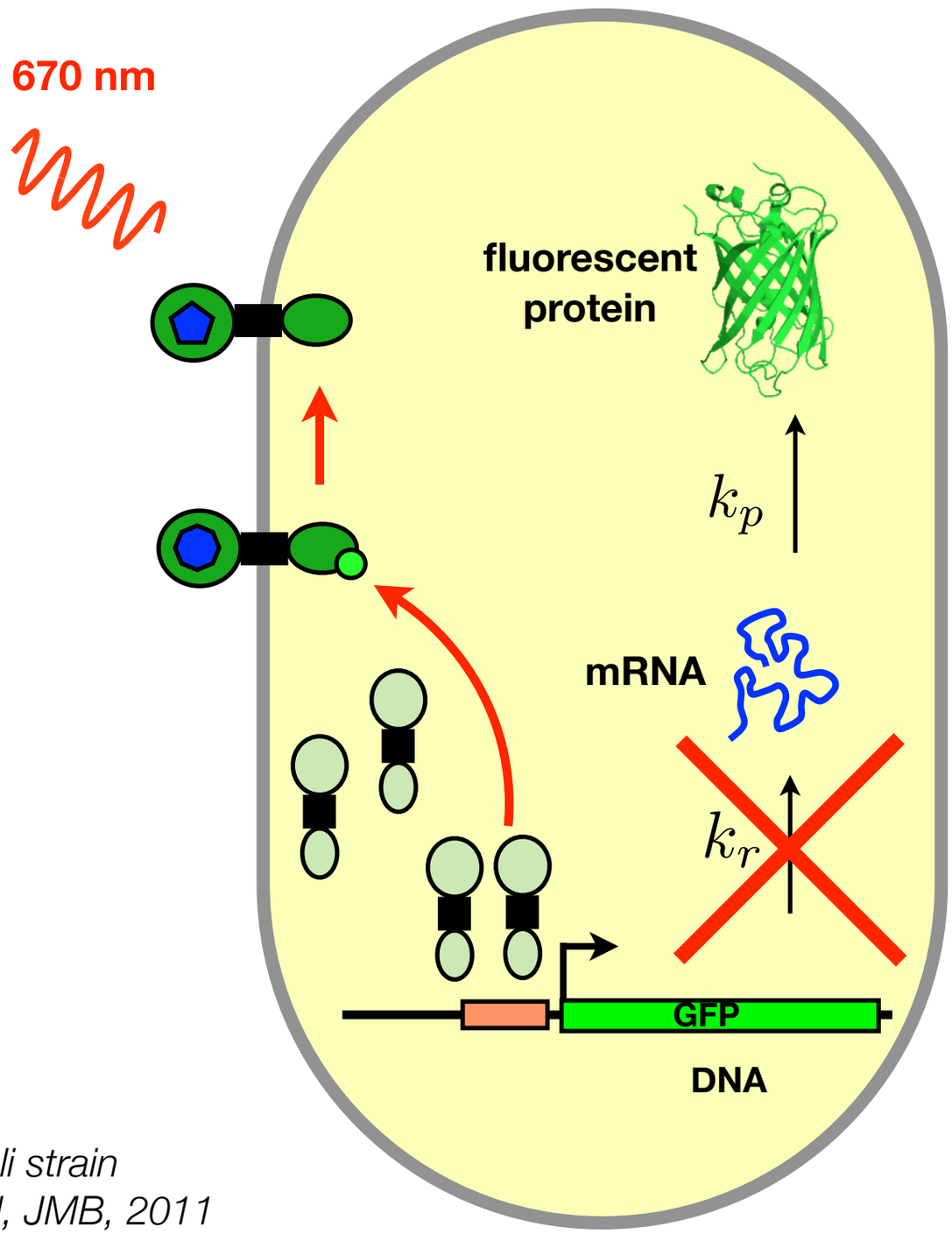
Bayesian:

$$P(\lambda|\{\mathbf{P}(t_i)\}) \propto \mathcal{L}(\lambda|\{\mathbf{P}(t_i)\}) \cdot P(\lambda)$$

Controlling Gene Expression Mean and Variance

Actuation with Light





E. coli strain
Tabor et. al, JMB, 2011

Closed-Loop Optogenetic Control

Turbidity Control

P1

M9 medium

P2

turbidity regulation

waste

Heating and Stirring

Auto Sampling

culture sample

air

V1

V2

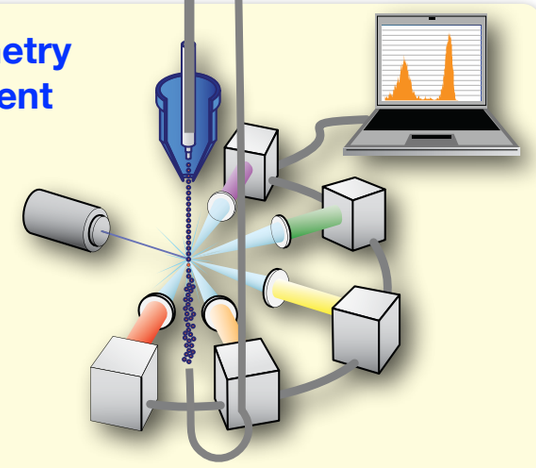
P1

P2

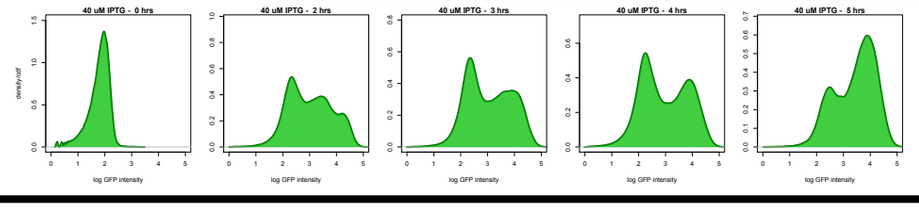
PBS

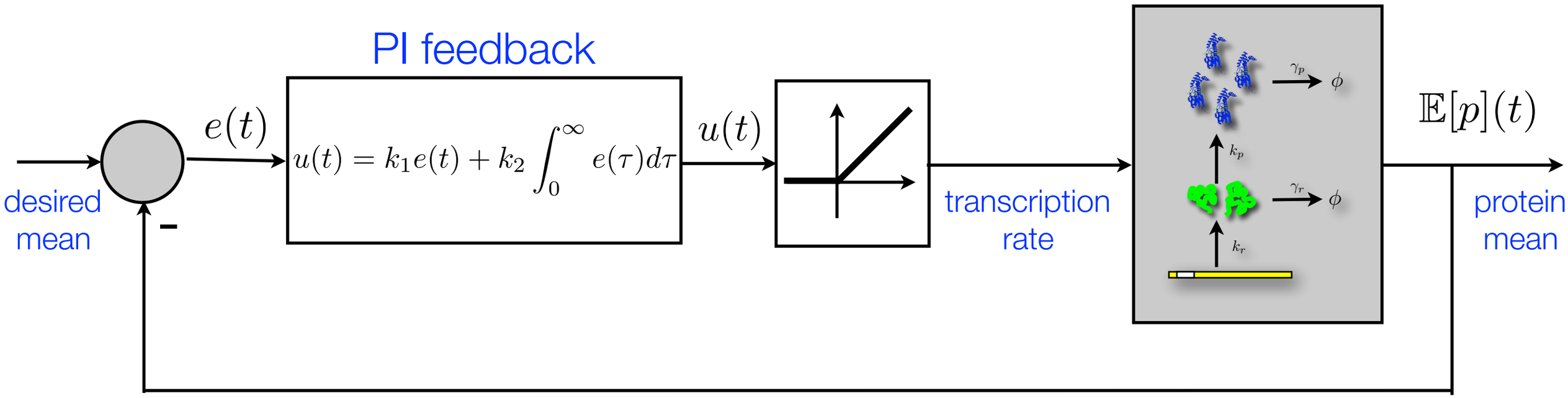
Waste

Flow Cytometry Measurement



setpoint





Controlling protein mean with PI feedback

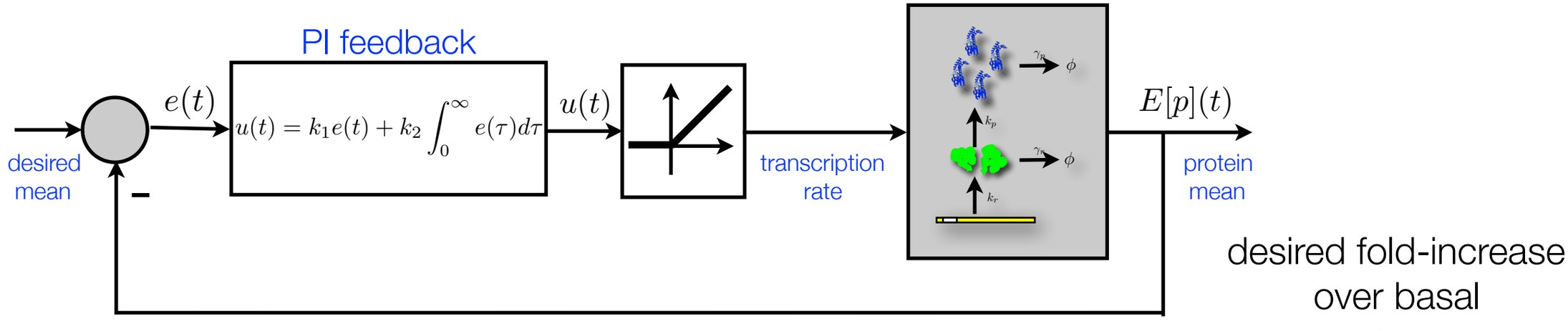
There *always* exists control parameters k_1 and k_2 such that the system is locally stable, and *the protein mean tracks asymptotically the desired mean.*

Local stability and asymptotic tracking are achieved iff

$$k_1 > \frac{k_2}{\gamma_p + \gamma_r} - \frac{\gamma_p \gamma_r}{k_p} \quad \text{and} \qquad k_2 > 0$$

Local stability iff *global* stability

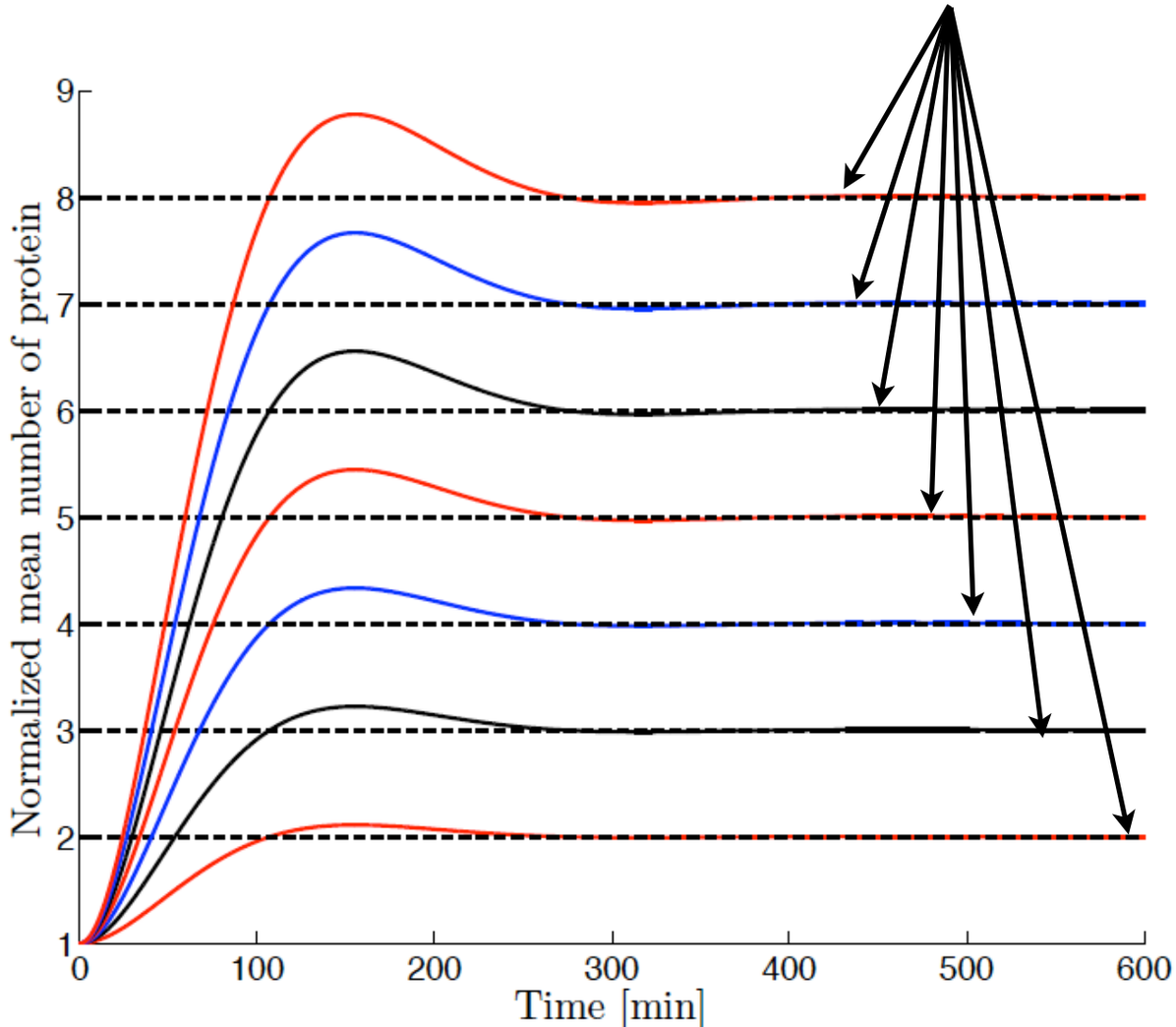
A Simulation Example



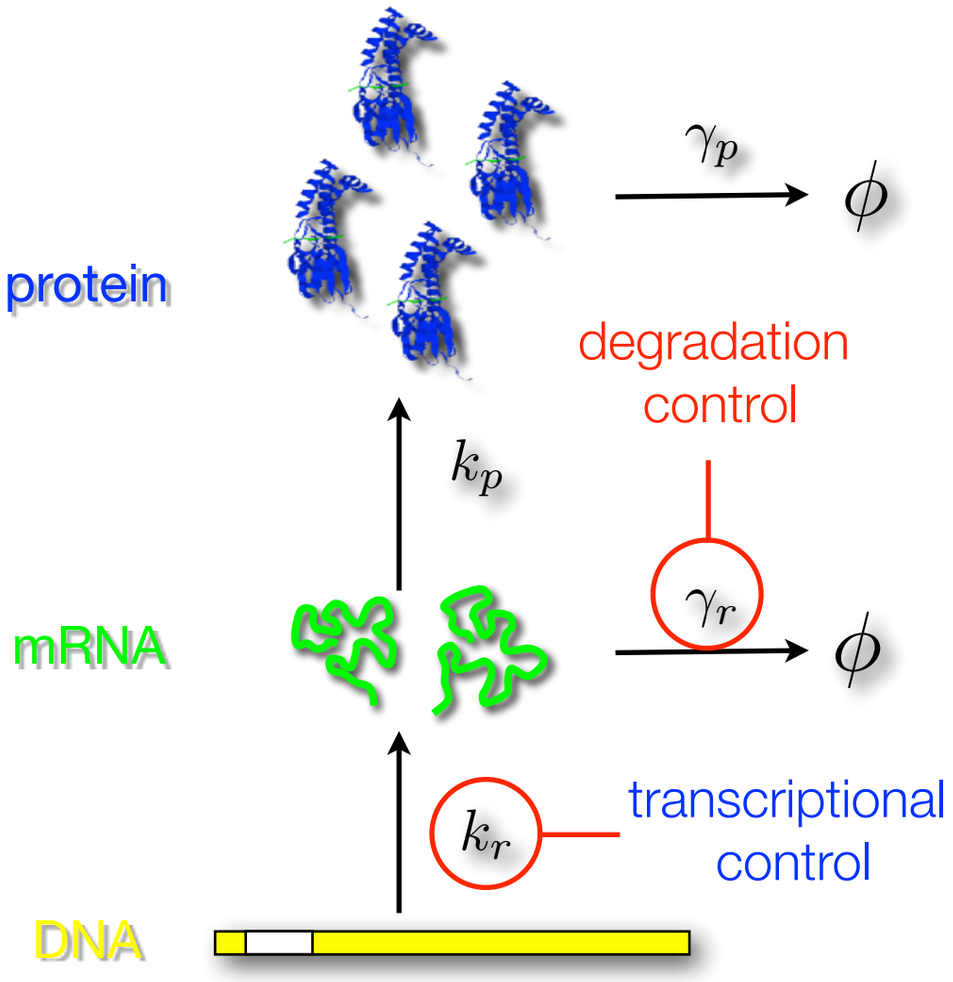
model

$$\frac{d\mathbb{E}[m]}{dt} = -\gamma_m \mathbb{E}[m] + b_m u(t) + r_m$$

$$\frac{d\mathbb{E}[p]}{dt} = k_p \mathbb{E}[m] - \gamma_p \mathbb{E}[p]$$



Mean and Variance Control

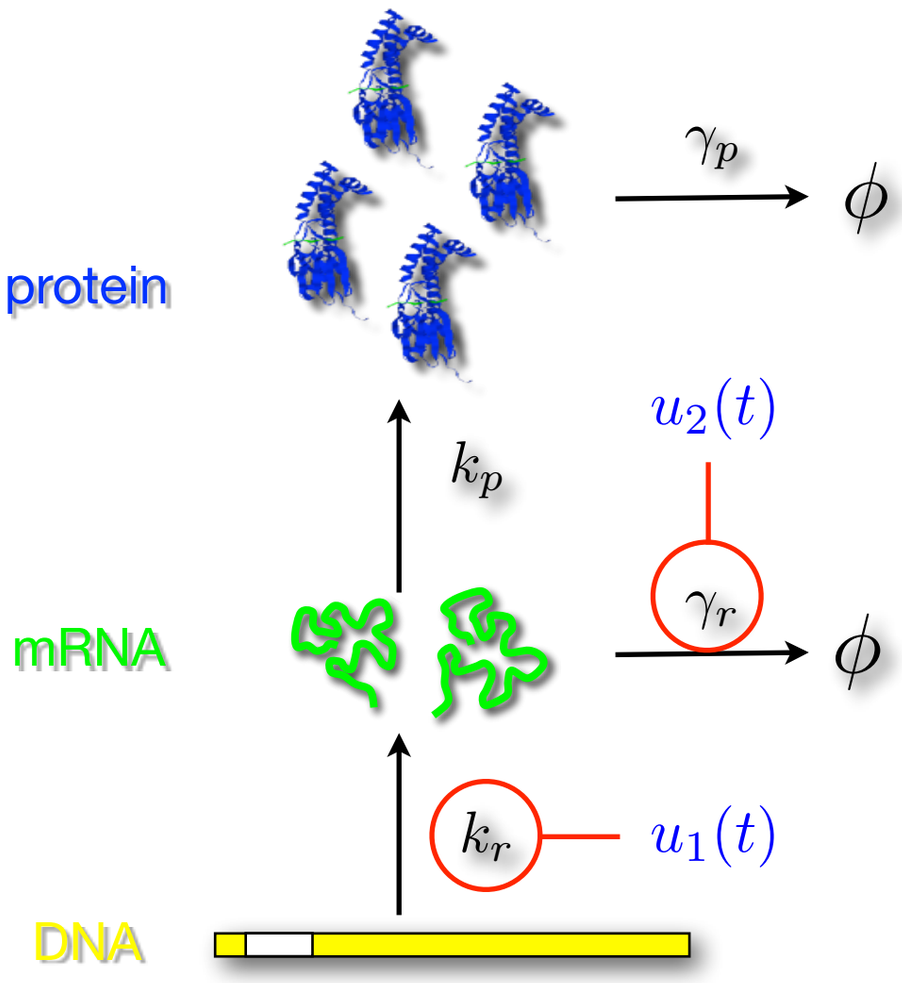


Goal: Control both protein mean and variance *independently*

Obstacle: We can prove that it is *impossible* to achieve this goal with transcriptional control alone

Possible solution: We explore the use of an additional independent control input: mRNA degradation

Mean and Variance Control



$$\frac{d\mu_m}{dt} = -u_2(t)\mu_m(t) + u_1(t)$$

$$\frac{d\mu_p}{dt} = k_p\mu_m(t) - \gamma_p\mu_p(t)$$

$$\frac{d\sigma_m^2}{dt} = u_2(t)(\mu_m(t) - 2\sigma_m^2(t)) + u_1(t)$$

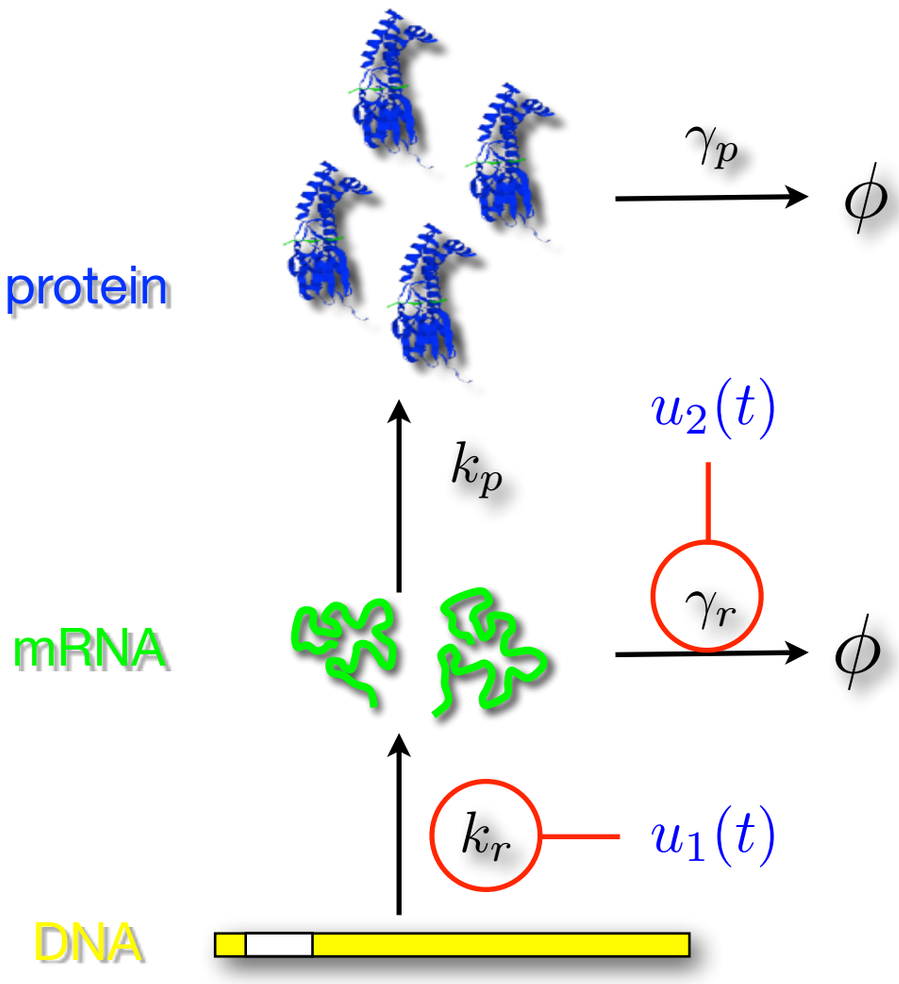
$$\frac{d\sigma_{mp}^2}{dt} = k_p\sigma_m^2(t) - \gamma_p\sigma_{mp}^2(t) - u_2(t)\sigma_{mp}(t)$$

$$\frac{d\sigma_p^2}{dt} = k_p\mu_m(t) + \gamma_p\mu_p(t) + 2k_p\sigma_{mp}^2(t) - 2\gamma_p\sigma_p^2(t)$$

Dynamical system is *bilinear*

$$\begin{aligned}
\mu_m &:= E[m]; \\
\mu_p &:= E[p]; \\
\sigma_m^2 &:= E[(m-\mu_m)^2]; \\
\sigma_{mp}^2 &:= E[(m-\mu_m)(p-\mu_p)]; \\
\sigma_p^2 &:= E[(p-\mu_p)^2]
\end{aligned}$$

Fundamental Limitations



Fact: Not all desired protein mean and variance are achievable.

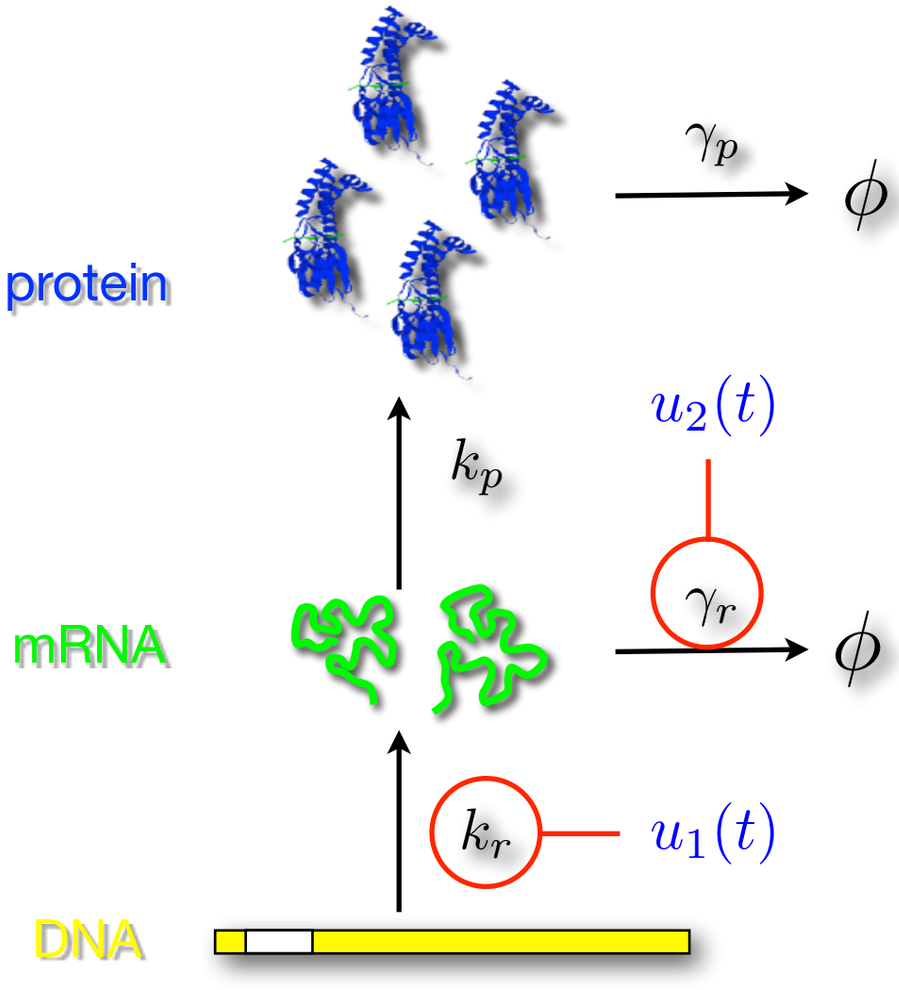
Let μ_{p*} be the *desired* protein mean

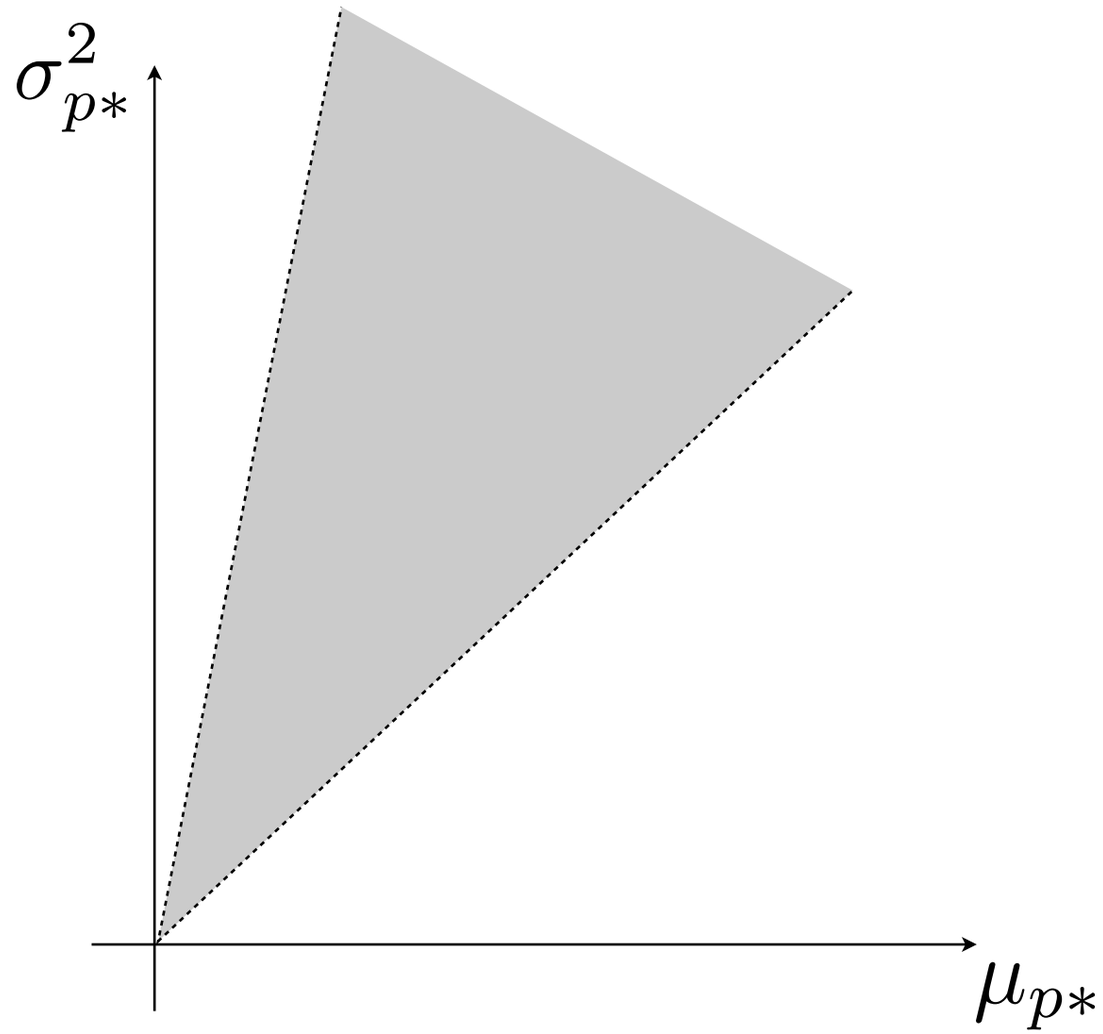
Let σ^2_{p*} be the *desired* protein variance

Fact: The set of achievable protein mean and variance is given by

$$\mu_{p*} < \sigma^2_{p*} < \left(1 + \frac{k_p}{\gamma_p}\right) \mu_{p*}$$

Fundamental Limitations

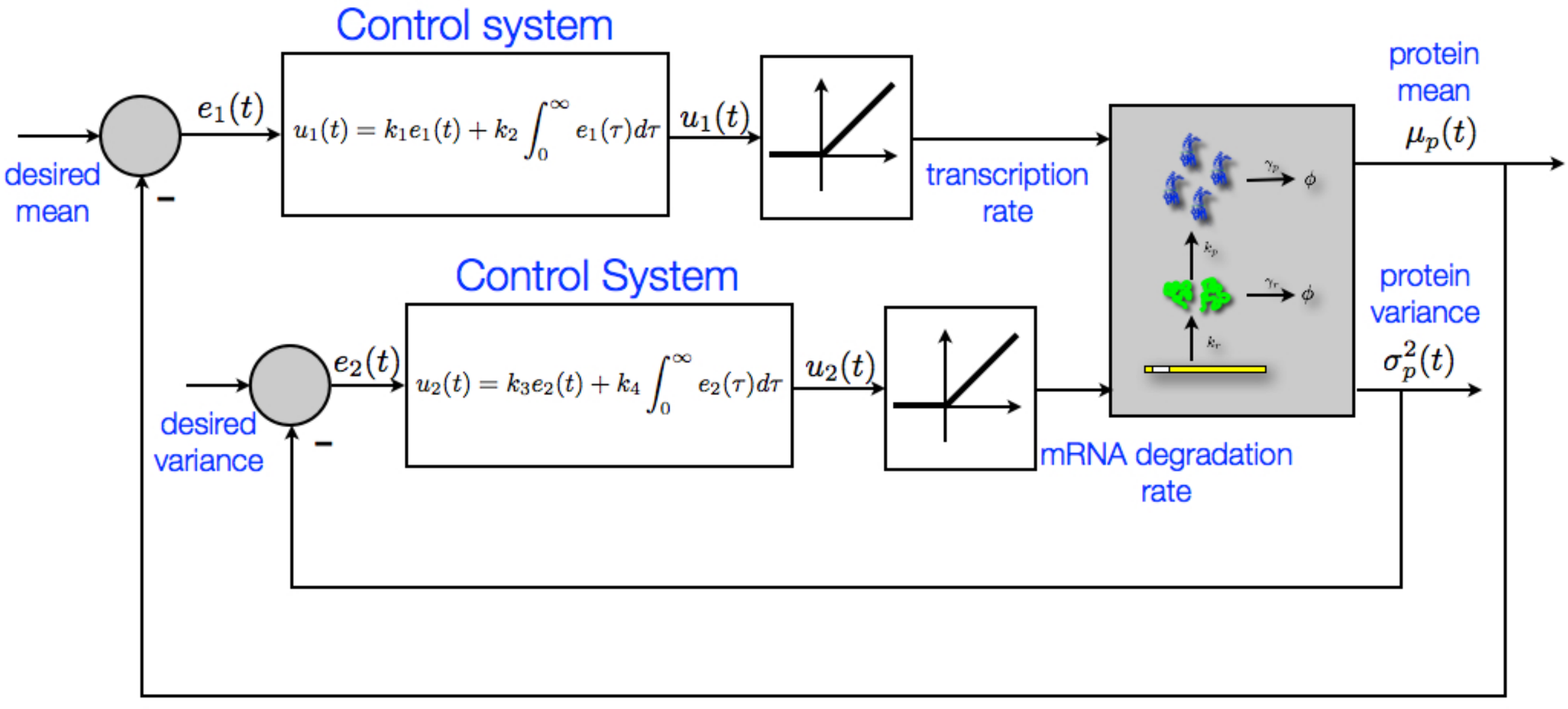


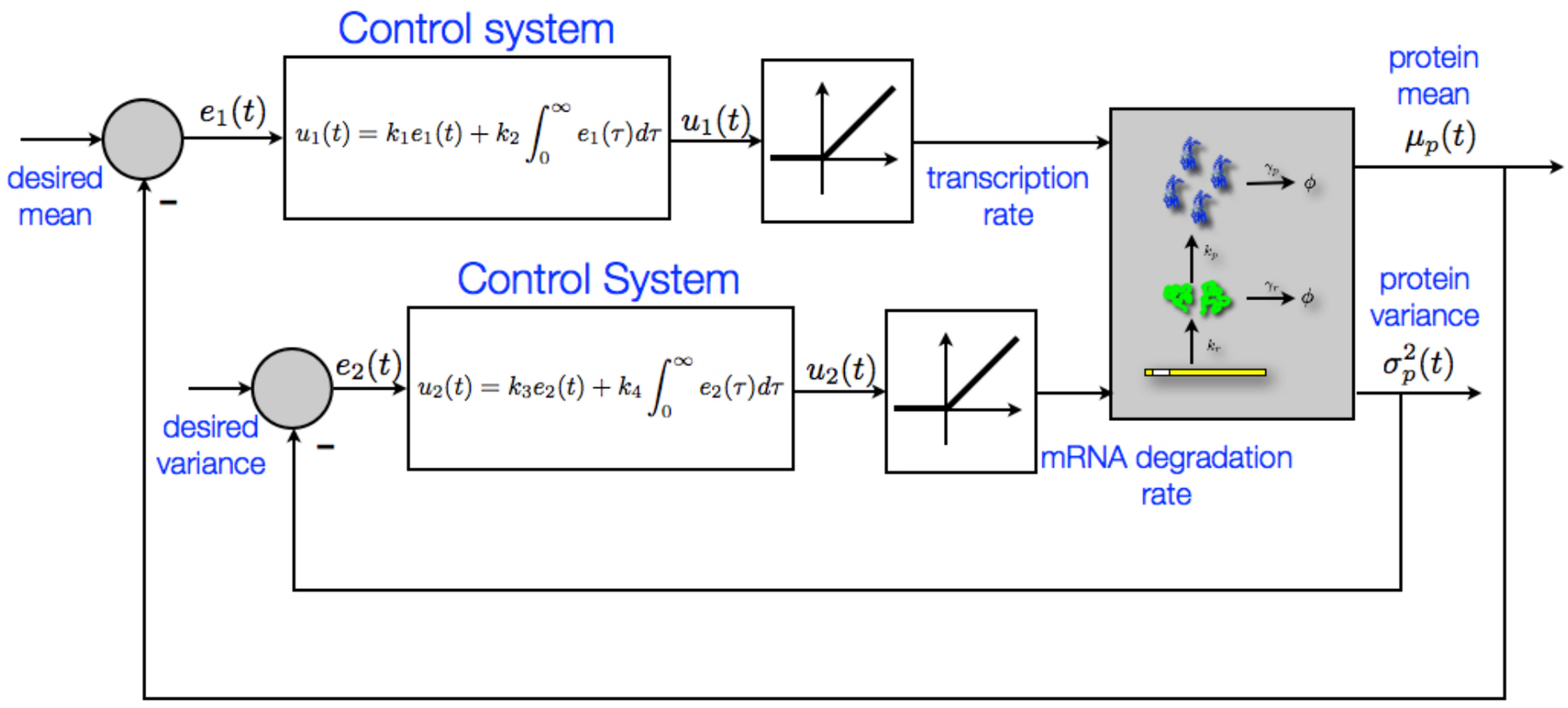


Fact: The set of achievable protein mean and variance is given by

$$\mu_{p*} < \sigma^2_{p*} < \left(1 + \frac{k_p}{\gamma_p}\right) \mu_{p*}$$

Feedback Control of Mean and Variance





Tracking of protein mean and variance with Multivariable PI feedback

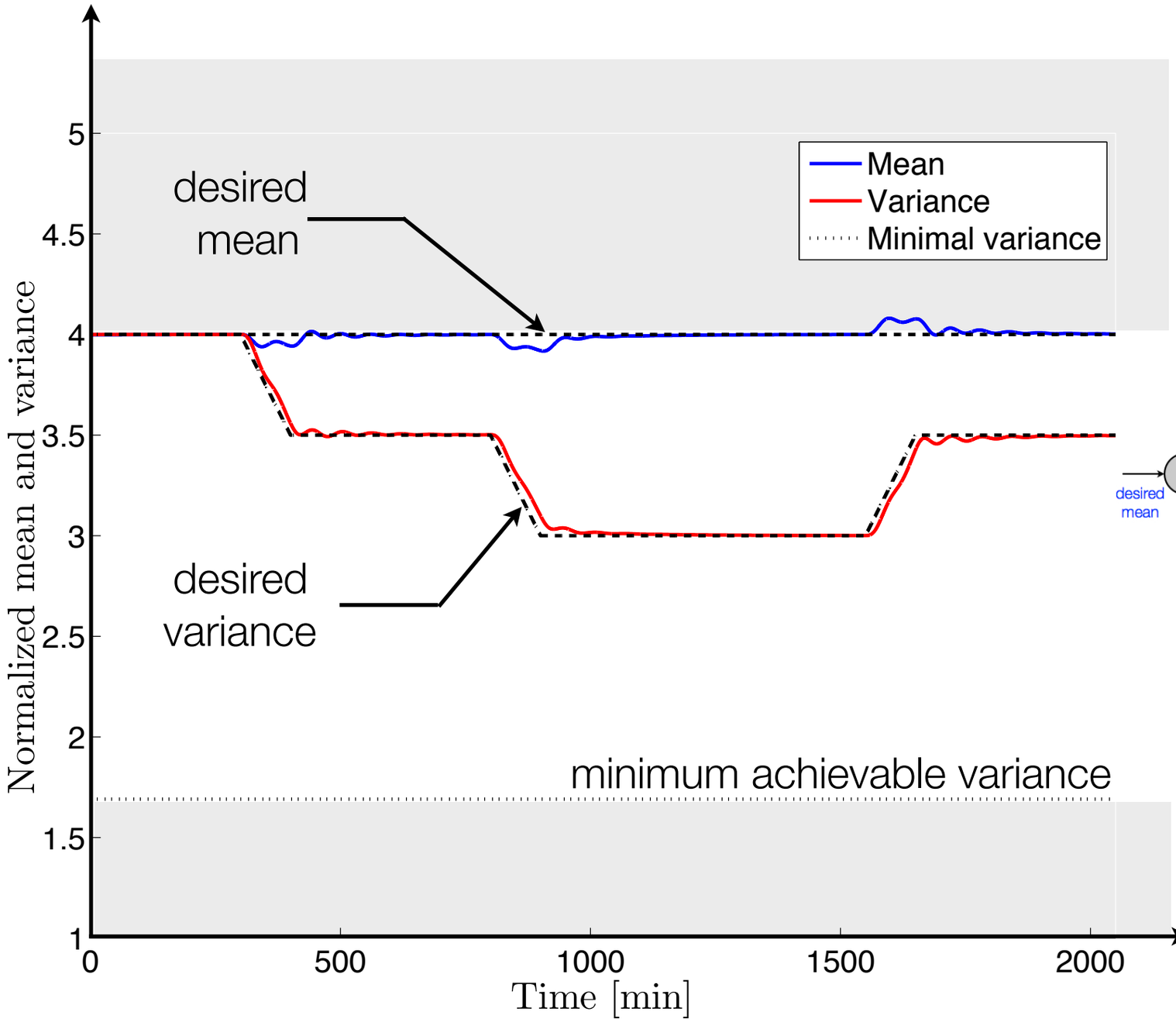
There *always* exists control parameters k_1, k_2 k_3, and k_4 such that the system is locally stable, and

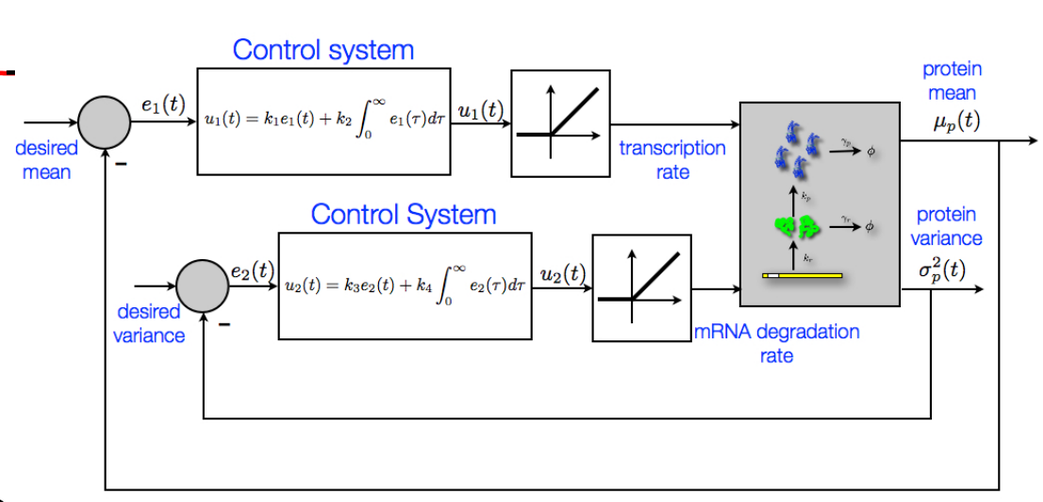
1. *the protein mean tracks asymptotically the desired mean* μ_{p*}; and
2. *the protein variance tracks asymptotically the desired mean* σ^2_{p*}

provided

$$\mu_{p*} < \sigma^2_{p*} < \left(1 + \frac{k_p}{\gamma_p}\right) \mu_{p*}$$

Simulation Example





Summary

- Gene expression is stochastic; this leads to population variability
- Variability plays an important biological role
- Probabilistic methods are required to model gene expression
- Population data can be used for statistical inference
- It is possible to control statistical properties of gene expression using external inputs