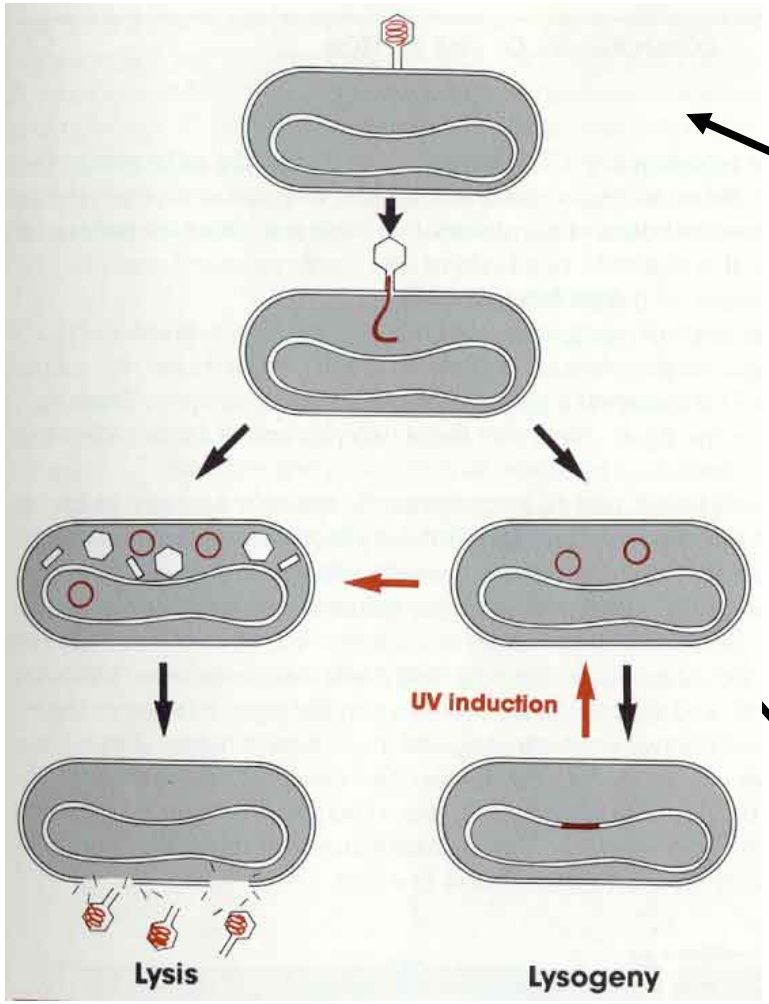
Stochastic Modeling in Systems Biology

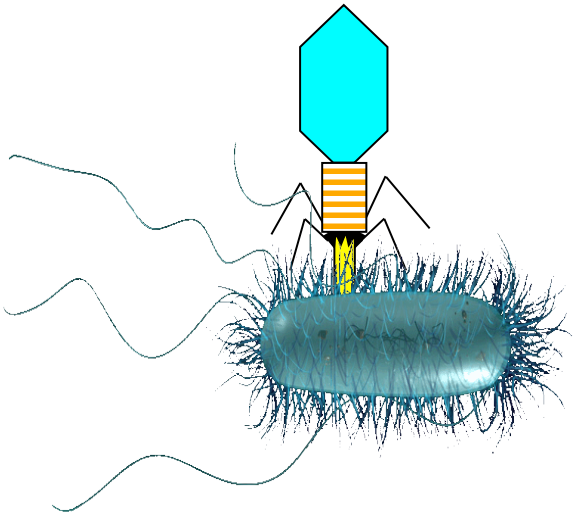
Yang Cao

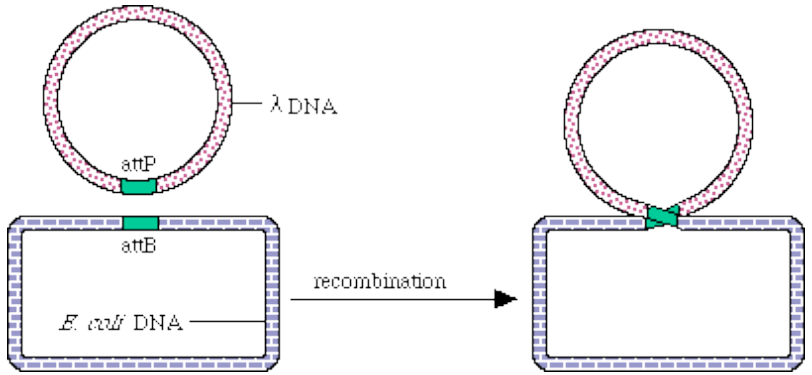
Department of Computer Science

VirginiaTech

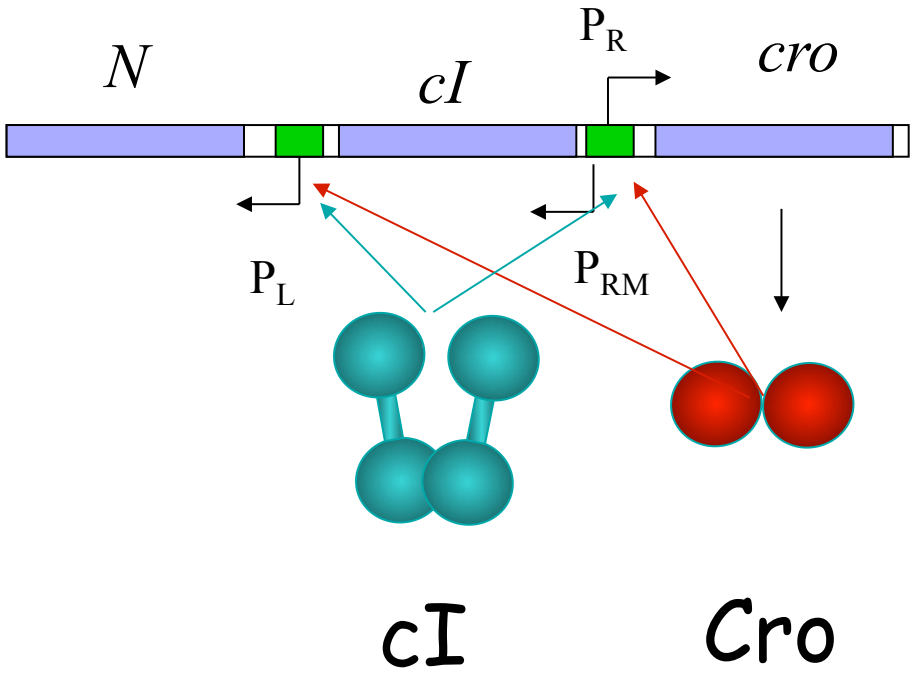
Lambda-phage affected E. Coli







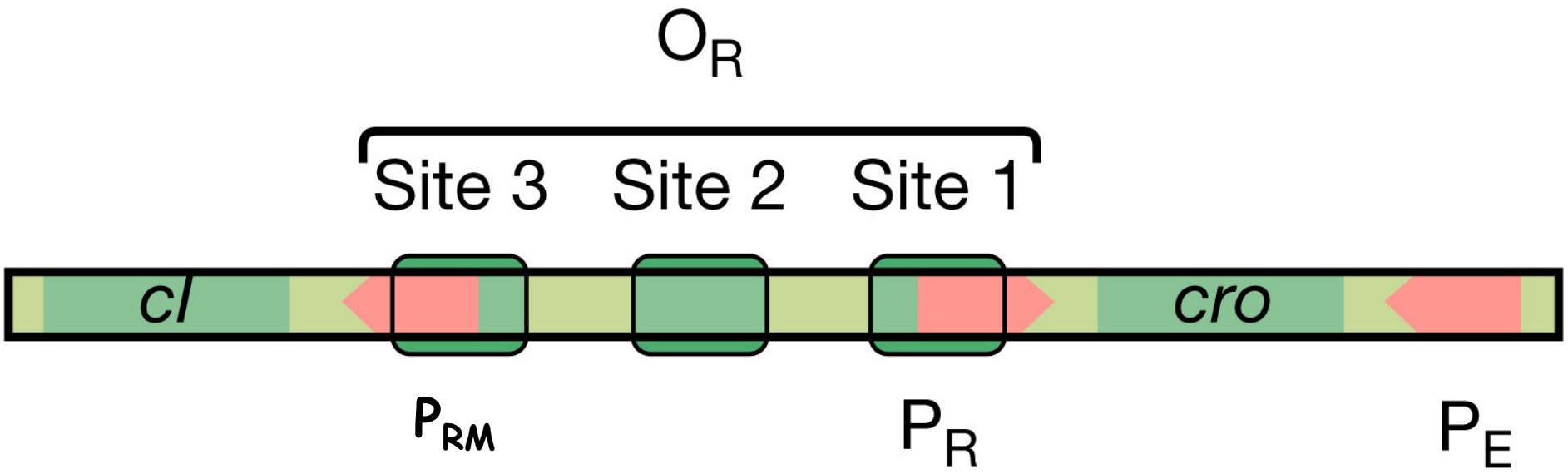
Highlight the lambda phage regulation



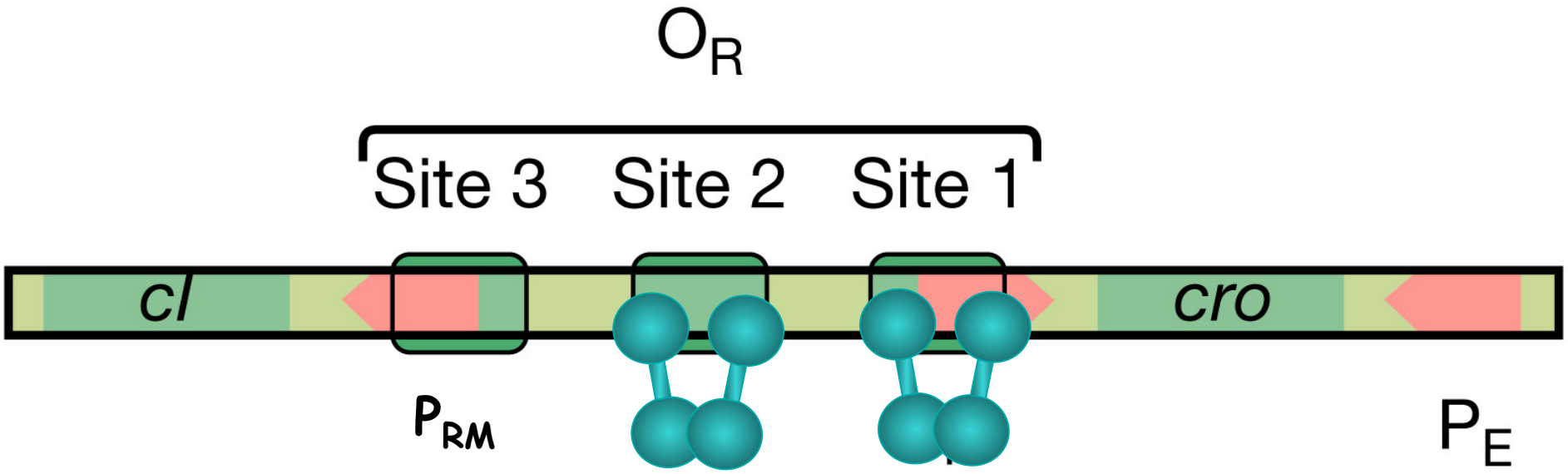
If cI wins, P_R and P_L are repressed and the cell enters lysogeny

If Cro wins, P_{RM} is repressed and the cells enters the lytic cycle

A close up on the right promoter- operator region



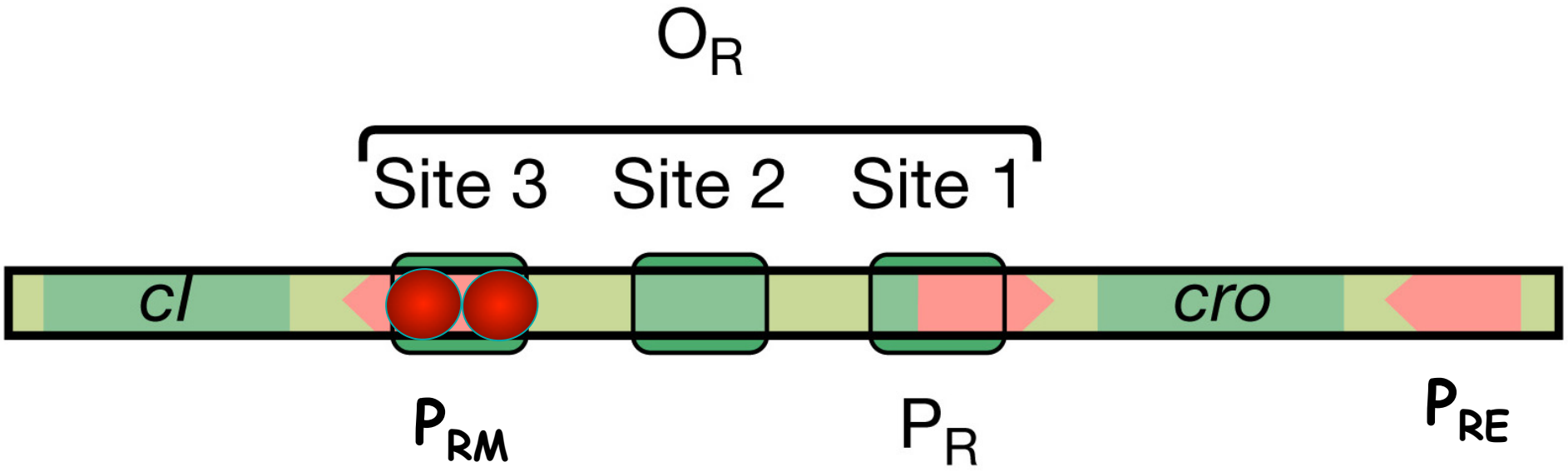
cI must bind to OR1 to repress rightwards transcription



cI represses P_R – shuts off *cro*

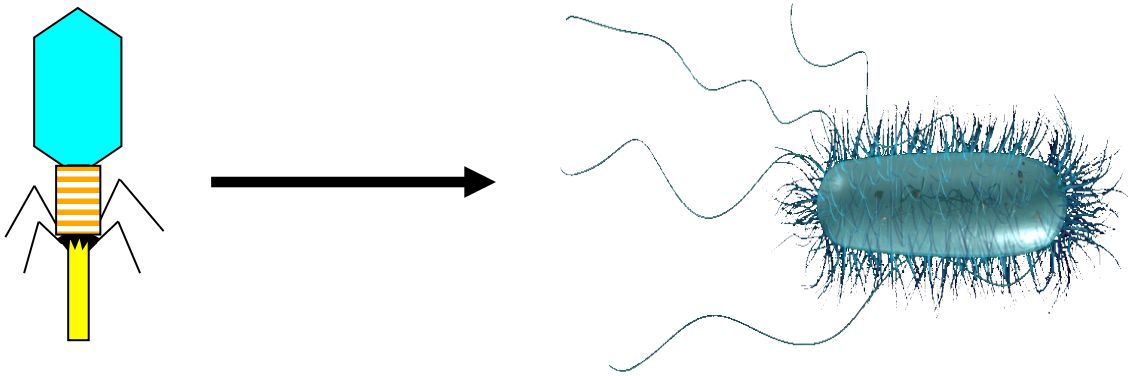
cI activates P_{RM} – expression of *cI*

Cro must bind to OR3 to repress expression of repressor by P_{RM}



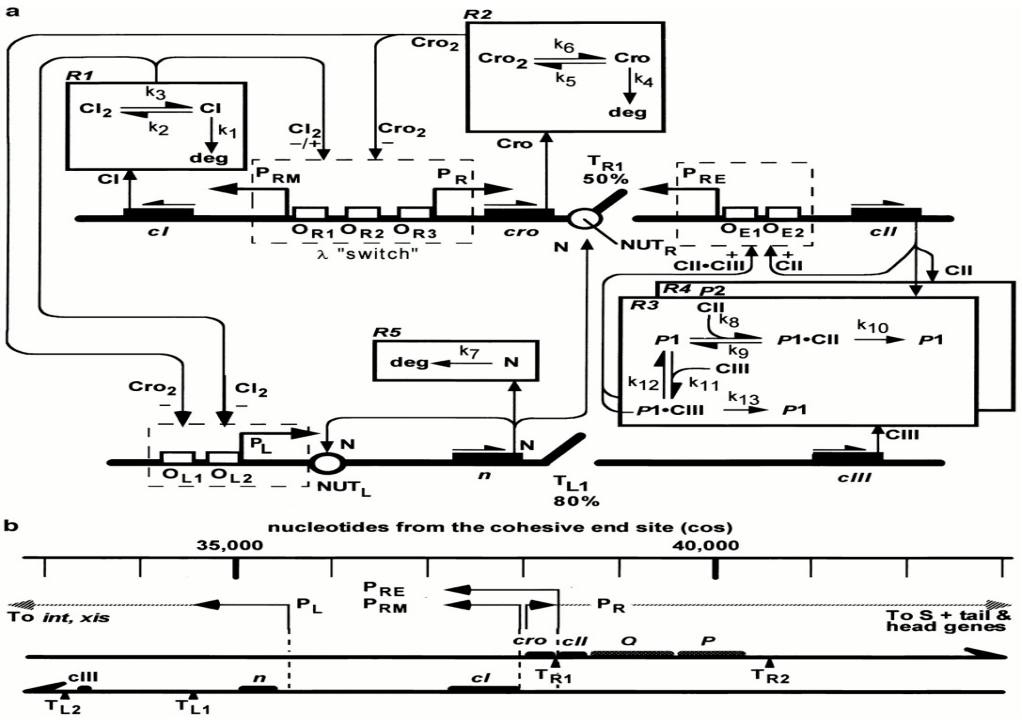
Cro represses P_{RM} – shuts off *cI* expression

Lambda-phage affected E. Coli



Lysis

Lysogeny



Stochastic effects play an important role in lytic/ lysogenic decision network

Arkin et al. 1997, 1998

Chemical Reacting System

A Chemically Reacting System

- Molecules of N ***chemical species*** $S_1, \ldots, S_N$.

  - In a volume Ω, at temperature T.

  - Different conformations or excitation levels are considered different species if they behave differently.

- M "elemental" ***reaction channels*** $R_1, \ldots, R_M$.

  - each R_j describes a *single instantaneous physical event*, which changes the population of at least one species. Thus, R_j is either

$$\varnothing \rightarrow S_i,$$

or

$$S_i \rightarrow \text{ something else},$$

or

$$S_i + S_{i'} \rightarrow \text{ something else}.$$

A General Question

<u>Question</u>: How does this system evolve in time?

The traditional answer, for *spatially homogeneous* systems:

"According to the ***reaction rate equation*** (RRE)."

- A set of coupled, first-order ODEs.
- Derived using ad hoc, phenomenological reasoning.
- Implies the system evolves *continuously* and *deterministically*.
- Empirically accurate for large systems.
- Often not adequate for small systems.

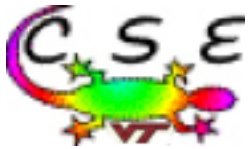
* * *

The question deserves a more carefully considered answer.

Molecular Dynamics (MD)

- The most exact way of describing the system's evolution.
- Tracks the position and velocity of *every* molecule in the system.
- Simulates *every* collision, *non-reactive* as well as *reactive*.
- Shows changes in species populations and their spatial distributions.
- ***But*** . . . it's *unfeasibly slow* for nearly all realistic systems.

Some Facts about Cell (got from Mark Paul)



Characteristic Length Scales

The radius of a single cell,

$$R_{cell} \approx 5\mu\text{m}$$

The radius of a water molecule,

$$R_{water} \approx 0.3\text{nm}$$

The volume of a single cell,

$$V_{cell} \approx \frac{1}{2}\text{pL}$$

If a cell only contained water, there would be,

$$N_{water} \approx 4 \times 10^{15}\text{molecules}$$

Characteristic Time Scales

Mean time for cell division,

$$t_{cell} \approx 100\text{min}$$

At room temperature the equipartition of energy yields the mean velocity of a molecule to be,

$$\frac{3}{2}k_B T = \frac{1}{2}mv^2$$

Using this, the mean time between collisions is,

$$t_c \lesssim 1\text{ps}$$

The number of collisions to occur per cell cycle is then,

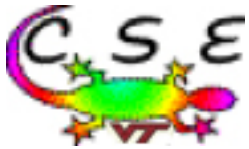
$$N \approx 3 \times 10^{31}$$

Molecular dynamics is out!

A great simplification occurs ***if*** successive *reactive* collisions tend to be separated in time by *very many non-reactive* collisions.

- The overall effect of the non-reactive collisions is to *randomize* the positions of the molecules (and also maintain thermal equilibrium).
- The non-reactive collisions merely serve to keep the system **well-stirred** or **spatially homogeneous** for the reactive collisions.
- Can describe the *state* of the system by $\mathbf{X}(t) \triangleq \big(X_1(t), \ldots, X_N(t) \big)$,

$$X_i(t) \triangleq \text{ the } \textit{number} \text{ of } S_i \text{ molecules at time } t\,.$$



But this ***well-stirred simplification***, which . . .

- ignores the non-reactive collisions,
- *truncates* the definition of the system's state,

. . . comes at a price:

X(t) ***must be viewed as a stochastic process*.**

- In fact, ***the system was never deterministic to begin with!***
  Even if molecules moved according to classical mechanics . . .
  - monomolecular reactions always involve QM.
  - bimolecular reactions require collisions, whose extreme sensitivity to initial conditions renders them essentially random.
  - bimolecular reactions usually involve QM too.
- But stochastic processes can be handled.

Characteristics of the System

For well-stirred systems, each R_j is completely characterized by:

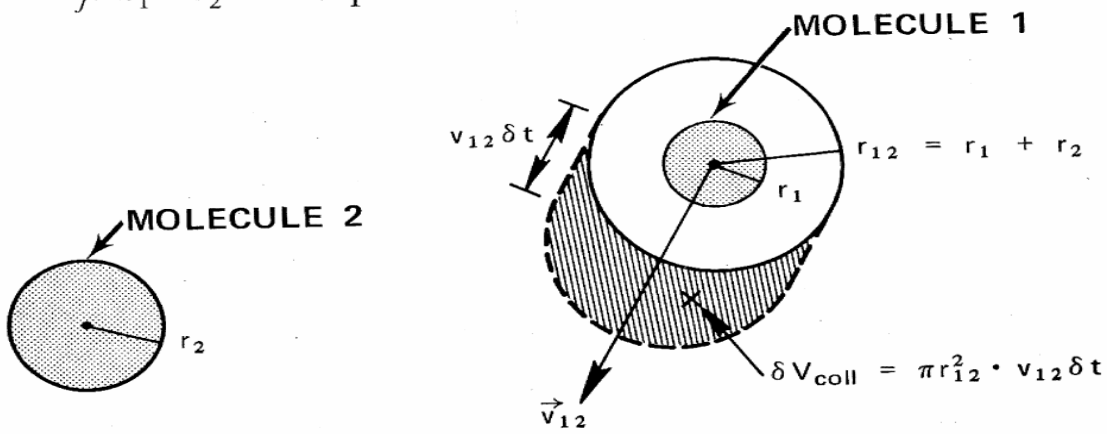
- a ***propensity function*** $a_j(\mathbf{x})$: Given the system in state $\mathbf{x}$, $a_j(\mathbf{x})\,dt \triangleq$ the *probability* that one R_j event will occur in the next dt.
  - The existence and form of $a_j(\mathbf{x})$ follow from kinetic theory.
  - $a_j(\mathbf{x})$ is roughly equal to, but is *not* derived from, the RRE "rate".
- a ***state change vector*** $\boldsymbol{\nu}_j \equiv (\nu_{1j},\ldots,\nu_{Nj})$: $\nu_{ij} \triangleq$ the *change* in the number of S_i molecules caused by one R_j event.
  - R_j induces $\mathbf{x} \to \mathbf{x}+\boldsymbol{\nu}_j$. $\{\nu_{ij}\} \equiv$ the "stoichiometric matrix."

E.g.,

$$S_1 + S_2 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} 2S_1 \quad \Rightarrow \quad \begin{cases} a_1(\mathbf{x}) = c_1 x_1 x_2, & \boldsymbol{\nu}_1 = (+1,-1,0,\ldots,0) \\ a_2(\mathbf{x}) = c_2 \dfrac{x_1(x_1-1)}{2}, & \boldsymbol{\nu}_2 = (-1,+1,0,\ldots,0) \end{cases}$$

One Reasoning (but not a proof)

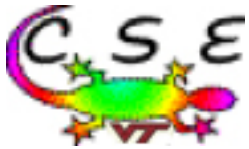
$$R_j:\ S_1 + S_2 \xrightarrow{c_j} \text{products}$$



$$\underbrace{\left\{\left(\frac{(\bar{v}_{12}dt)(\pi r_{12}^2)}{\Omega}\right)\times\exp\left(-E_j/k_BT\right)\right\}}_{\text{Probability that a randomly chosen } S_1\text{-}S_2 \text{ pair will react according to } R_j \text{ in next } dt.}\times x_1x_2 = \underbrace{\left(\frac{\bar{v}_{12}\pi r_{12}^2}{\Omega}\exp\left(-E_j/k_BT\right)\right)}_{c_j} x_1x_2\,dt$$

$$\triangleq \underbrace{\left(c_j\,x_1x_2\right)}_{a_j(\mathbf{x})}dt \triangleq a_j(\mathbf{x})\,dt$$

Propensity Function vs the Reaction Rate



- **Difference**
  - Propensity function describes the probability while reaction rate describes the changing rate.
  - Propensity functions are defined based on population of species while the reaction rates are defined based on the concentration of species
- **Connection**
  - For simple system, they have similar format.
    For reaction like $A \rightarrow B$
    - Reaction rate: $k[A]$
    - Propensity function cx_A and we have $c = k$
  - For reaction like $A + B \rightarrow C$
    - Reaction rate $k[A][B]$
    - :
    - Propensity function $cx_A x_B$ and we have $c = k / V$

Chemical Master Equation

Two exact, rigorously derivable consequences . . .

➢ **1.** The ***chemical master equation*** (CME):

$$\frac{\partial P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)}{\partial t}=\sum_{j=1}^{M}\Big[a_j(\mathbf{x}-\boldsymbol{\nu}_j)P(\mathbf{x}-\boldsymbol{\nu}_j,t\,|\,\mathbf{x}_0,t_0)-a_j(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\Big].$$

- Gives $P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\triangleq \text{Prob}\{\mathbf{X}(t)=\mathbf{x}, \text{ given } \mathbf{X}(t_0)=\mathbf{x}_0\}$ for $t\geq t_0$.
- The CME follows from the *probability* statement

$$P(\mathbf{x},t+dt\,|\,\mathbf{x}_0,t_0)=P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\times\left[1-\sum_{j=1}^{M}\Big(a_j(\mathbf{x})dt\Big)\right]$$
$$+\sum_{j=1}^{M}P(\mathbf{x}-\boldsymbol{\nu}_j,t\,|\,\mathbf{x}_0,t_0)\times\Big(a_j(\mathbf{x}-\boldsymbol{\nu}_j)dt\Big).$$

- But it's practically always *intractable* (analytically and numerically).

➢ **2.** The ***stochastic simulation algorithm*** (SSA):

- A procedure for constructing *sample paths* or *realizations* of $\mathbf{X}(t)$.
- Approach: Generate the time to the *next* reaction and the index of that reaction.
- Theoretical justification: With $p(\tau, j|\mathbf{x},t)$ defined by

$p(\tau, j|\mathbf{x},t)\,d\tau \triangleq$ prob, given $\mathbf{X}(t)=\mathbf{x}$, that the next reaction will occur in $[t+\tau, t+\tau+d\tau)$, *and* will be an R_j,

can prove that

$$p(\tau, j|\mathbf{x},t) = a_j(\mathbf{x}) \exp\left(-a_0(\mathbf{x})\tau\right), \quad \text{where } a_0(\mathbf{x}) \triangleq \sum_{j'=1}^{M} a_{j'}(\mathbf{x}).$$

This implies that the *time* τ to the next reaction event is an exponential random variable with mean $1/a_0(\mathbf{x})$, and the *index* j of that reaction is an integer random variable with prob $a_j(\mathbf{x})/a_0(\mathbf{x})$.

Direct Method (DM)

The "Direct" Version of the SSA

1. With the system in state $\mathbf{x}$ at time t, evaluate $a_0(\mathbf{x}) \triangleq \sum_{j'=1}^{M} a_{j'}(\mathbf{x})$.

2. Draw two unit-interval uniform random numbers r_1 and r_2, and compute τ and j according to

- $\tau = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{r_1}\right)$,
- j = the *smallest integer* satisfying $\sum_{j'=1}^{j} a_{j'}(\mathbf{x}) > r_2\, a_0(\mathbf{x})$.

3. Replace $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$.

4. Record $(\mathbf{x}, t)$. Return to Step **1**, or else end the simulation.

(Inverse Generation Method)

First Reaction Method (FRM)

- A different but equivalent simulation method for SSA
- Generate a firing time for each reaction channel

$$\tau_j = \frac{1}{a_j(x)} \ln\left(\frac{1}{r_j}\right),$$

- Find the minimum of all the firing time and throw off all the others

μ = the index satisfying

$$\tau_\mu = \min_j \tau_j,$$

- In theory this method is equivalent to the DM
- The biggest concern is the computational cost

Next Reaction Method (NRM)

- Derived by Gibson and Bruck 2000
- Keep the randomly generated firing time if the corresponding propensity function is not changed (Use a dependent graph (DG) to achieve this goal)
- Use absolute time instead of relative time

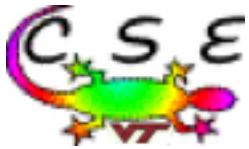
$$\tau_\alpha = \frac{1}{a_\alpha(t)} \log\left(\frac{1}{r}\right) + t.$$

- Reuse the unapplied uniform random number

$$\tau_\alpha \leftarrow \left(a_{\alpha,old}/a_{\alpha,new}\right)\left(\tau_\alpha - t\right) + t.$$

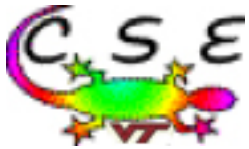
- Use a priority tree (heap array) to conduct the search
- Only need to generate one uniform random number each step
- The computational cost to maintain the data structure can be hug!

Optimized Direct Method (ODM)



- Analyze the time profile of DM and NRM. Conclude that the data structure maintaining cost is very high for most problems.
- Sort the reaction channel so that more frequent firing reaction channels have smaller index. (Run a few pre-simulation to collect info for the problem.)
- If necessary, use the dependent graph (DG) to avoid recalculating propensity functions for reaction channels that are not affected by the last reaction.
- Efficient for multiscale problem
- Among all test problems we tried, ODM is faster than NRM.

Sorted Direct Method (SDM)



- The pro-simulation procedure is troublesome for an automatic code.
- The index should be changed dynamically during the simulation
- Bubble sorting technique
  If a reaction just fired, move its index one step up. After a while, the reaction index will be automatically well-sorted.

- Less (but almost the same) efficient than ODM but much easier to code and maintain.

The SSA . . .

- Is *exact*.
- Is equivalent to (but is not derived from) the CME.
- Does *not* entail approximating "dt" by "Δt".
- Is *procedurally simple*, even when the CME is intractable.

- ***Remains too slow for most practical problems***: Simulating *every* reaction event, *one* at a time, is just too much work if any reactant is present in very large numbers.

A Model for Prokaryotic Gene Expression

1. **Transcription Initiation (the binding and initiation)**

$$\mathrm{P+RNAP} \rightarrow \mathrm{P \bullet RNAP} \qquad k_1 = 10^8\,\mathrm{M}^{-1}s^{-1}$$

$$\mathrm{P \bullet RNAP} \rightarrow \mathrm{P+RNAP} \qquad k_2 = 10s^{-1}$$

$$\mathrm{P \bullet RNAP} \rightarrow \mathrm{TrRNAP} \qquad k_3 = 1s^{-1}$$

2. **Elongation (RBS is available before elongation terminates**

$$\mathrm{TrRNAP} \rightarrow \mathrm{RBS + P + ElRNAP} \qquad k_4 = 1s^{-1}$$

3. **Translation Initiation**

$$\mathrm{Ribosome + RBS} \rightarrow \mathrm{RibRBS} \qquad k_5 = 10^8\,\mathrm{M}^{-1}s^{-1}$$

$$\mathrm{RibRBS} \rightarrow \mathrm{Ribosome + RBS} \qquad k_6 = 2.25s^{-1}$$

$$\mathrm{RibRBS} \rightarrow \mathrm{ElRib + RBS} \qquad k_7 = 0.5s^{-1}$$

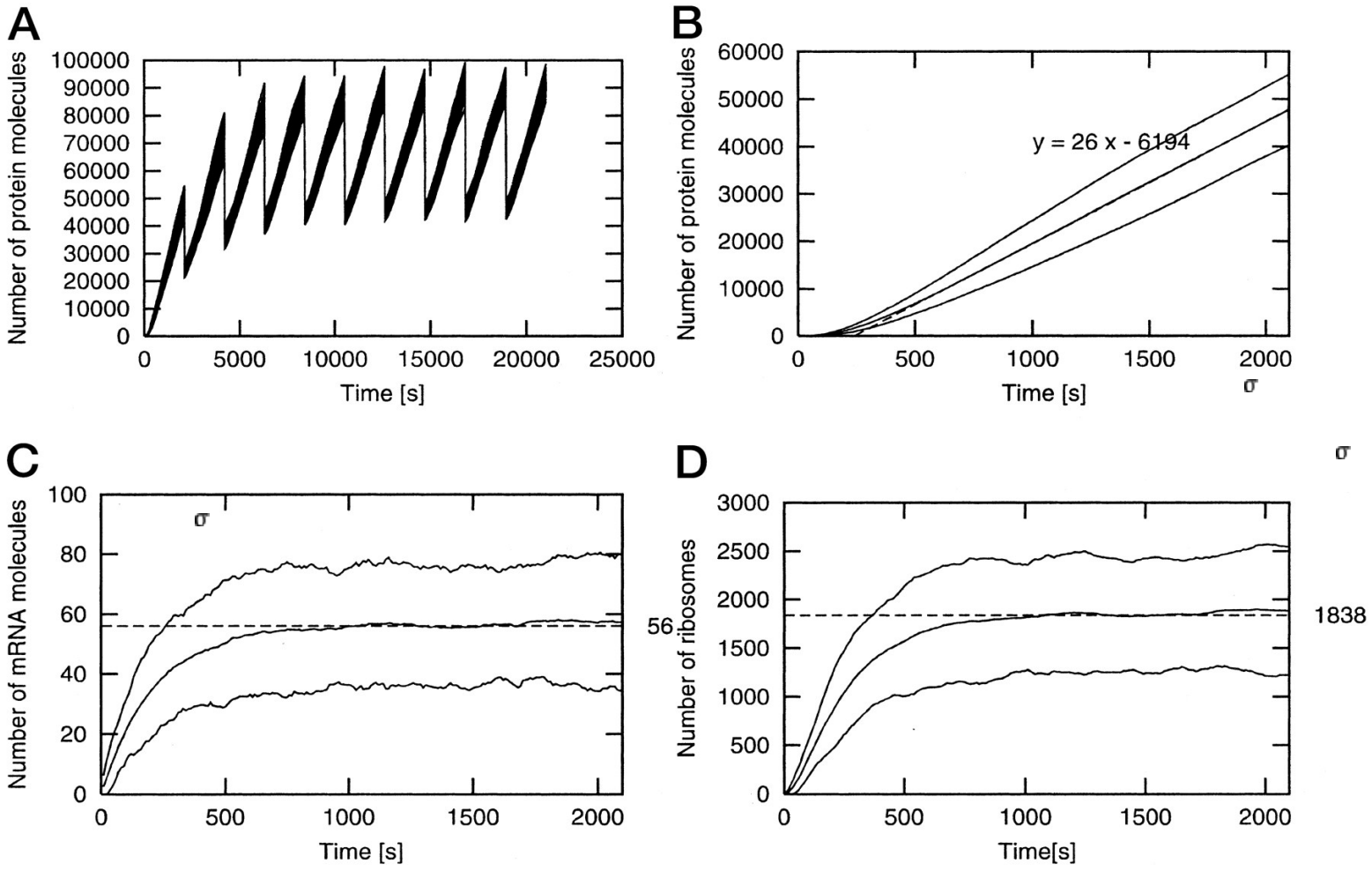
$$\mathrm{RBS} \rightarrow \mathrm{decay} \qquad k_8 = 0.3s^{-1}$$

4. **Elongation**

$$\mathrm{ElRib} \rightarrow \mathrm{Protein} \qquad k_9 = 0.015s^{-1}$$

$$\mathrm{Protein} \rightarrow \mathrm{decay} \qquad k_{10} = 6.42 \times 10^{-5}s^{-1}$$

Simulation Results



Kierzek, A. M. et al. J. Biol. Chem. 2001;276:8165-8172

Schlögl Model and Error Measurement

- Reactions:

$$B_1 + 2X \underset{c_2}{\overset{c_1}{\leftrightarrow}} 3X$$

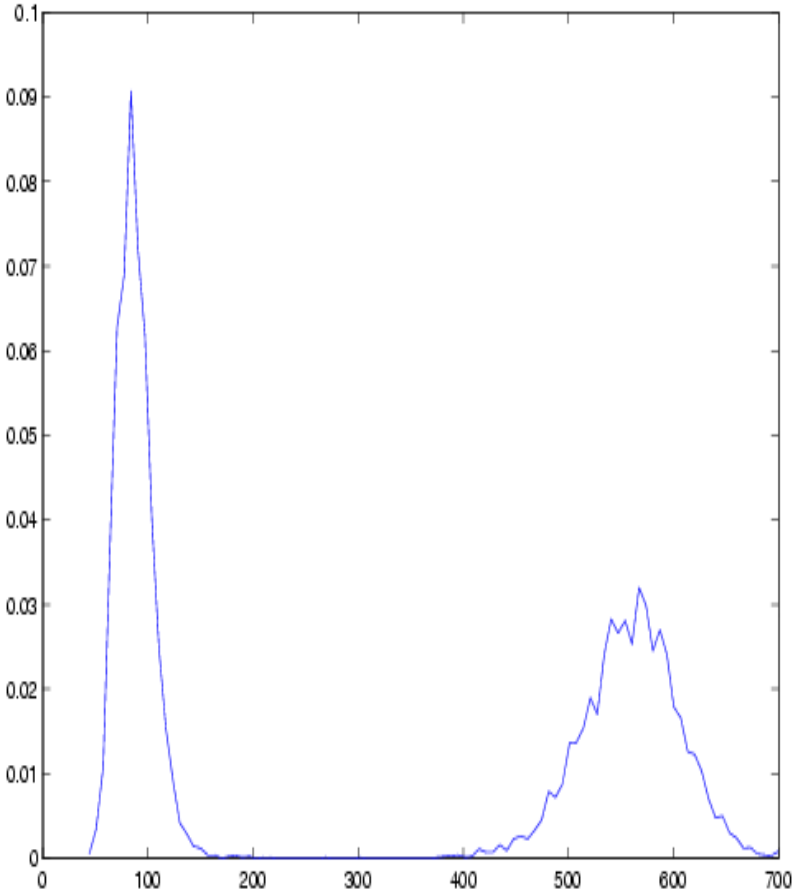
$$B_2 \underset{c_4}{\overset{c_3}{\leftrightarrow}} X$$

- Propensity functions:

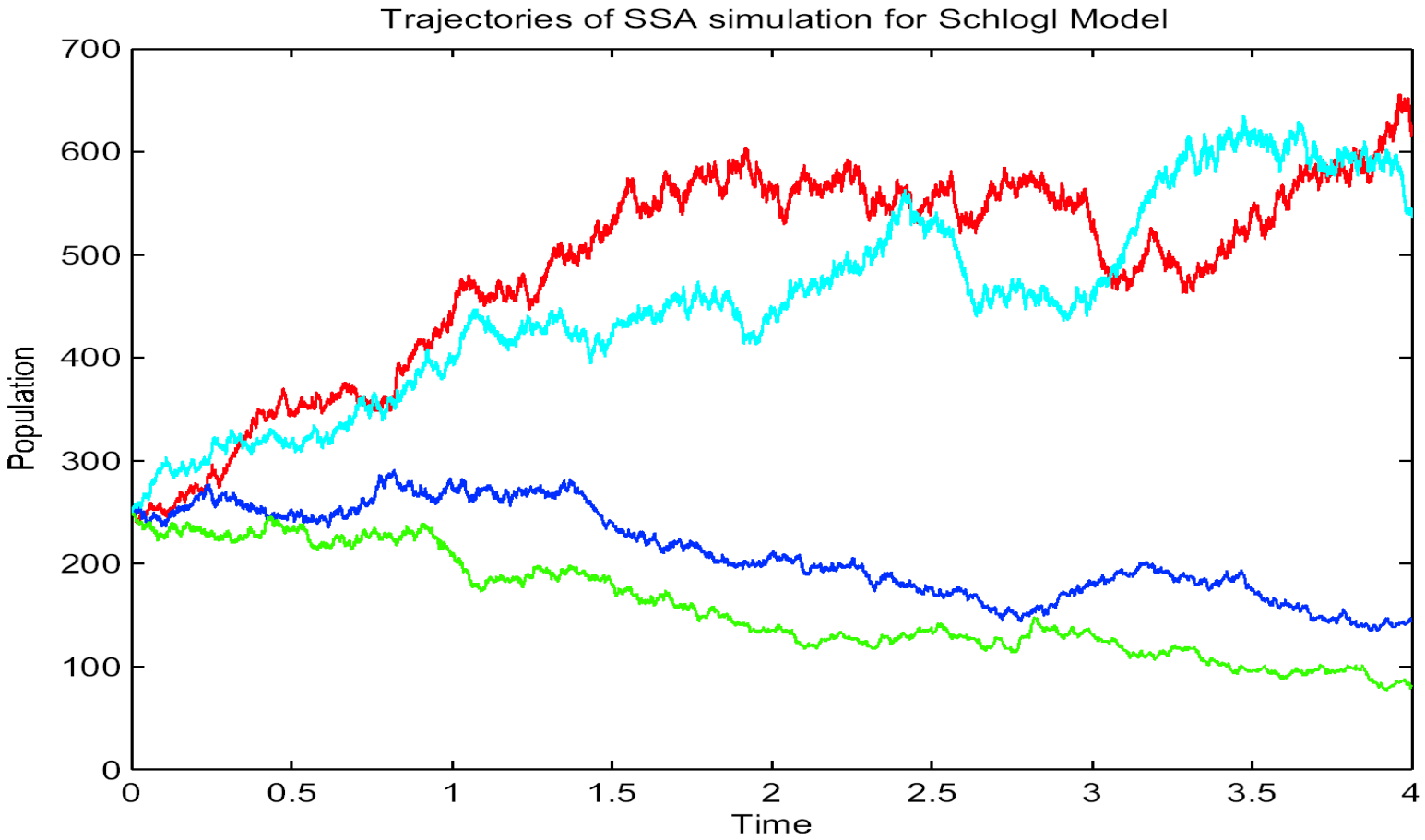
$$a_1(x) = \frac{c_1}{2} N_1 x(x-1),$$
$$a_2(x) = \frac{c_2}{6} x(x-1)(x-2),$$
$$a_3(x) = c_3 N_2,$$
$$a_4(x) = c_4 x.$$

- Bistable distribution

Histogram plot of the state in Schlögl model



Single Simulation Results



Stochastic Modeling

Lotka reactions:

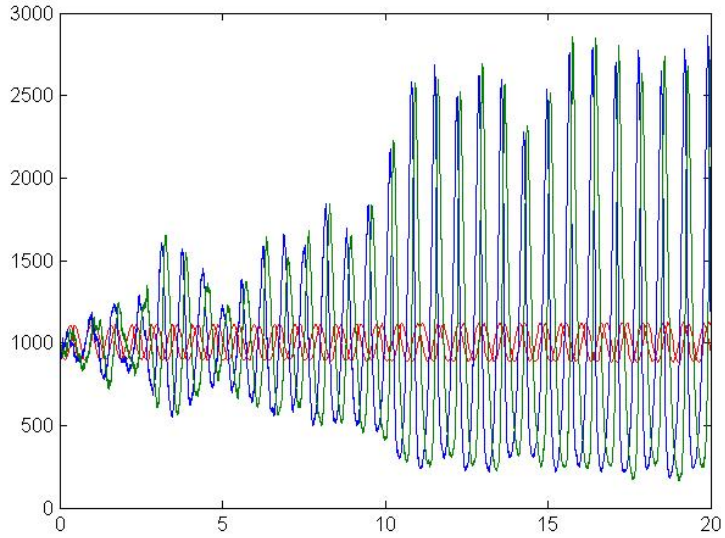
$$A + X \xrightarrow{c_1} 2X$$
$$X + Y \xrightarrow{c_2} 2Y$$
$$Y \xrightarrow{c_3} Z$$

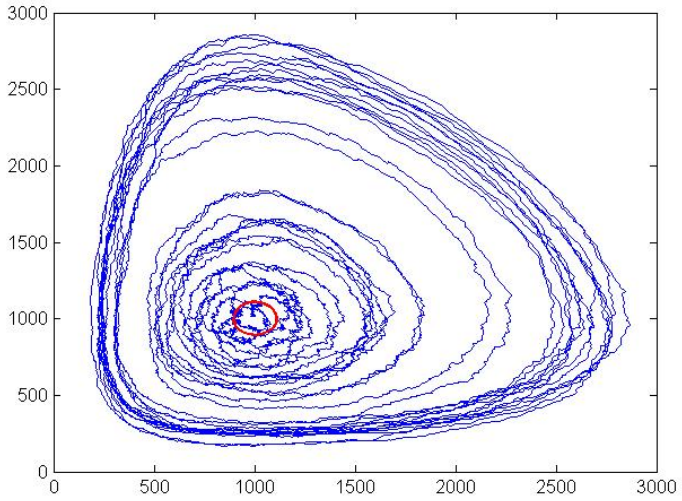
Lead to ODEs

$$\begin{cases} \dot{x} = (c_1 A - c_2 y)x \\ \dot{y} = (-c_3 + c_2 x)y \end{cases}$$

The stochastic simulation generates interestin

$$c_1 A = 10,$$
$$c_2 = 0.01,$$
$$c_3 = 10$$

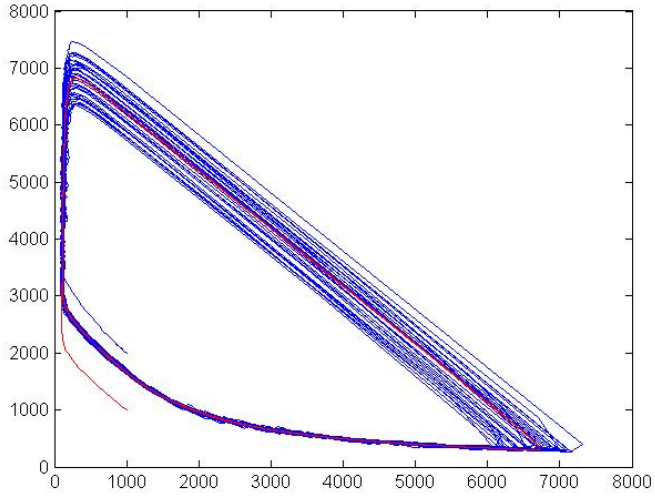




Brusselator

$$A \xrightarrow{c_1} X$$
$$B + X \xrightarrow{c_2} Y + C$$
$$2X + Y \xrightarrow{c_3} 3Y$$
$$Y \xrightarrow{c_4} D$$

$c_1 A = 5000,$
$c_2 B = 50,$
$c_3 = 0.00005,$
$c_4 = 5.$



Lead to ODEs

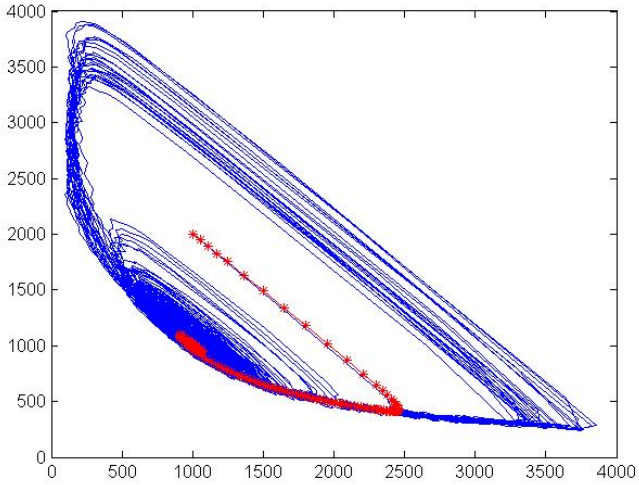
$$\begin{cases} \dot{x} = c_1 A - c_2 B y + \frac{c_3}{2} x^2 y - c_4 x \\ \dot{y} = c_2 B y - \frac{c_3}{2} x^2 y \end{cases}$$

Bifurcation happens around the condition:

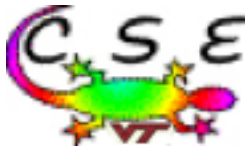
$$\frac{2c_2 B}{c_3} = \frac{(c_1 A)^2}{c_4^2} + \frac{2c_4}{c_3}$$

J. Tyson's 1973, 1974 paper

$c_1 A = 5000,$
$c_2 B = 50,$
$c_3 = 0.0001,$
$c_4 = 5.$



Break the Assumption



Spatially Inhomogeneous Systems

- Spatial homogeneity does *not* require that all equal-size subvolumes of Ω contain the same number of molecules!
- The CME and SSA require only that the center of a *randomly* chosen S_i molecule be found with *equal probability* at any point inside Ω.
- A system consisting of only one molecule can be "well-stirred".

➢ **But the well-stirred assumption can't always be made.**

In that case, we must do something different; however, the traditional ***reaction-diffusion equation*** (RDE) is *not always* the answer:

- The RDE (like the RRE) is *continuous* and *deterministic*.
- It assumes that each $d\Omega$ contains a spatially homogeneous mixture of *infinitely* many molecules.
- Not the case in most cellular systems, where spatial inhomogeneity arises not from slow mixing but rather from compartmentalization caused by highly heterogeneous structures within the cell.