Biology of the *noisy* gene

Universidad Autónoma de Madrid

Jan 2008

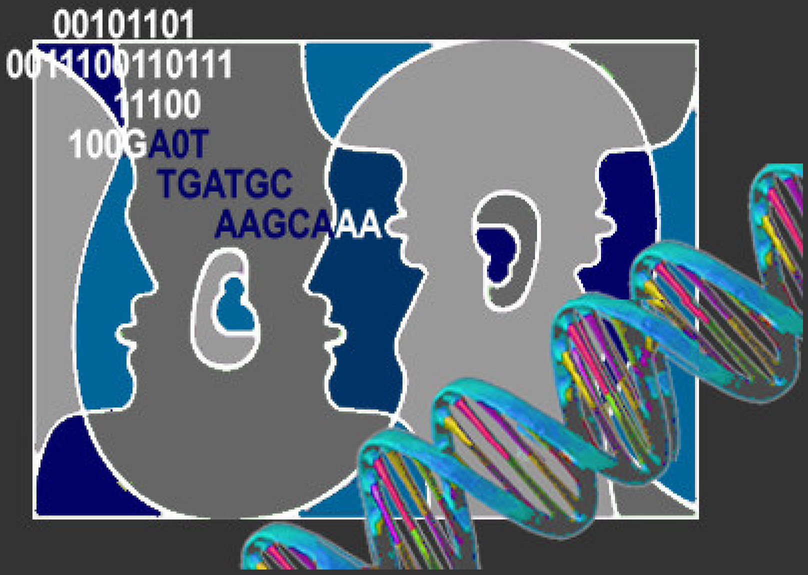
Juan F. Poyatos

Logic of Genomic Systems laboratory

CSIC

Spanish National Biotechnology Centre (CNB)

day II: modeling noise



- Two-step gene expression model
- Translational Bursting
- Gillespie’s algorithm
- Langevin equations

Stochastic dynamics of gene expression, summary

- Genes are expressed by means of chemical reactions

- Chemical reactions are stochastic processes (collisions, etc)

- Gene expression is noisy: ***intrinsic*** noise (fluctuating reaction rates) TODAY
extrinsic noise (molecules involved in gene expression)

A simple model of gene expression, summary

$$\begin{aligned} \emptyset &\xrightarrow{k} P, \\ P &\xrightarrow{\delta} \emptyset, \end{aligned}$$

Poisson process (birth and death)

- protein produced on average every 1/k seconds (birth)

$$\frac{d[P]}{dt} = k - \delta[P]$$

- protein decays with rate δ (death)

What is the steady state distribution of this model?

What is noise then?

A simple model of gene expression, summary

Recall; a new framework, the master equation

$$\frac{dp(n)}{dt} = -p(n)(r_k + n_P r_\delta) + p(n-1)r_k + p(n+1)(n_P + 1)r_\delta$$

Some comments:

- All moments of the distribution p(n) can be derived from it
- It is a linear equation in p(n).
- Solving the master equation can be done for simple systems, however only normally at steady state.
- In connection with experiments, p(n) would represent the fraction of cells having n copies of some given protein

A simple model of gene expression, summary

Poisson distribution

mean $\langle n \rangle = \langle n \rangle_{ss}$

variance $\sigma^2 = \langle n \rangle_{ss}$

Macroscopic statistics

What is noise then?

standard deviation

definition-1 (coe. variat.) = $n_1 = \frac{\sigma}{\langle n \rangle}$ (= $1/\sqrt{\langle n \rangle}$. Poisson distribution, noise increases as the number of molecules decreases)

definition-2 (Fano factor) = $n_2 = \frac{\sigma^2}{\langle n \rangle}$ (= 1, Poisson distribution mean = variance)

Recovering the deterministic law

Equation of the mean; emergence of deterministic law

$$\boxed{\frac{d\langle n\rangle}{dt} \quad = \quad \sum_n n\frac{dp_n}{dt}}$$

$$\begin{aligned}
&= \sum_n n[-p_n(r_k + nr_\delta) + p_{n-1}r_k + p_{n+1}(n+1)r_\delta] \\
&= -r_k\langle n\rangle - r_\delta \sum_n n^2 p_n + r_k \sum_n p_{n-1}n + r_\delta \sum_n n(n+1)p_{n+1} \\
&= \underline{r_k - r_\delta\langle n\rangle}.
\end{aligned}$$

Recovering the deterministic law

Considering that $[P] = \dfrac{\langle n \rangle}{V}$

We can rewrite the deterministic equation as

$$\frac{d\langle n \rangle}{dt} = Vk - \delta V[P] = Vk - \delta \langle n \rangle .$$

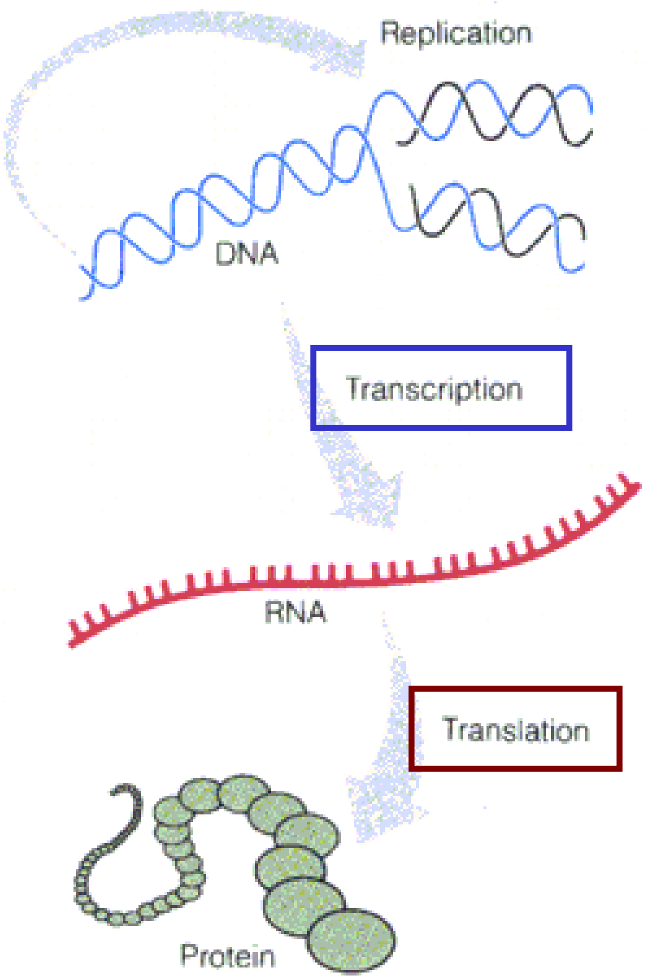
And thus

$$\boxed{\begin{aligned} r_k &= Vk \\ r_\delta &= \delta \end{aligned}}$$

pseudofirst-order reaction

first order reaction

A more detail model of gene expression



Deterministic model

$$\begin{aligned} \frac{d[mRNA]}{dt} &= k_R - \delta_R[mRNA] \\ \frac{d[P]}{dt} &= k_P[mRNA] - \delta_P[P] \end{aligned}$$

$$\emptyset \xrightarrow{k_R} mRNA, \qquad \text{P}(k_R \text{ reaction}) = \mathbf{r}_k\ \delta\text{t}$$

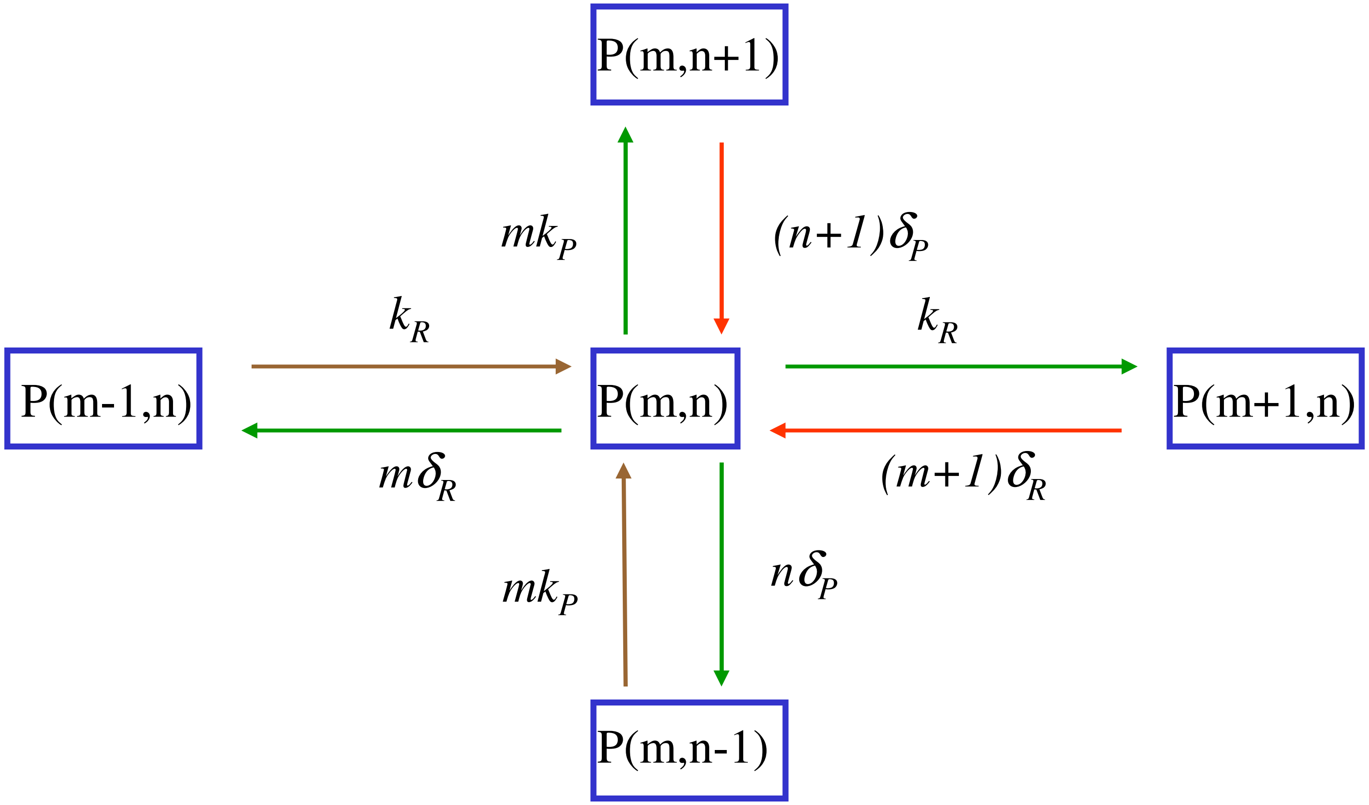
$$mRNA \xrightarrow{\delta_R} \emptyset, \qquad \text{P}(\delta_R \text{ reaction}) = \text{m}_\text{R}\ \mathbf{r}_\delta\ \delta\text{t}$$

$$mRNA \xrightarrow{k_P} P, \qquad \text{P}(k_P \text{ reaction}) = \text{m}_\text{R}\ \mathbf{r}_\kappa\ \delta\text{t}$$

$$P \xrightarrow{\delta_P} \emptyset, \qquad \text{P}(\delta_P \text{ reaction}) = \text{n}_\text{P}\ \mathbf{r}_\delta\ \delta\text{t}$$

How does the probability of having, say, m mRNA molecules and n *P* molecules, p(m,n), change with time?

r_κ's as k's to simplify notation
(this could also imply that $V = 1$)



$$\frac{dp_{m,n}}{dt} = -p_{m,n}[m\delta_R + mk_P + k_R + n\delta_P]$$ 🟩

$$+ \quad p_{m,n+1}(n+1)\delta_P + p_{m+1,n}(m+1)\delta_R$$ 🟥

$$+ \quad p_{m,n-1}k_P m + p_{m-1,n}k_R$$ 🟫

Equation of the mean; emergence of deterministic laws

note first, a useful equation for a given function *f(n,m)*

$$\begin{aligned}\frac{d\langle f_{n,m}\rangle}{dt} &= -\langle f_{n,m}m\rangle\delta_R - \langle f_{n,m}m\rangle k_P - \langle f_{n,m}\rangle k_R - \langle f_{n,m}n\rangle\delta_P \\ &+ \langle f_{n-1,m}n\rangle\delta_P + \langle f_{n,m-1}m\rangle\delta_R + \langle f_{n+1,m}m\rangle k_P + \langle f_{n,m+1}\rangle k_R\end{aligned}$$

thus, we get

$$\frac{d\langle m\rangle}{dt} = k_R - \delta_R\langle m\rangle \longrightarrow$$

this is the equation the very same equation we obtained for the simple model, i.e., it implies steady state Poisson statistics for mRNA

$$\frac{d\langle n\rangle}{dt} = k_P\langle m\rangle - \delta_P\langle n\rangle \longrightarrow$$

what kind of protein macroscopic steady state statistic characterizes protein dynamics?

we make use of the following equations ...

$$\begin{aligned}\frac{d\langle n^2\rangle}{dt} &= -2\langle n^2\rangle\delta_P + \langle n\rangle\delta_P + 2\langle nm\rangle k_P + \langle m\rangle k_P \\ \frac{d\langle nm\rangle}{dt} &= -\langle nm\rangle(\delta_P + \delta_R) + \langle m^2\rangle k_P + \langle n\rangle k_R\end{aligned}$$

... to get the final expressions for the macroscopic statistics

Fano Protein = $\dfrac{\langle n^2\rangle - \langle n\rangle^2}{\langle n\rangle} = 1 + \dfrac{k_P/\delta_R}{1+\delta_P/\delta_R} \approx \boxed{1 + \dfrac{k_P}{\delta_R}}$ translation efficiency influences noise

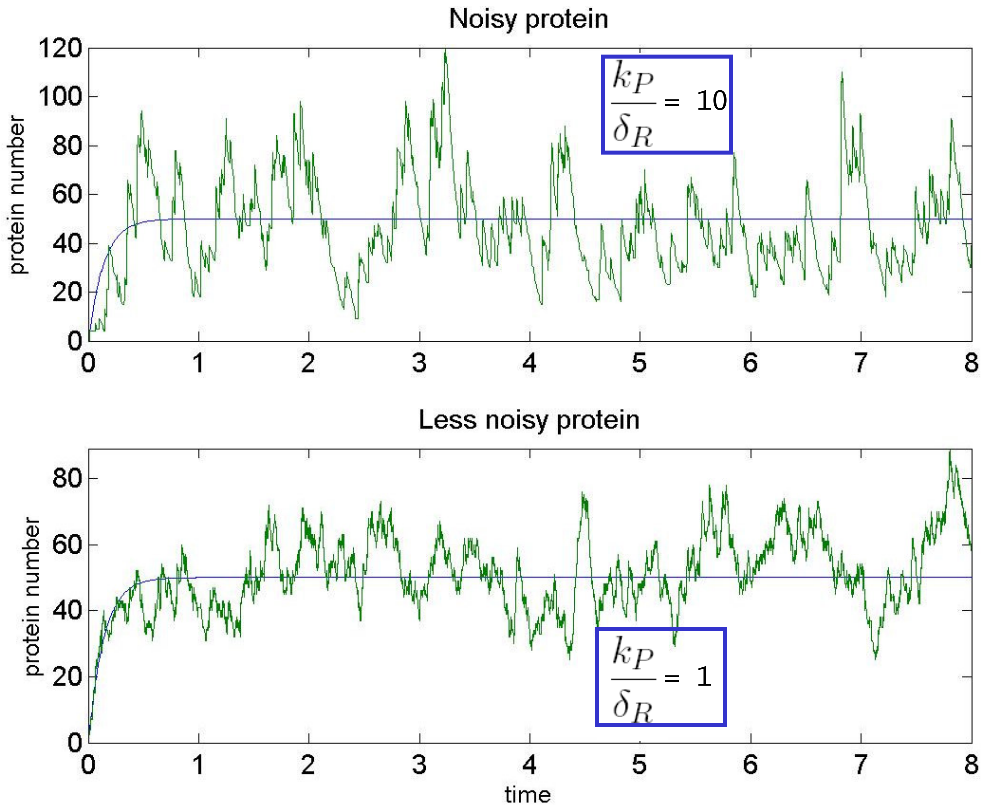
Fano mRNA = $\boxed{1}$

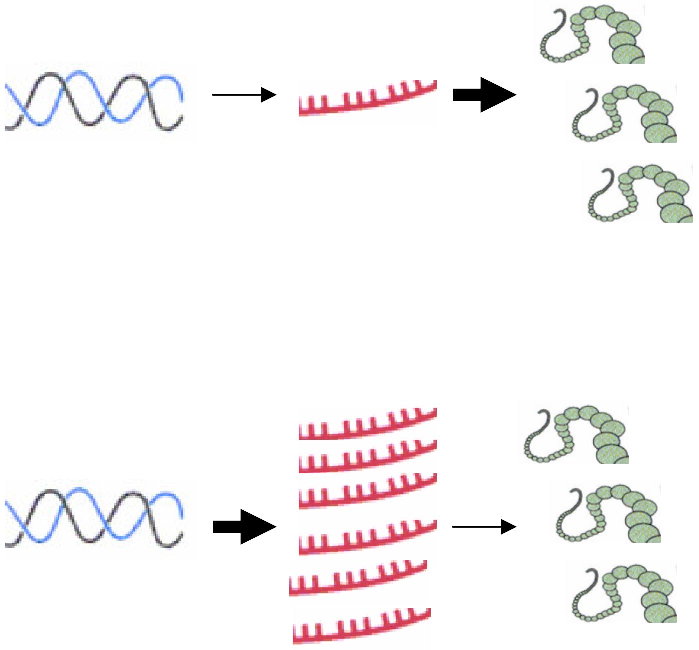
transcription efficiency does not influence noise

protein half-lifetime ~ hours
mRNA half-lifetime ~ minutes
thus

$t_{1/2} = \log 2/\delta$ and $\delta_P \ll \delta_R$

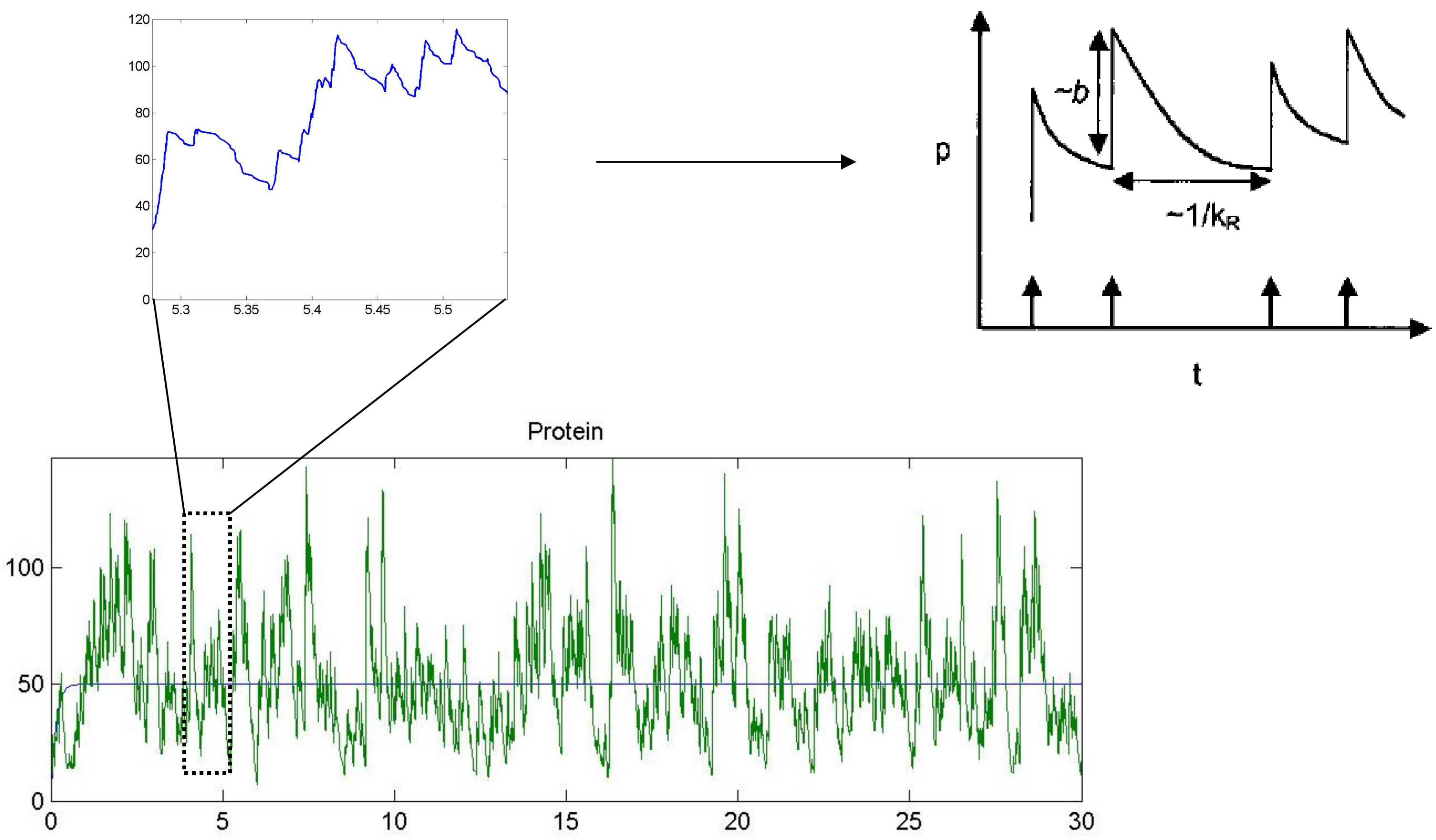
Independently adjust average and dispersion of protein





“Random bursts model”

$$b = \frac{k_P}{\delta_R}$$



Master equations and gene expression

- Genes are generally regulated by complex nonlinear functions. Analitical studies become difficult.

- Two types of approximation methods

1) Numerical Simulation → Gillespie's algorithm

2) Perturbation Methods → Langevin equations, ...

Simulating Stochastic Reactions

Two key questions: When will the next reaction occur?
What kind of reaction will it be?

$P(\tau,\mu)d\tau$ = probability that, given the state $(X_1,\ldots X_N)$ at time t, the next reaction in V occurs in the infinitesimal time interval $(t + \tau, t + \tau + d\tau)$ and it will be a R_μ reaction.

propensity function, e.g., $n_1 n_2 \mathbf{r}$

$$P(\tau,\mu)d\tau = P_0(\tau)\, a_\mu d\tau,$$

here,

$P_0(\tau)$: the probability that no reaction happens in the time interval $(t, t + \tau)$

$a_\mu d\tau$: the probability that reaction R_μ will happen in the time interval $(t + \tau, t + \tau + d\tau)$

The function $P_0(t)$: (no reaction)

$P_0(t+dt) = P_0(t)(1 – a_0 dt)$ $a_0 = \sum_{j=1,M} a_j$

$(P_0(t+dt) – P_0(t))/dt = -a_0 P_0(t)$

$d P_0 /dt = -a_0 P_0(t)$

$P_0(t) = \exp(-a_0t)$

The reaction probability density function:

$P(\tau,\mu)d\tau = P_0(\tau) a_\mu d\tau = a_\mu \exp(-a_0\tau) d\tau$ $\mu = 1,...,M$ $\tau \in(0,+ \infty)$

It is possible to write $P(\tau,\mu)$ as a product of $P(\tau)$ and $P(\mu)$:

$P(\tau,\mu)d\tau = a_\mu \exp(-a_0\tau) d\tau = (a_\mu / a_0) a_0 \exp(-a_0\tau) d\tau$

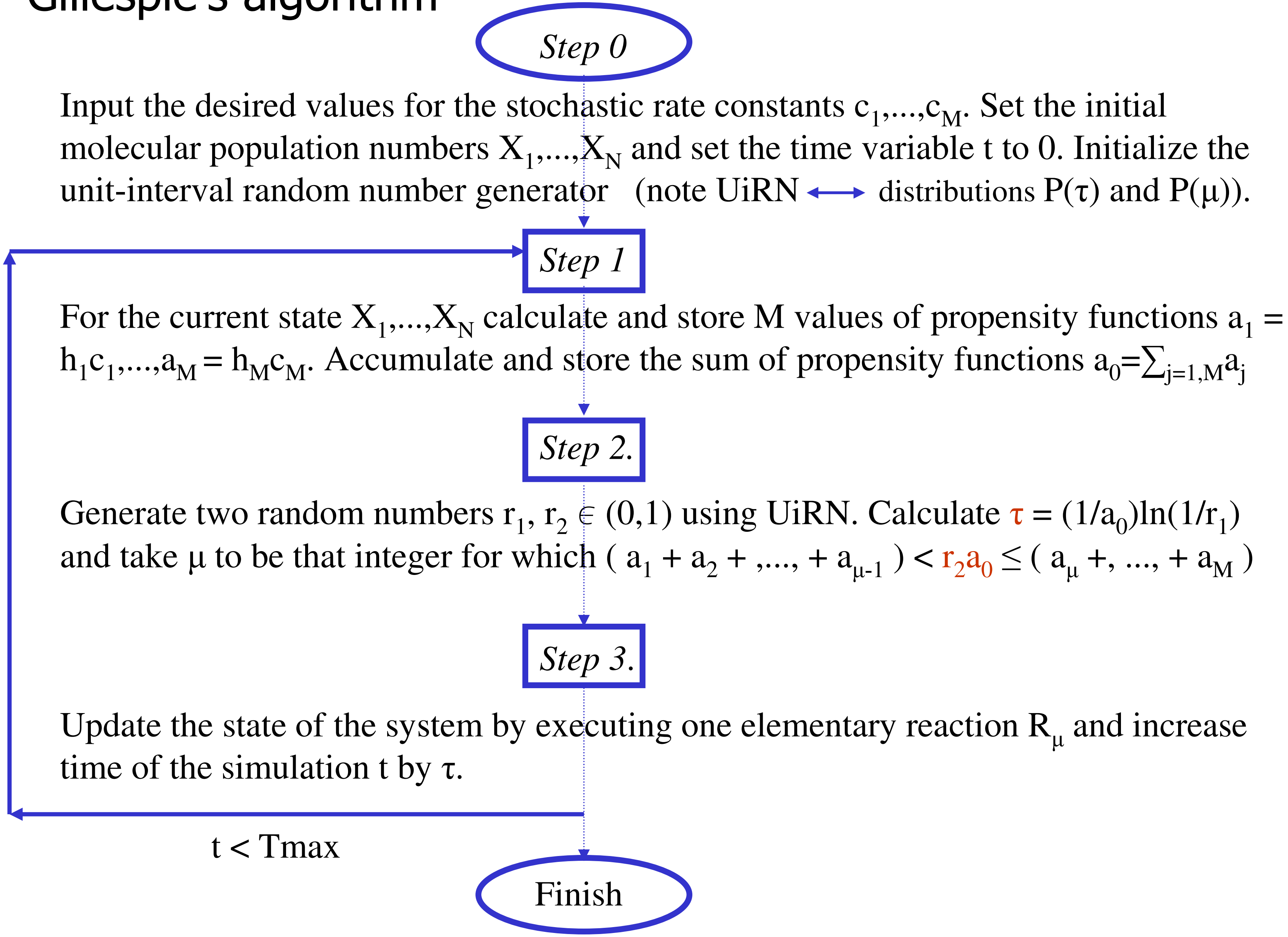
$P(\mu) = (a_\mu / a_0)$

$P(\tau) = a_0 \exp(-a_0\tau) d\tau$

Therefore, we may determine the waiting time for the next reaction by generating two random numbers following distributions $P(\tau)$ and $P(\mu)$.

Note that the algorithm is a rigurous consequence of the Fundamental Hypothesis

Gillespie’s algorithm



Step 0

Input the desired values for the stochastic rate constants $c_1,\ldots,c_M$. Set the initial molecular population numbers $X_1,\ldots,X_N$ and set the time variable t to 0. Initialize the unit-interval random number generator (note UiRN ⟷ distributions $P(\tau)$ and $P(\mu)$).

Step 1

For the current state $X_1,\ldots,X_N$ calculate and store M values of propensity functions $a_1 = h_1c_1,\ldots,a_M = h_Mc_M$. Accumulate and store the sum of propensity functions $a_0=\sum_{j=1,M}a_j$

Step 2.

Generate two random numbers r_1, $r_2 \in (0,1)$ using UiRN. Calculate $\tau = (1/a_0)\ln(1/r_1)$ and take μ to be that integer for which $(a_1 + a_2 + ,\ldots, + a_{\mu-1}) < r_2a_0 \leq (a_\mu +, \ldots, + a_M)$

Step 3.

Update the state of the system by executing one elementary reaction R_μ and increase time of the simulation t by τ.

$t <$ Tmax

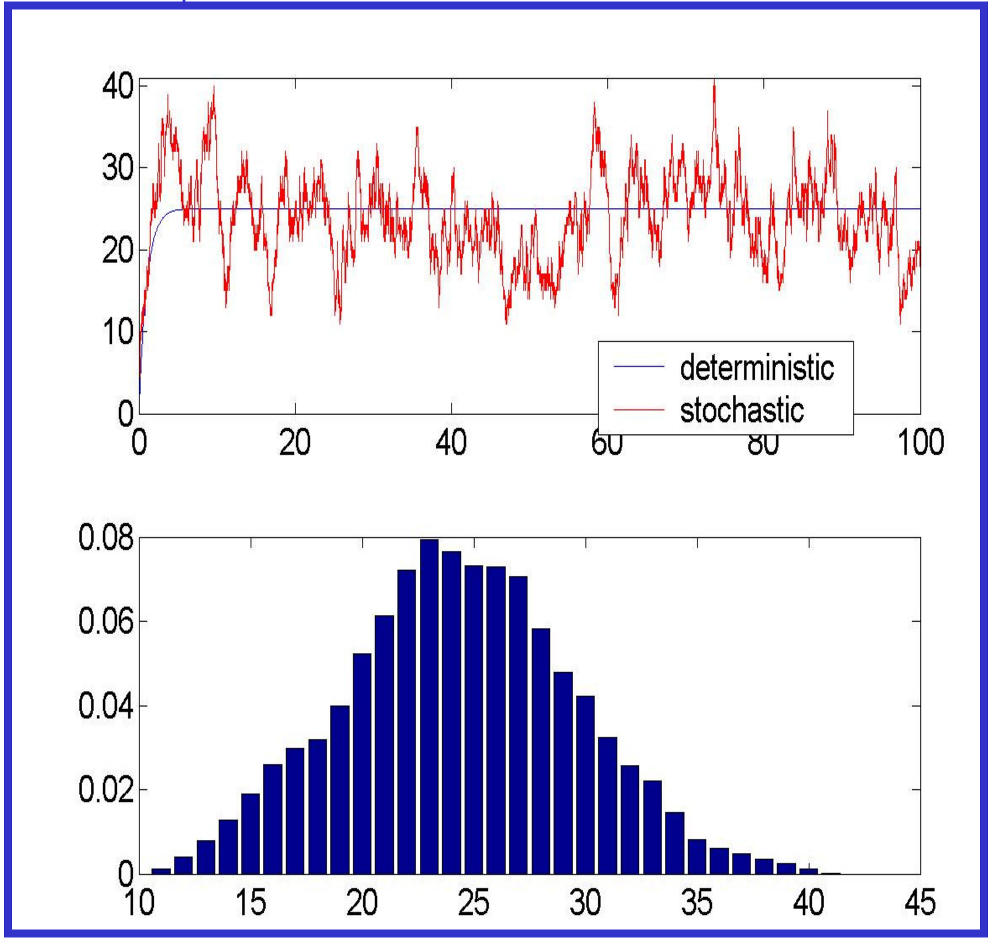
Finish

MATLAB code 2

```
% .. code1stoch.m
% .. simple gene expression stochastic and deterministic

clear all
k = 25;
delta = 1;

% .. stochastic eqs. Gillespie's algorithm
P = 0;
Pstochastic = P;
tmax = 10;
t = 0;
tspan = t;
```

```
while t < tmax

    % .. a's
    a = [k, delta*P(1)];
    a0 = sum(a);
    % .. determine time of next reaction
    r1 = rand;
    tau = -log(r1)/a0;
    t = t + tau;
    % .. determine nature of next reaction
    r2 = rand;
    acumsum = cumsum(a)/a0;
    chosen_reaction = min(find(r2 <= acumsum));

    if chosen_reaction == 1;
        P(1) = P(1) + 1;
    else
        P(1) = P(1) - 1;
    end

    tspan = [tspan,t];
    Pstochastic = [Pstochastic;P];

end
```

```
% .. deterministic eqs.
P0 = 0;
options = [];
[t P] = ode23(@code1equations,tspan,P0,options,k,delta);

% .. plot
subplot(211)
plot(t,P,t,Pstochastic,'r')
legend('deterministic','stochastic')
axis([0 tmax 0 max(Pstochastic)]);

subplot(212)
Pst = Pstochastic;
hh = histc(Pst,min(Pst):max(Pst));
bar(min(Pst):max(Pst),hh/sum(hh))
```



```
meanhist = 24.3909

varihist = 26.7338

fano = 1.0961
```

Langevin equations

- Developed originally to the study of Brownian motion

- Alternative mathematical framework to that of the Master Equation

- Better suited for an intermediary ("not very noisy") regime

- Based on adding explicitely noise terms to the deterministic (macroscopic) equations

- Need to characterize the noise distribution of the added noise

$$\begin{aligned}\frac{d[mRNA]}{dt} &= k_R - \delta_R[mRNA] + \underline{\xi_R} \\ \frac{d[P]}{dt} &= k_P[mRNA] - \delta_P[P] + \underline{\xi_P}\end{aligned}$$

$\xi_R,\ \xi_P$ added stochastic variables.

These equations are fully specified when the probability distributions for the stochastic variables are given.

Valid to describe an intermediate situation where fluctuations are important even though the number of particles is big enough.

what are the properties of ξ ?

- we would like to know mean, variance, characteristic fluctuation times, ...

Fluctuation time:
we can ask how much correlates the variation of ξ with respect to its mean:
autocorrelation function

$$\begin{aligned} C_\xi(t_1,t_2) &= \Big\langle \Big[\xi(t_1) - \langle \xi(t_1)\rangle\Big]\Big[\xi(t_2) - \langle \xi(t_2)\rangle\Big]\Big\rangle \\ &= \langle \xi(t_1)\xi(t_2)\rangle - \langle \xi(t_1)\rangle\langle \xi(t_2)\rangle \end{aligned}$$

at equal times (t_1=t_2) we recover the variance.

Often there exists a characteristic time τ_c for which $\underline{C_\xi(t_1,t_2)=0.}$

τ_c is known as the autocorrelation time

White noise

Langevin originally applied to brownian motion: no reason why thermal fluctuations should favour a particular reaction:

$\xi_R \quad \xi_P$ defined such that $\langle \xi(t) \rangle \quad = \quad 0$

thus, $C_\xi(t_1, t_2) \quad = \quad \langle \xi(t_1))\xi(t_2)) \rangle$

collision time is faster than time-scale of change of molecule numbers, noise is uncorrelated

$$C_\xi(t_1, t_2) \quad = \quad \Gamma e^{\frac{-(t_1 - t_2)}{\tau}}$$

noise strength, variance at equal times

$\tau \quad \rightarrow 0$

very small autocorrelation times: $\boxed{C_\xi(t_1, t_2) \quad = \quad \Gamma \delta(t_1 - t_2)}$

White noise:

- noise variable with *zero* autocorrelation time

- white?

$$S(\omega) \quad = \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega\tau}\langle\xi(t))\xi(t+\tau)\rangle d\tau = 1$$

all frequencies contribute equally

Color noise

- noise variable with *finite* autocorrelation time

Stochastic differential equations are very irregular, modified numerical methods to solve them: one to the simplest Euler-Maruyana method