Stochastic Gene Expression in Prokaryotes: A Point Process Approach

Emanuele LEONCINI

INRIA Rocquencourt - INRA Jouy-en-Josas

Mathematical Modeling in Cell Biology

March 27th 2013

Central role of protein production

- Proteins are the core of biologic processes: *enzymes*, DNA replication machinery, ...
- $\sim$ 50% of the bacteria dry weight
- $\sim$ 3.5 millions of proteins in each cell
- $\sim$ 2000 types of proteins produced at any time at any growth condition (volume growth)
- proteins ranging from few dozens up to 10^5

Central role of protein production

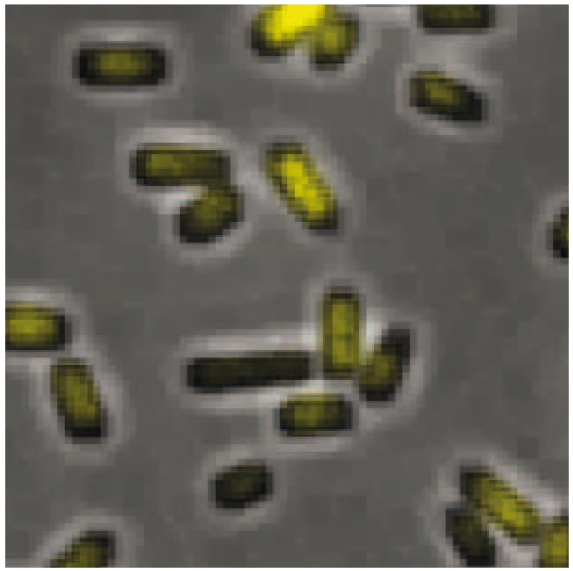
- Proteins are the core of biologic processes: *enzymes*, DNA replication machinery, ...
- $\sim$ 50% of the bacteria dry weight
- $\sim$ 3.5 millions of proteins in each cell
- $\sim$ 2000 types of proteins produced at any time at any growth condition (volume growth)
- proteins ranging from few dozens up to 10^5

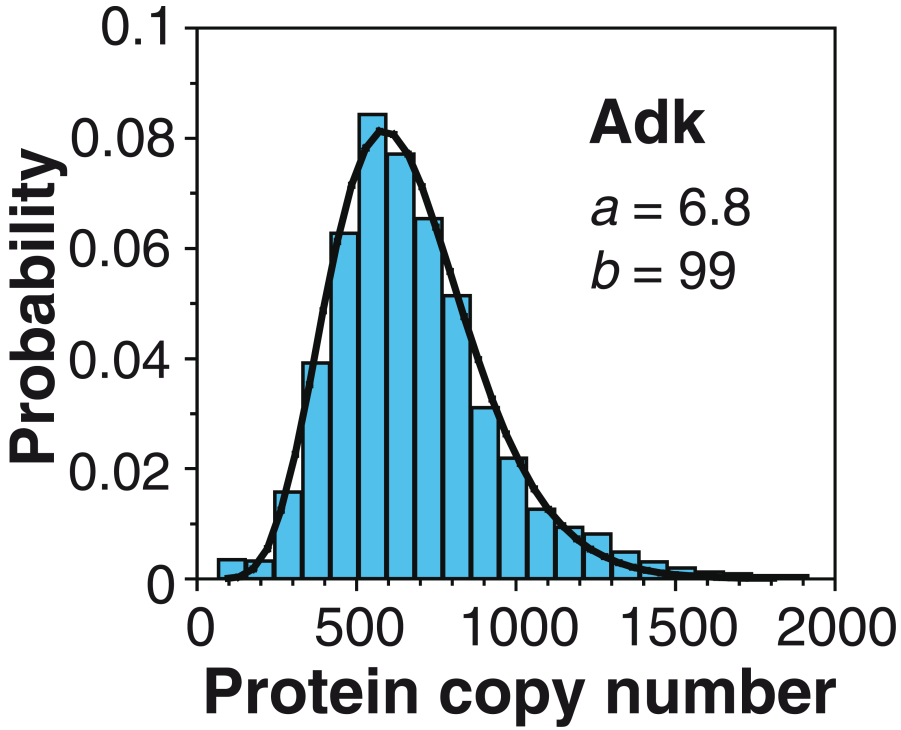
A highly consuming process:

- at each generation, bacteria has to duplicate all proteins
- more than 85% of cell resources

Stochasticity in protein production: experimental viewpoint

ADk cytoplasm protein[1]





[1] Yuichi Taniguchi et al. *Science* (2010), pp. 533–538.

Stochasticity in bacteria

Sources of stochasticity:

- bacterial *cytoplasm*: disordered medium
- main cellular motility mechanism: diffusion in a stiff medium
- most cellular processes require the encounter of macromolecules (*Poisson* process)

Stochasticity in bacteria

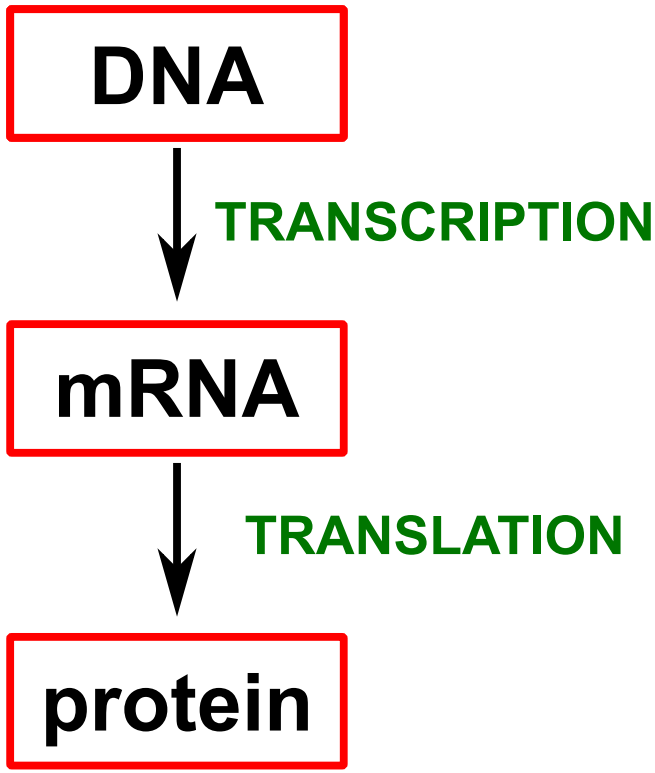
Sources of stochasticity:

- bacterial *cytoplasm*: disordered medium
- main cellular motility mechanism: diffusion in a stiff medium
- most cellular processes require the encounter of macromolecules (*Poisson* process)

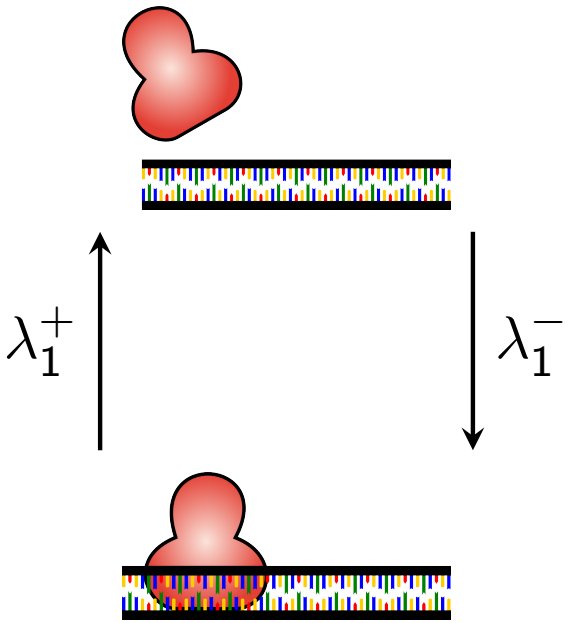
Protein production: inherently stochastic process.

Model

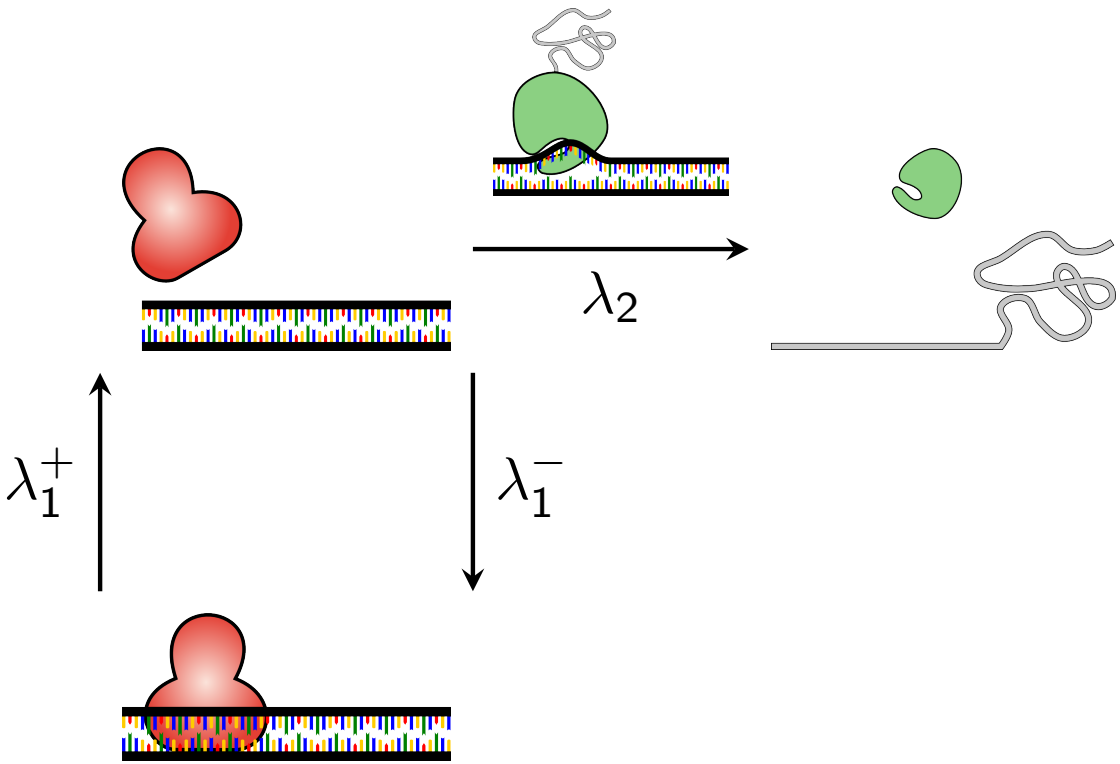
Central Dogma of molecular biology



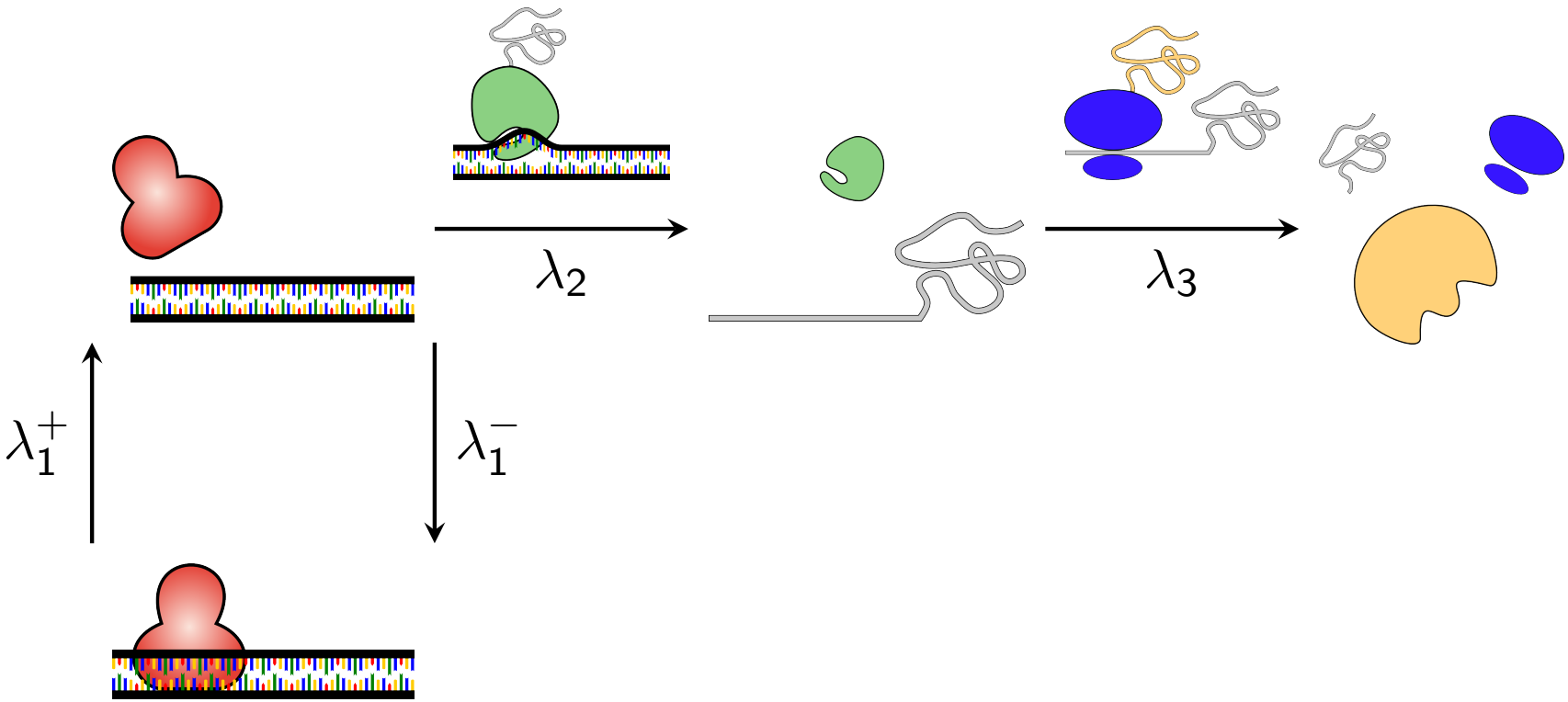
3-stage model: activation



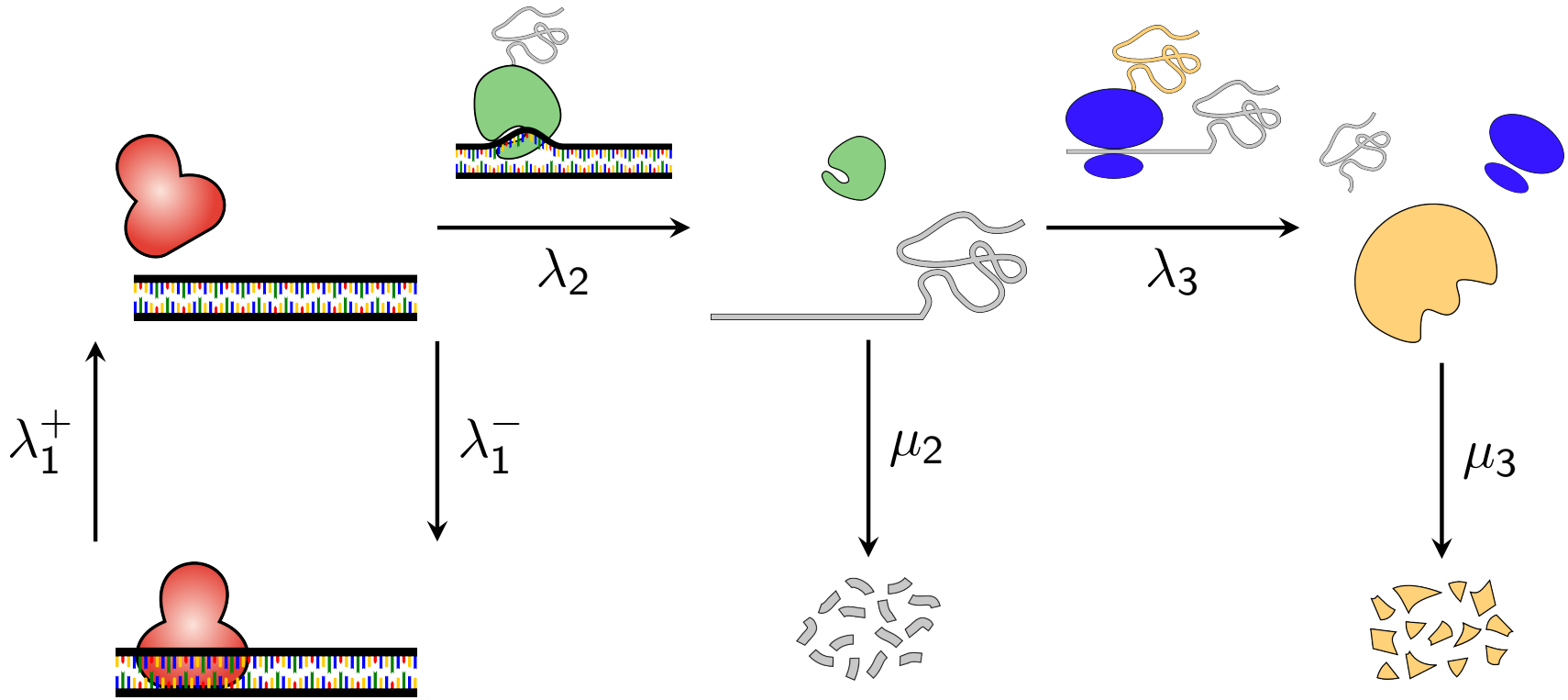
3-stage model: transcription



3-stage model: translation

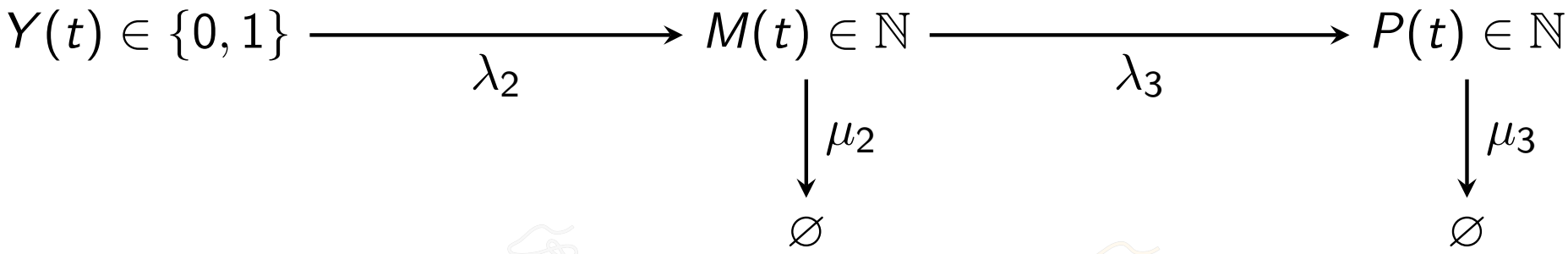


3-stage model[2]



[2] Johan Paulsson. *Physics of life reviews* (2005), pp. 157–175.

3-stage model



Goal: Characterize mean and variance of the number of proteins P at equilibrium

Classic models: Paulsson's survey[3]

Properties

- Assumption: each step has *exponentially distributed* duration
- *Markovian description* of the protein production

[3] Johan Paulsson. *Physics of life reviews* (2005), pp. 157–175.

Classic models: Paulsson's survey[3]

Properties

- Assumption: each step has *exponentially distributed* duration
- *Markovian description* of the protein production

Tools

- Markov processes
- Fokker-Plank equations

explicit analytic formulas of mean and variance as function of the main parameters

[3] Johan Paulsson. *Physics of life reviews* (2005), pp. 157–175.

Classic models: Paulsson's survey[3]

Properties

- Assumption: each step has *exponentially distributed* duration
- *Markovian description* of the protein production

Tools

- Markov processes
- Fokker-Plank equations

→ explicit analytic formulas of mean and variance as function of the main parameters

→ biologists classically use models because of the explicit characterization of protein fluctuations

[3] Johan Paulsson. *Physics of life reviews* (2005), pp. 157–175.

Classic approach:

Exponential assumption:

Not each described process has an exponentially distributed duration.

Classic approach:

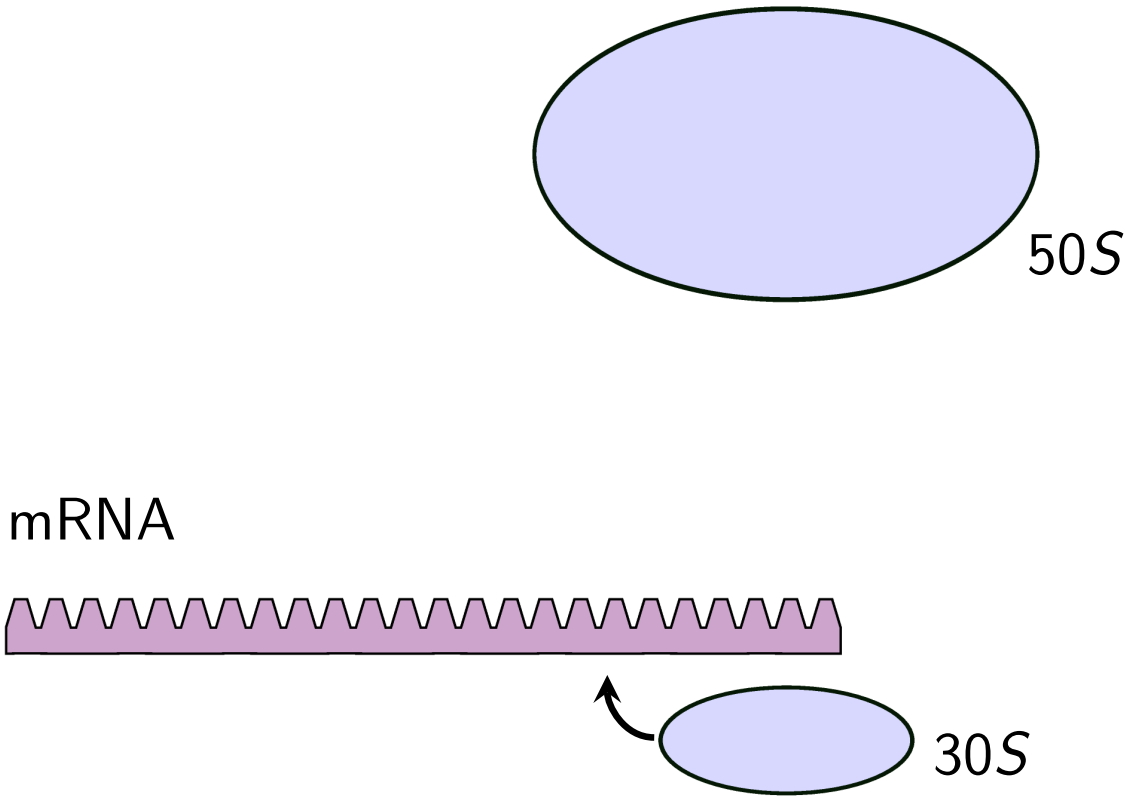
Exponential assumption:

Not each described process has an exponentially distributed duration.

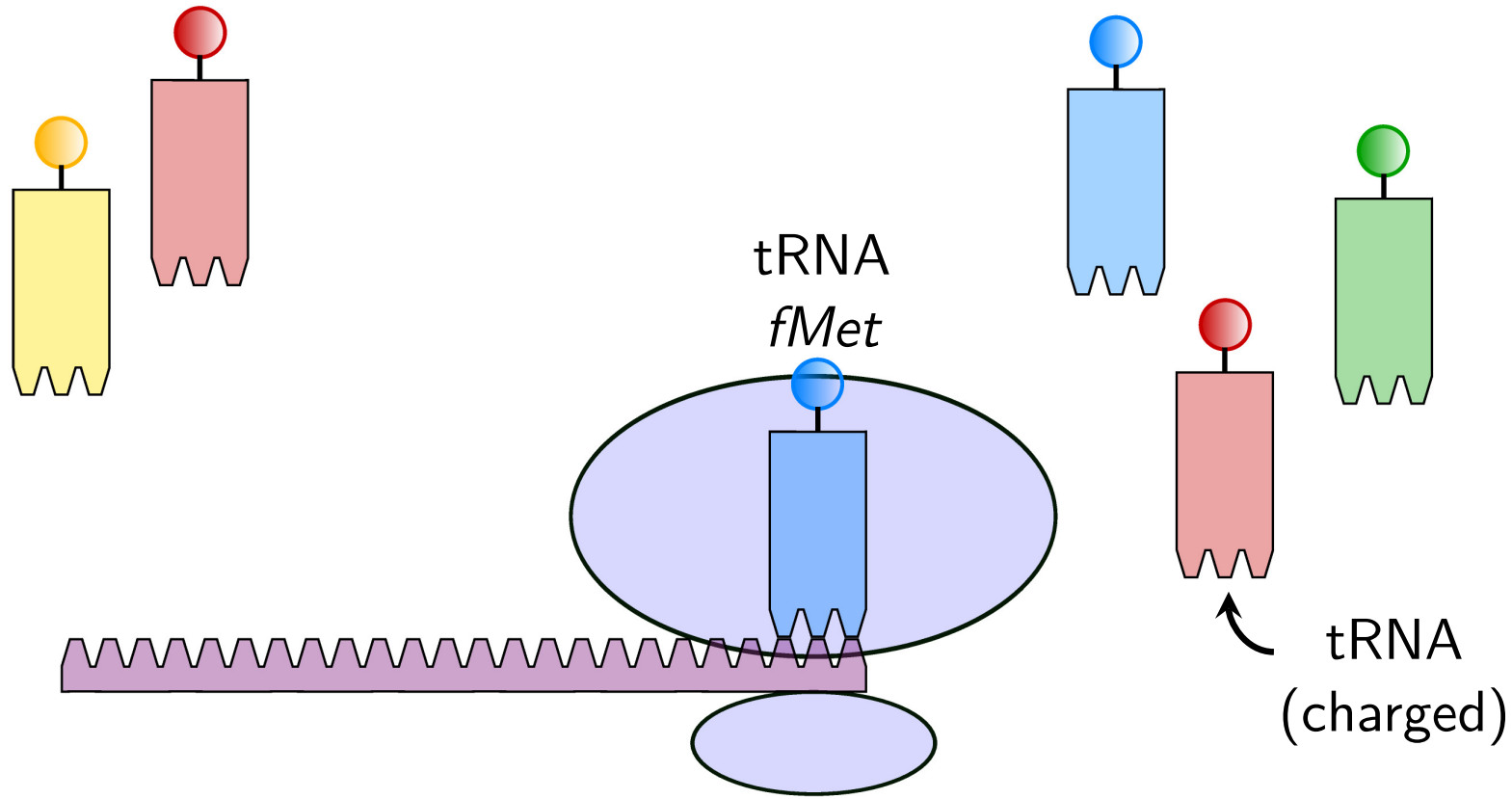
The duration of the following processes is not exponential

- *protein elongation*
- *mRNA elongation*
- *protein degradation*

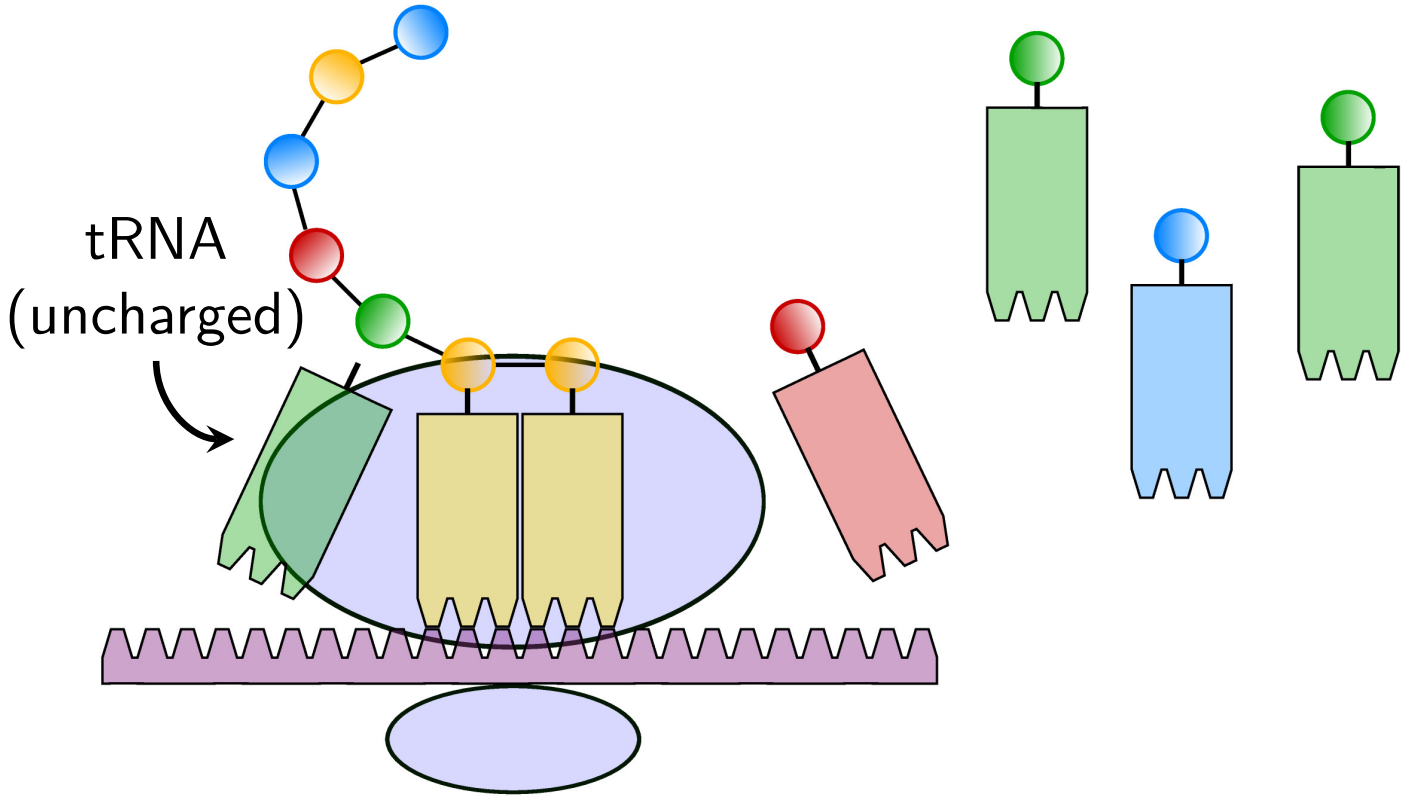
Protein chain elongation



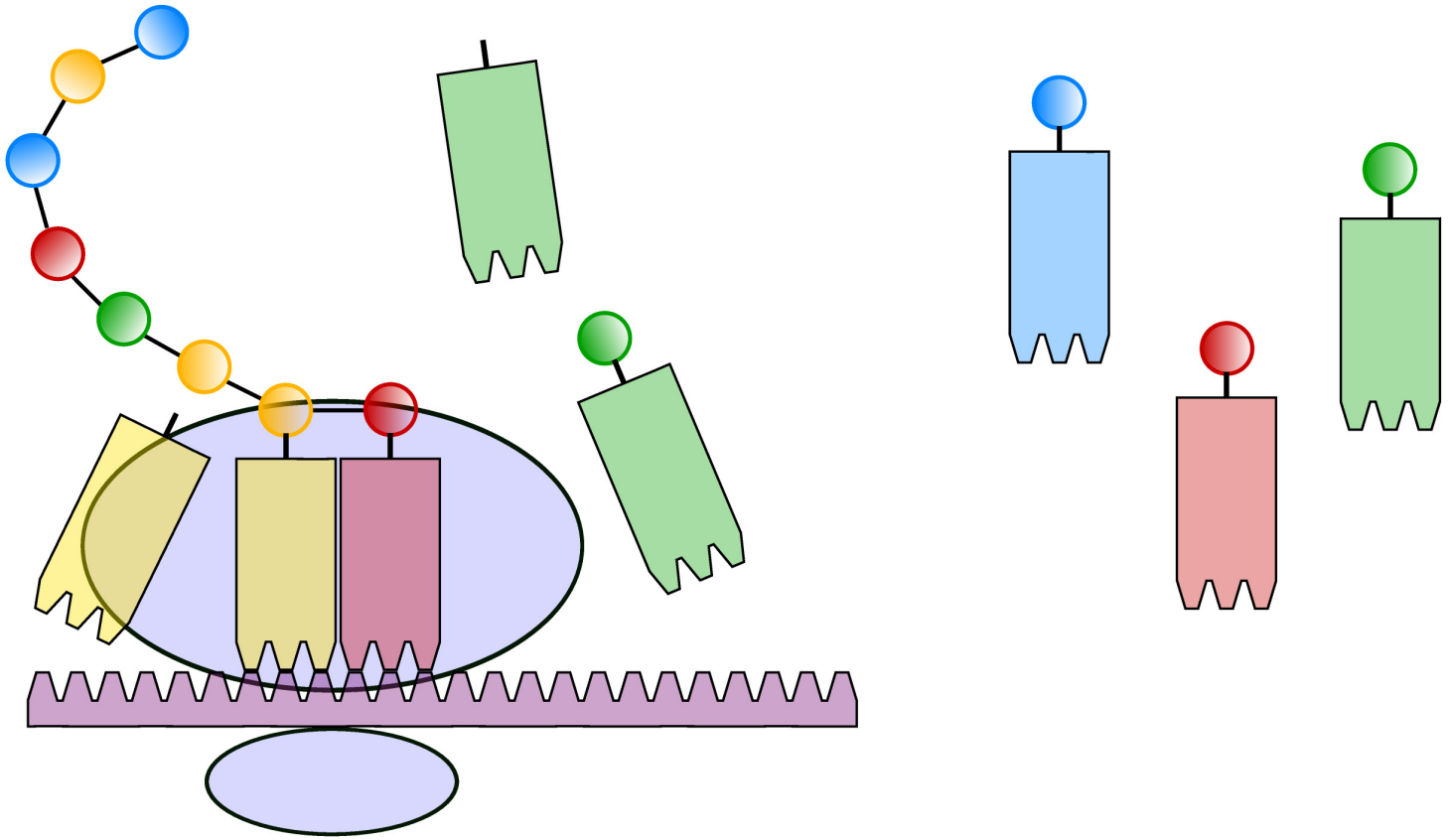
Protein chain elongation



Protein chain elongation

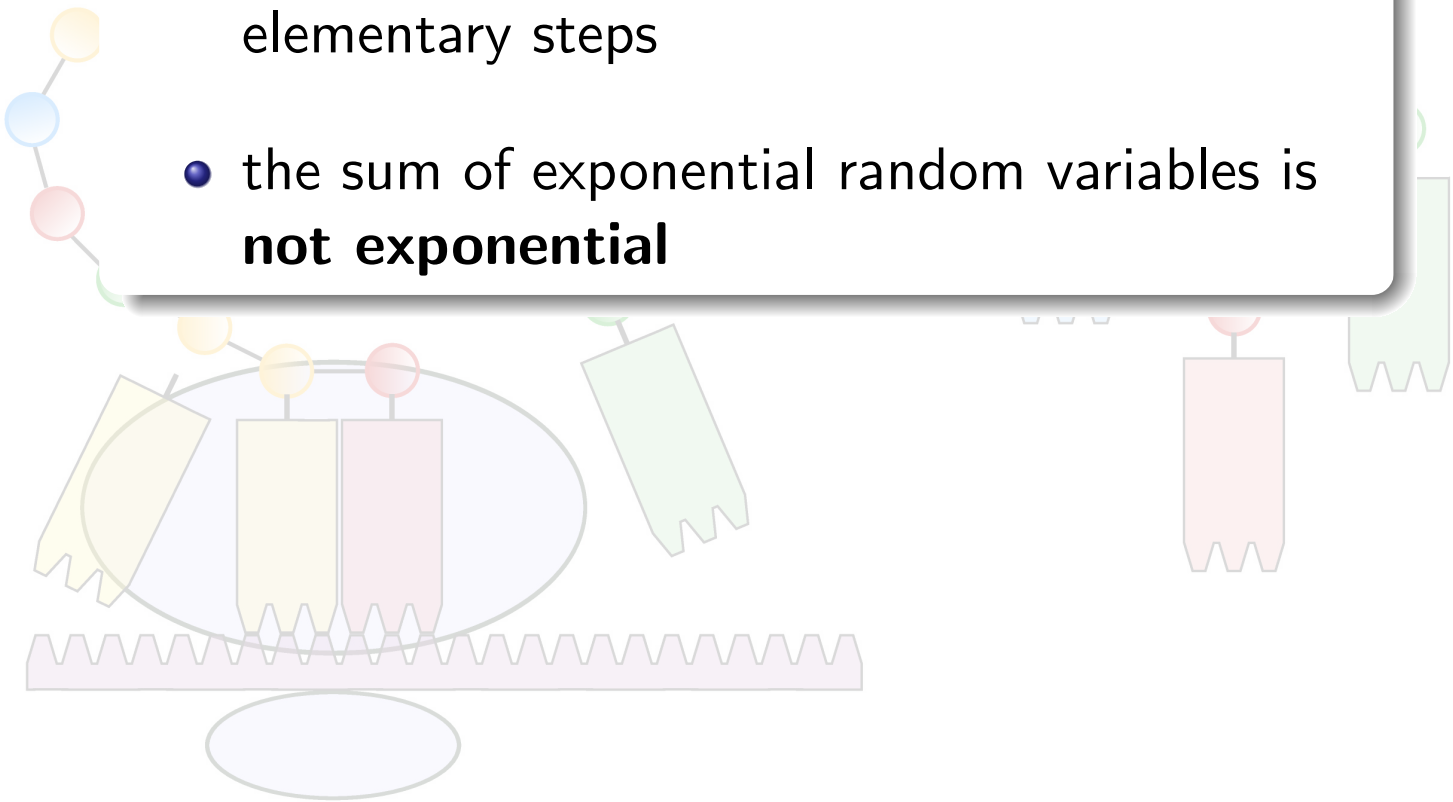


Protein chain elongation



Protein chain elongation

- **Elongation:** exponentially distributed elementary steps
- the sum of exponential random variables is **not exponential**



Protein chain elongation

- **Elongation:** exponentially distributed elementary steps
- the sum of exponential random variables is **not exponential**

Large number of identical steps ($N \approx 400$ a.a.)

➡ elongation time described as normal random variable

Classic approach: a not completely satisfying description

The duration of the following processes is not exponential

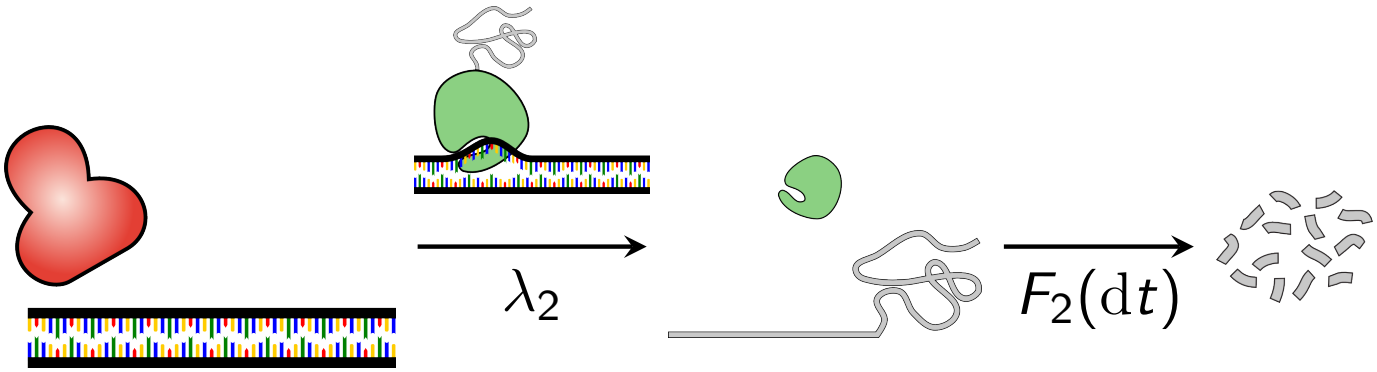
- *protein elongation*
- *mRNA elongation*
- *protein degradation*

New description

Gene expression with non-exponential steps requires a new approach.

Marked Poisson Point Process (MPPP): new description of gene expression

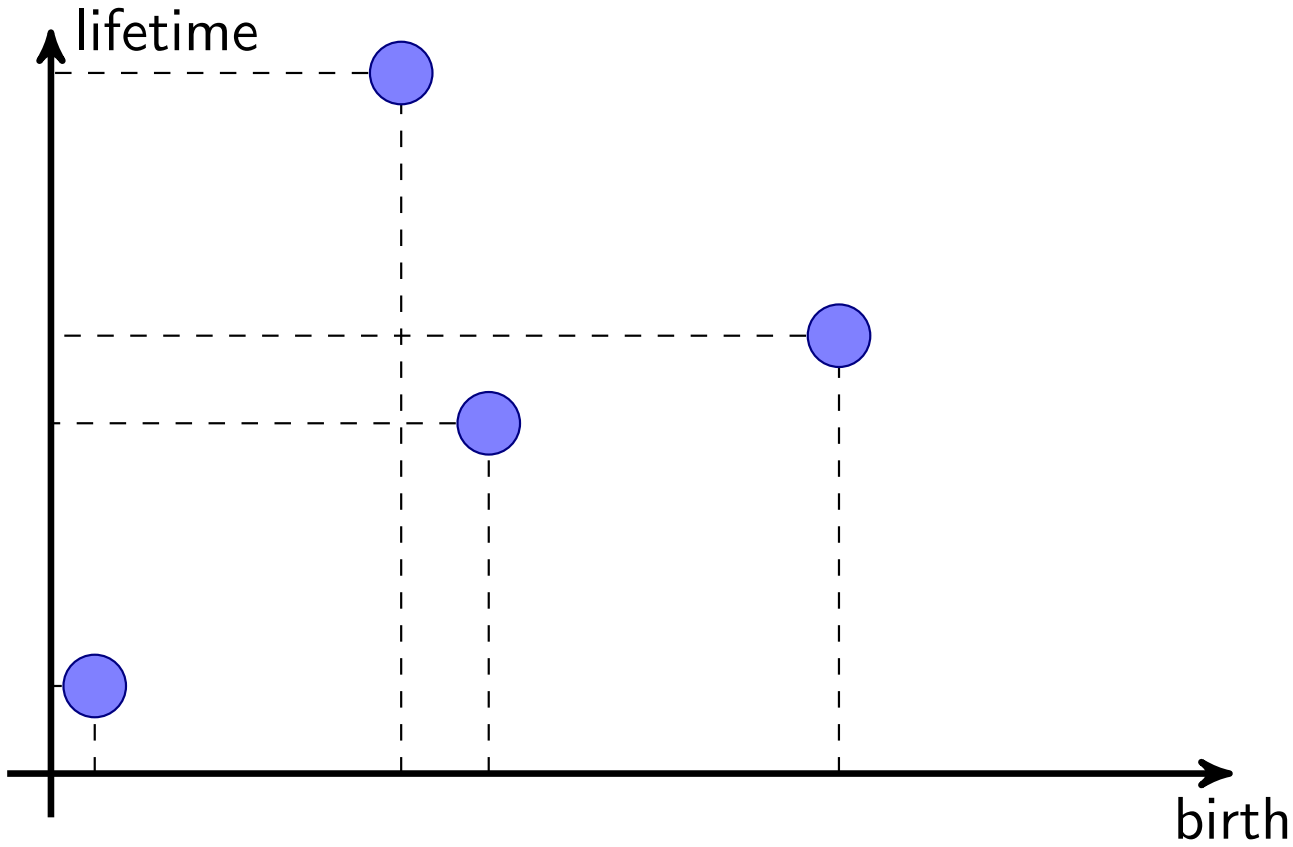
General distributions: introducing MPPP



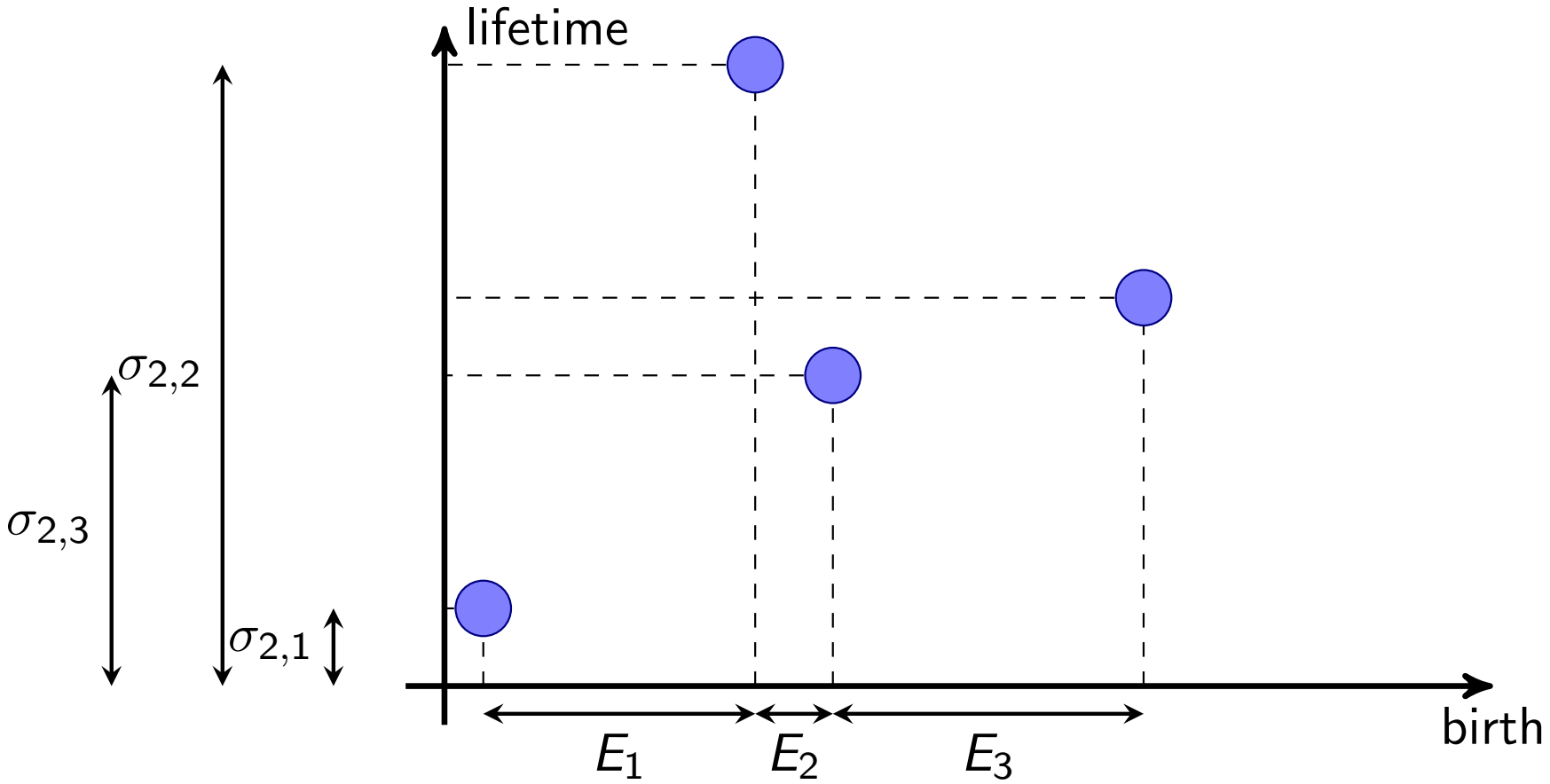
Assumptions:

- mRNA births (s_n) follow a Poisson process of parameter λ_2
- mRNA lifetimes $(\sigma_{2,n})$ have distribution $F_2(\mathrm{d}t)$

mRNA



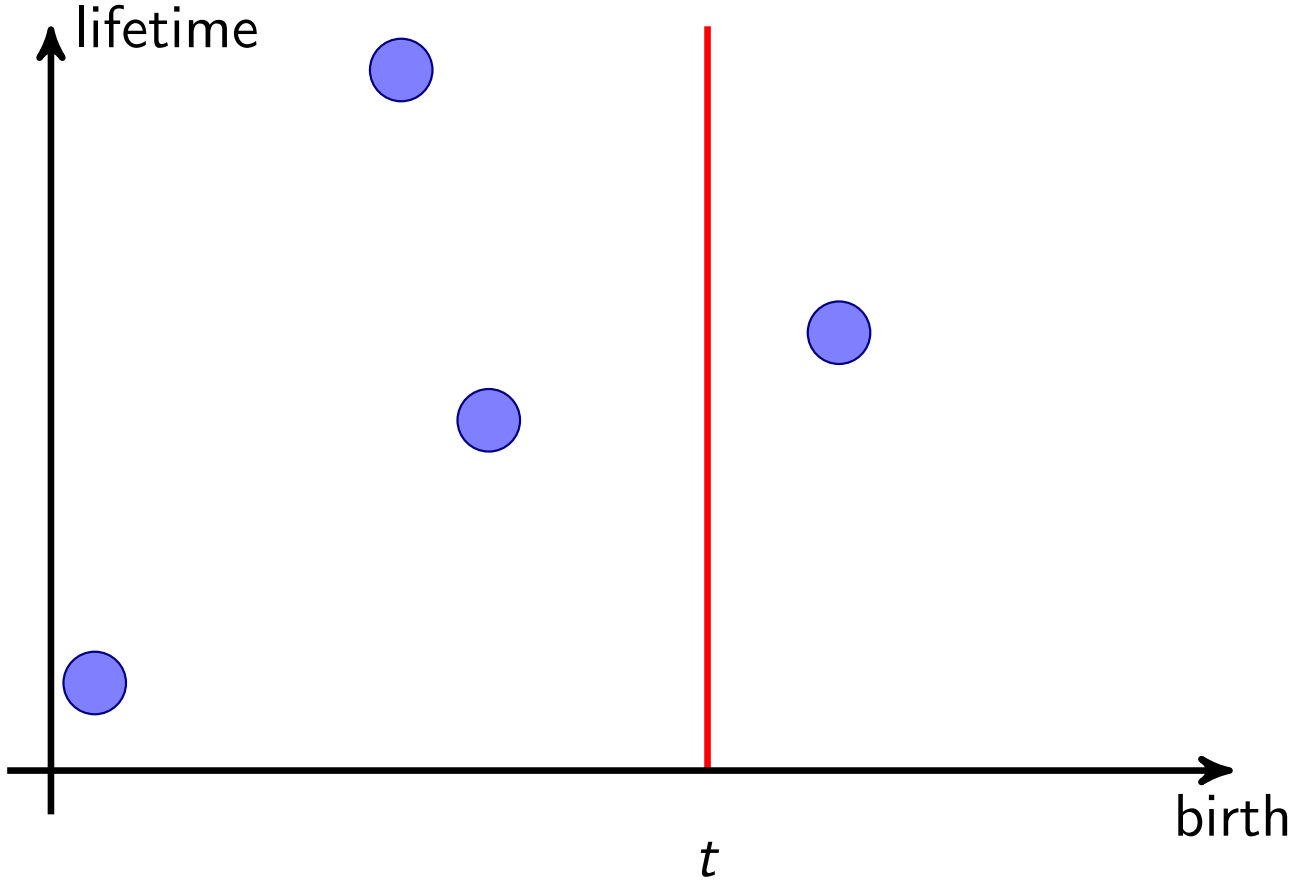
mRNA



E_i exponential random variables of parameter λ_2.

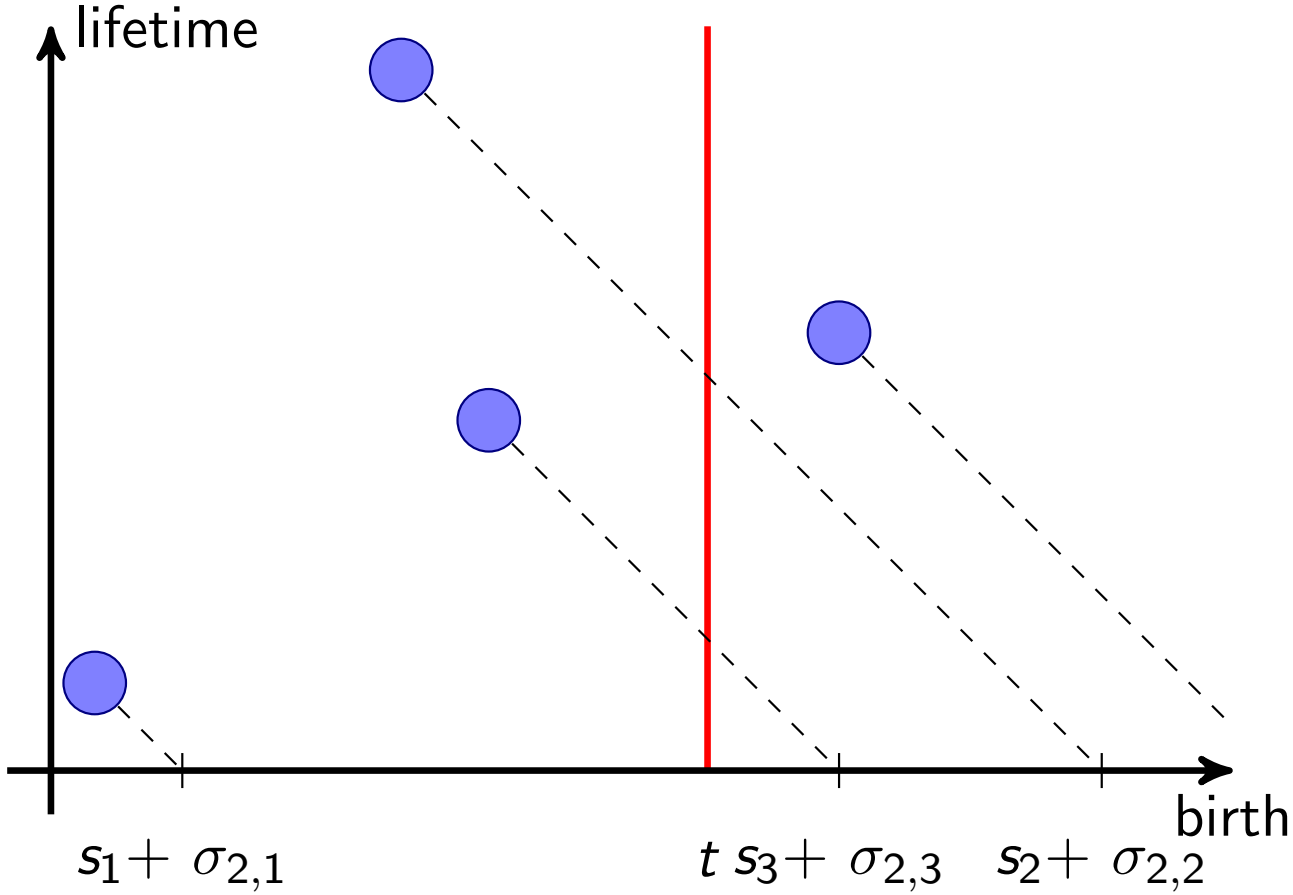
mRNA

How many mRNAs at time t ?



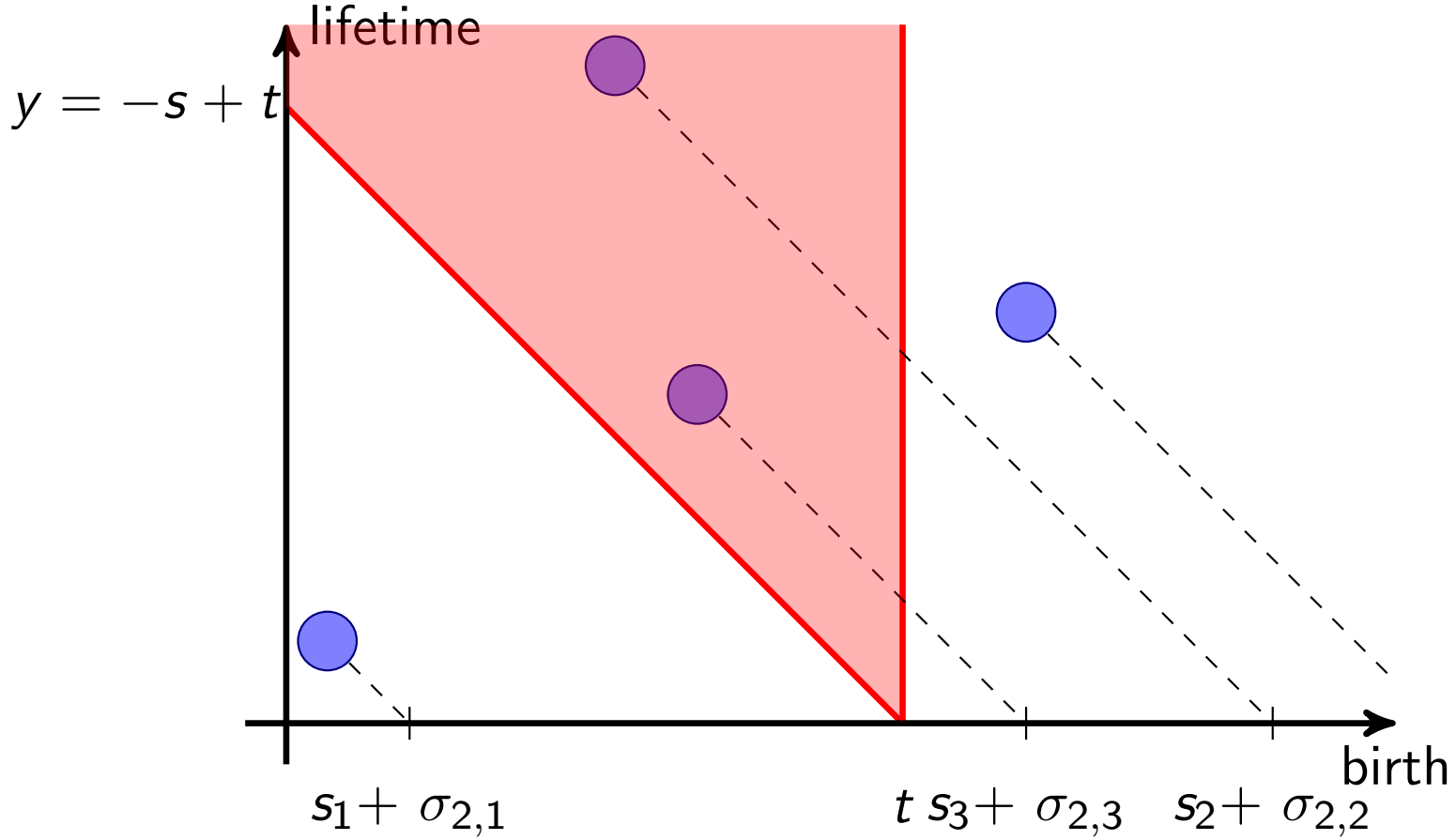
mRNA

How many mRNAs at time t ?



mRNA

How many mRNAs at time t ?



mRNA: general results

mRNAs at equilibrium:

$$M = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\leq 0\leq u+v\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v)$$

mRNA: general results

mRNAs at equilibrium:

$$M = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\leq 0\leq u+v\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v)$$

Proposition

$$\mathbb{E}[M] = \delta_+ \lambda_2 \mathbb{E}[\sigma_2]$$

$$\mathrm{var}(M) = \mathbb{E}[M] + 2\lambda_2^2 \delta_+ (1-\delta_+) \cdot \cdot \int_0^{+\infty} \int_{-u}^{0} e^{-\Lambda v} (1 - F_2(u))(1 - F_2(u+v)) \,\mathrm{d}u \,\mathrm{d}v$$

where $F_2(x) = F_2([0,x])$, $\Lambda = \lambda_1^+ + \lambda_1^-$ and $\delta_+ = \lambda_1^+/\Lambda$

Proteins: general results

Proteins at equilibrium:

$$P = \int_{\mathbb{R}\times\mathbb{R}_+} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v) \left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\leq x\leq u+v\}} \mathbb{1}_{\{x\leq 0\leq x+y\}} \mathcal{N}^u_{\lambda_3}(\mathrm{d}x, \mathrm{d}y) \right]$$

Proteins: general results

Proteins at equilibrium:

$$P = \int_{\mathbb{R}\times\mathbb{R}_+} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v) \left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\le x\le u+v\}} \mathbb{1}_{\{x\le 0\le x+y\}} \mathcal{N}^u_{\lambda_3}(\mathrm{d}x, \mathrm{d}y) \right]$$

Proposition

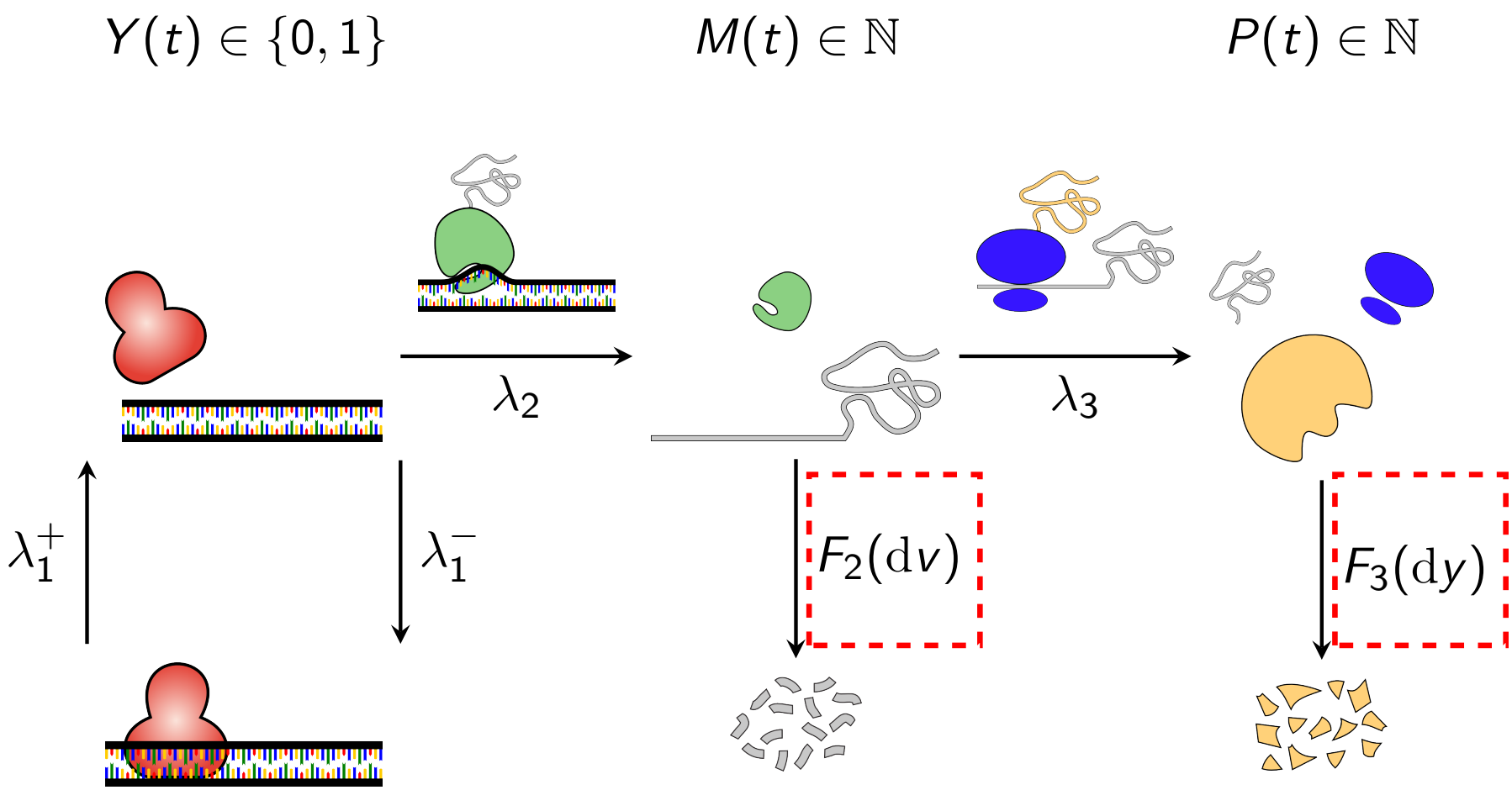
$$\mathbb{E}[P] = \delta_+ \lambda_2 \lambda_3 \mathbb{E}[\sigma_2]\, \mathbb{E}[\sigma_3]$$

$$\mathrm{var}(P) = \mathbb{E}[P] + \lambda_2 \lambda_3^2 \delta_+ \int_{\mathbb{R}_+^2} \left(\int_{-s}^{(-s+t)\wedge 0} F_3(u)\,\mathrm{d}u \right) F_2(t)\,\mathrm{d}t\,\mathrm{d}s$$

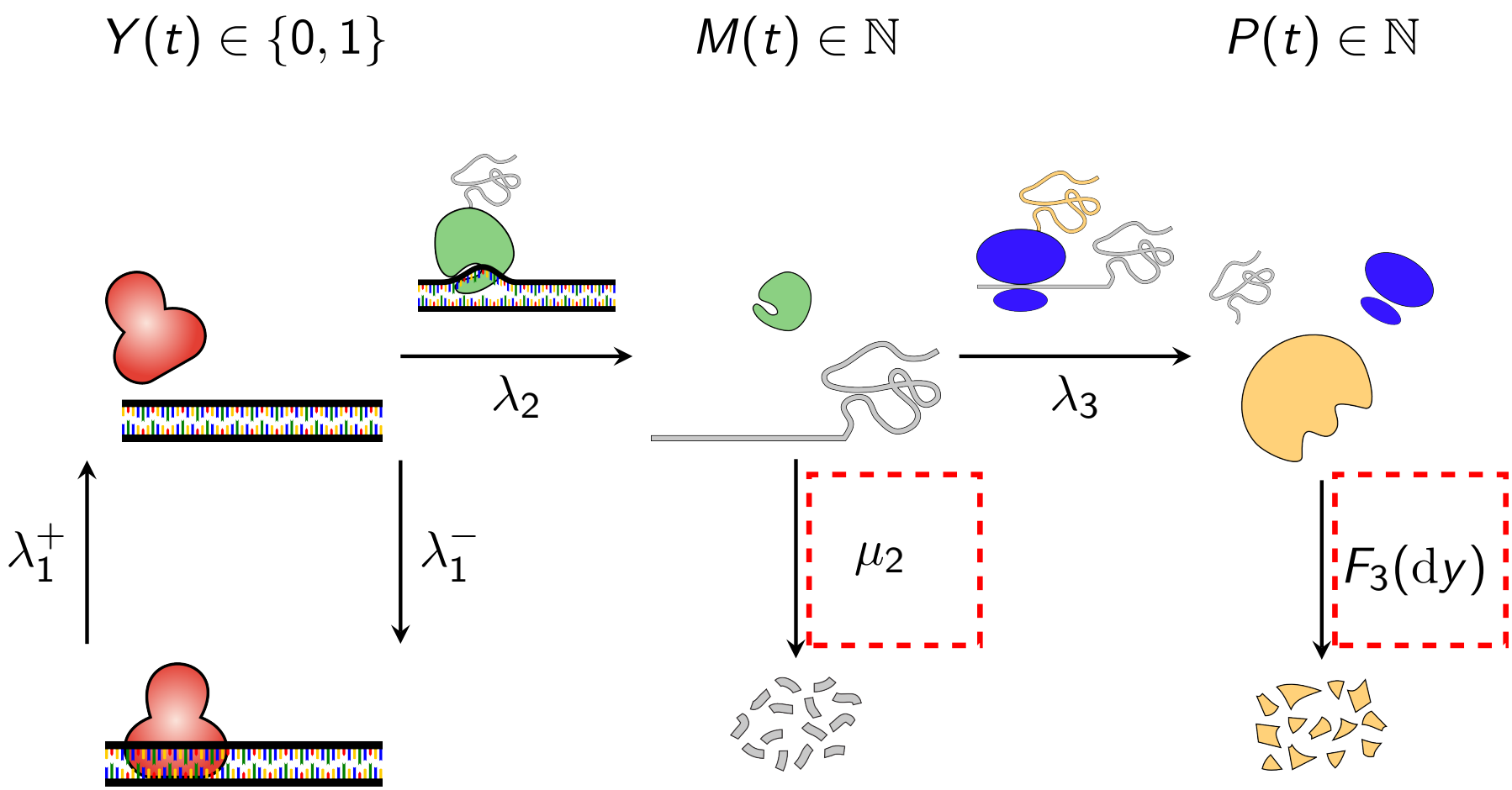
$$+ \lambda_2^2 \lambda_3^2 \delta_+ (1-\delta_+) \int_{R_+^4} e^{-\Lambda|u_1-u_2+v_1-v_2|} \prod_{i=1}^{2} F_2(u_i) F_3(v_i)\,\mathrm{d}u_i\,\mathrm{d}v_i$$

MPPP Applications & Model Extensions

Application: MPPP 3-Stage Model



Application: MPPP 3-Stage Model



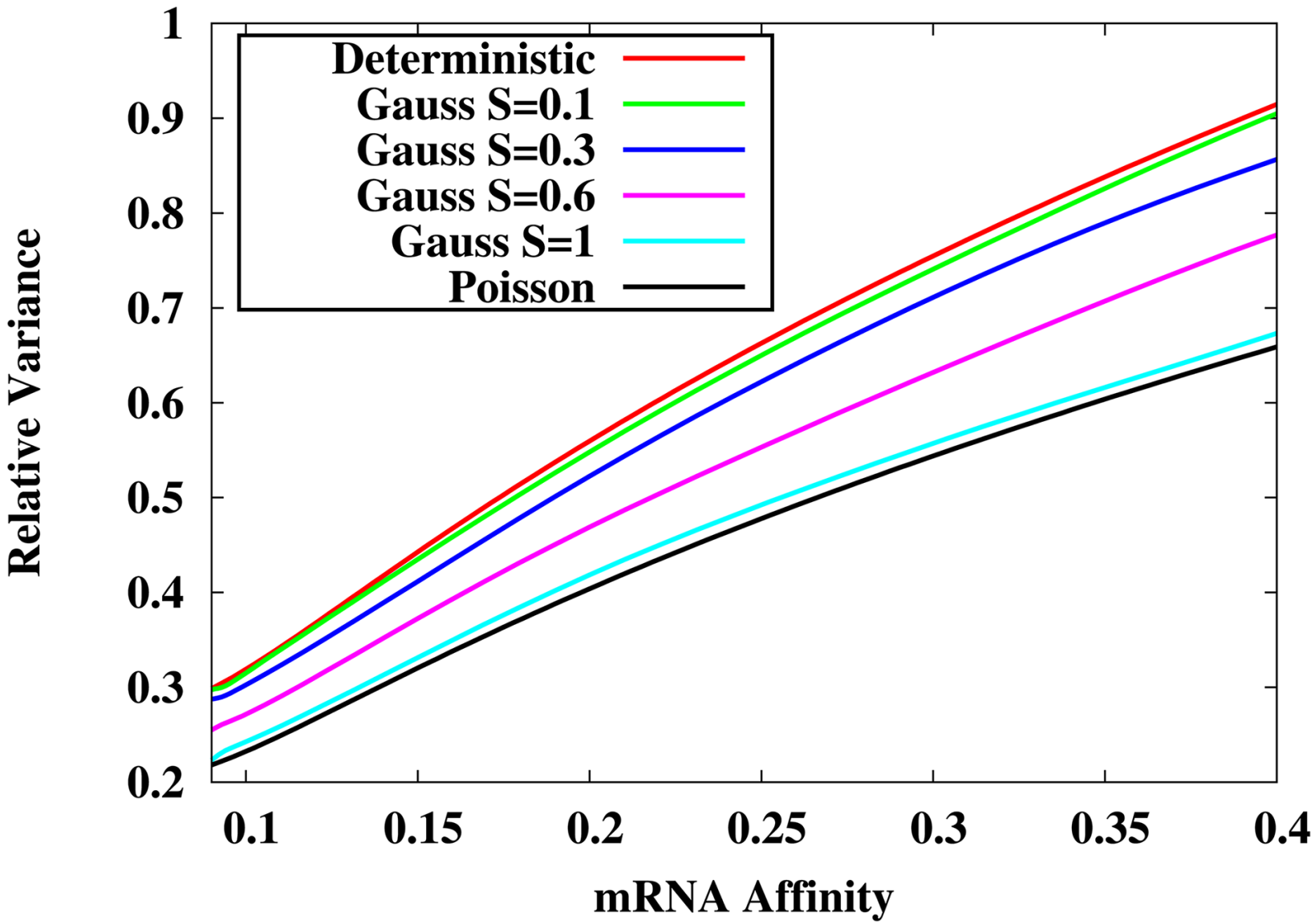
Application: MPPP 3-Stage Model

$Y(t) \in \{0,1\}$ $M(t) \in \mathbb{N}$ $P(t) \in \mathbb{N}$

Choices for $F_3(\mathrm{d}y)$

Exponential	➡	explicit close formula depending on model parameters
Normal	➡	analytic formula
Deterministic	➡	explicit close formula depending on model parameters (limit case)

Deterministic vs Exponential



Application: MPPP 3-Stage Model

3-Stage model with deterministic elongation

$$\begin{aligned}\mathrm{var}_D(P) = \mathbb{E}(P)\Bigg[1 + 2\frac{\lambda_3}{\mu_2}\left(1 - \frac{\mu_3}{\mu_2}(1 - e^{-\mu_2/\mu_3})\right) \\ + \frac{2\lambda_2\lambda_3(1-\delta_+)\mu_2}{\Lambda^2 - \mu_2^2}\left(\frac{\mu_3}{\Lambda^2}\left[1 - e^{-\Lambda/\mu_3}\right]\right. \\ \left. - \frac{\mu_3}{\mu_2^3}\Lambda\left[1 - e^{-\mu_2/\mu_3}\right] + \Lambda\left[\frac{1}{\mu_2^2} - \frac{1}{\Lambda^2}\right]\right)\Bigg]\end{aligned}$$

3-Stage model with exponential elongation

$$\mathrm{var}_E(P) = \mathbb{E}(P)\left(1 + \frac{\lambda_3}{\mu_2 + \mu_3} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda + \mu_2 + \mu_3)}{(\mu_2 + \mu_3)(\Lambda + \mu_2)(\Lambda + \mu_3)}\right)$$

where $\Lambda = \lambda_1^+ + \lambda_1^-$, $\delta_+ = \lambda_1^+/\Lambda$

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression
- averages independent of the chosen distribution

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression
- averages independent of the chosen distribution
- analytic form formula for any distribution
  explicit formula depending on the model parameters for specific and interesting distributions

Conclusions

Biological consequences:

- the computed variance may be underestimated
- deterministic protein elongation might be an upper-bound for protein variance
- possibility to compute protein variance under more realistic assumptions

Few more results

More realistic description of gene expression

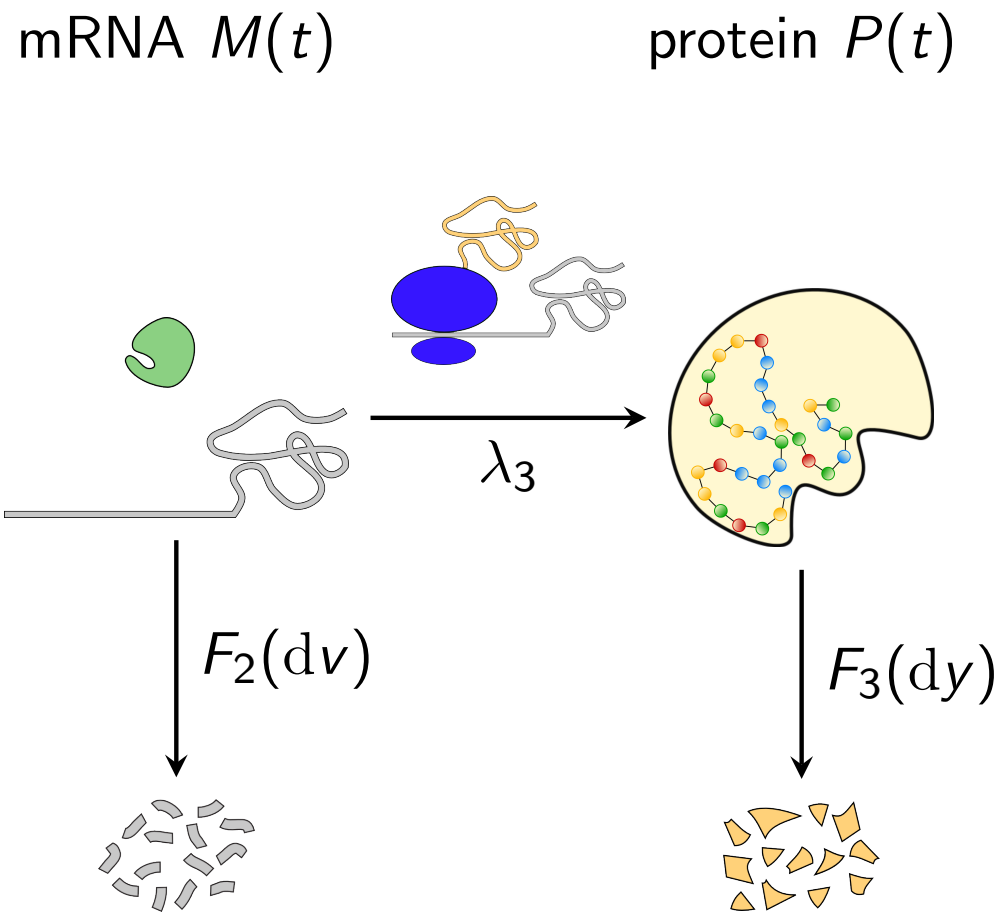
- 4-Stage model and general distribution for protein elongation
- analysis and proof of the correct assumption for protein degradation (*proteolysis*/volume dilution)
- counter-intuitive: $\mathrm{var}_{\mathrm{DET}}(P) > \mathrm{var}_{\mathrm{EXP}}(P)$

V. Fromion, E. Leoncini, and P. Robert. "Stochastic Gene Expression in Cells: A Point Process Approach". In: *SIAM Journal on Applied Mathematics* 73.1 (2013), pp. 195–211

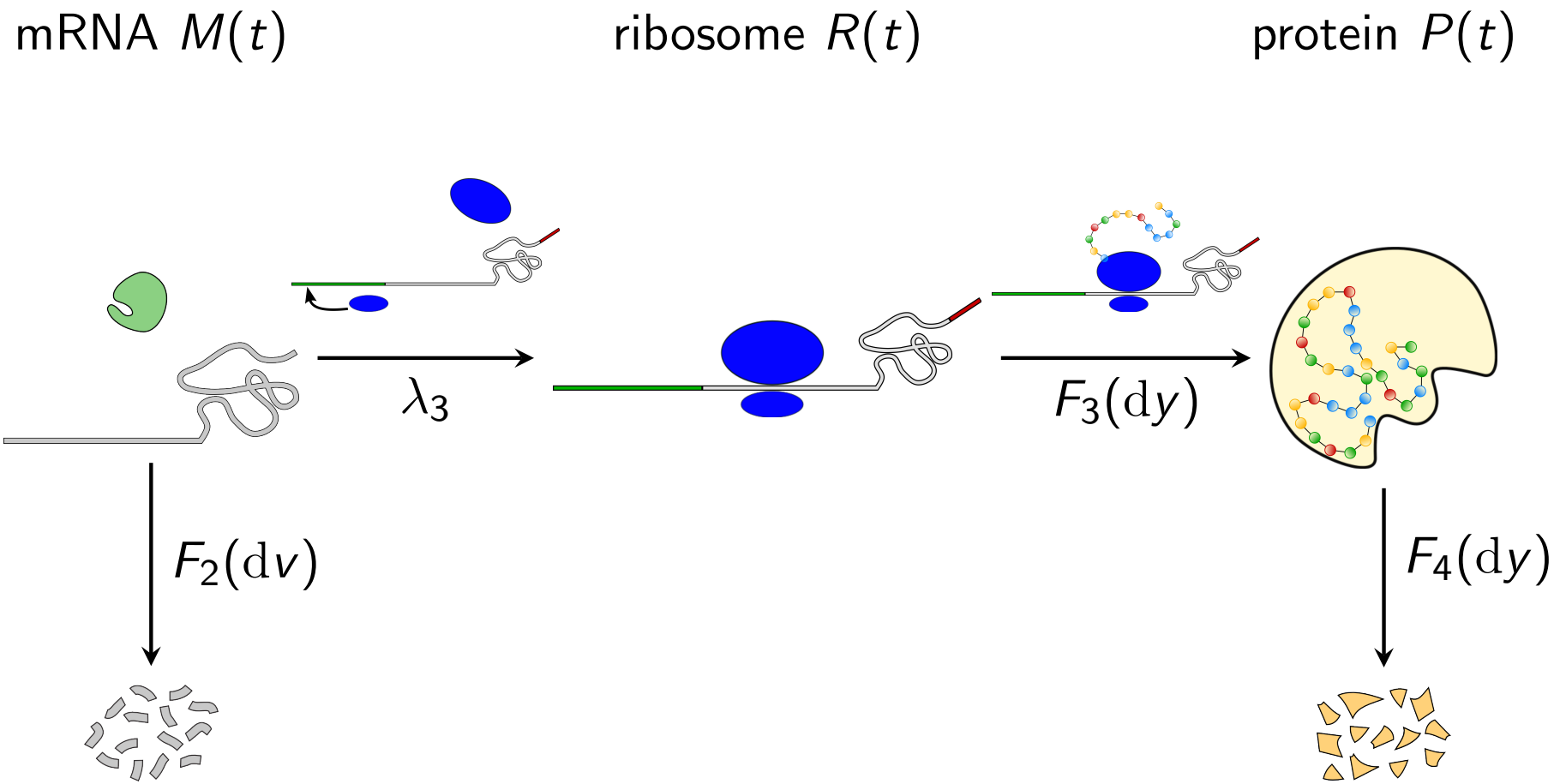
Thanks.

One more thing: realistic model of gene expression

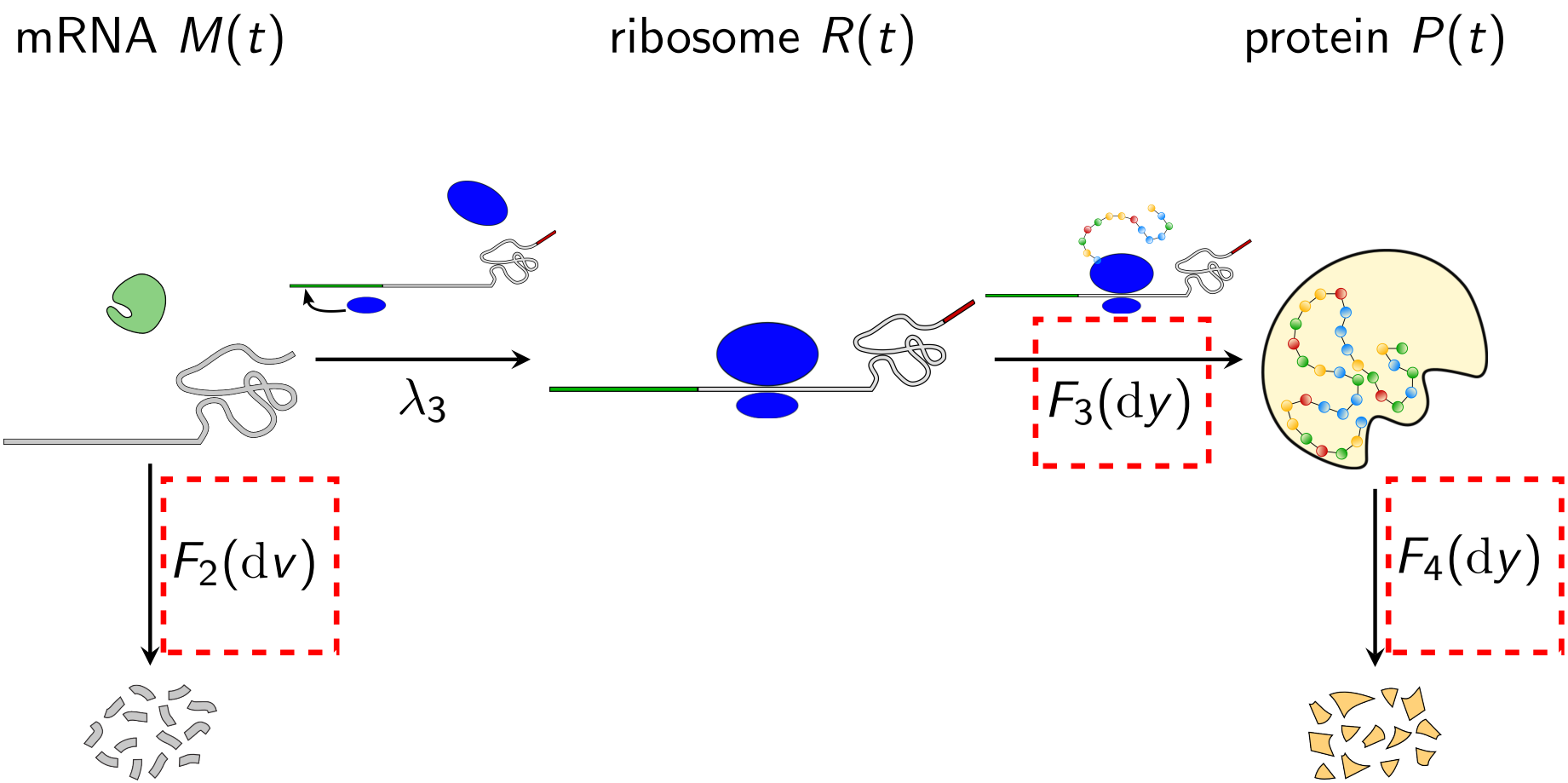
4-Stage Model:



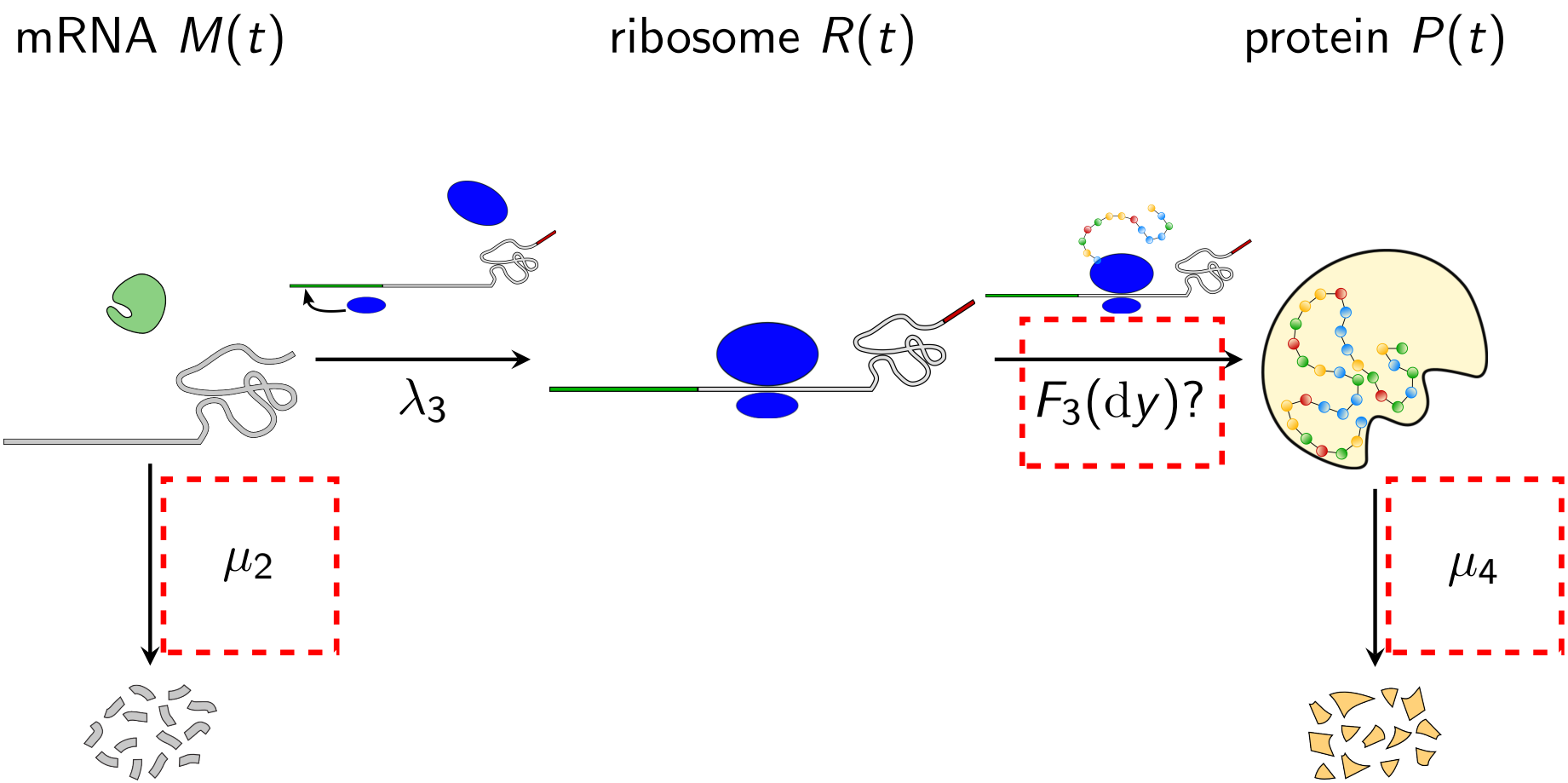
4-Stage Model:



4-Stage Model:



4-Stage Model:



4-Stage Model:

3-Stage model all exponential steps

$$\mathrm{var}^{(3)}(P) = \mathbb{E}[P]\left[1 + \frac{\lambda_3}{\mu_2 + \mu_3}\right] \qquad \text{(active gene)}$$

4-Stage model all exponential steps

$$\mathrm{var}^{(4)}(P) = \mathbb{E}[P]\left[1 + \frac{\lambda_3\mu_3(\mu_2 + \mu_3 + \mu_4)}{(\mu_2 + \mu_3)(\mu_2 + \mu_4)(\mu_3 + \mu_4)}\right] \qquad \text{(active gene)}$$

4-Stage Model: protein elongation

- lots of identical steps ($\sim$ 400 amino acids per protein)
- each step is exponentially distributed

The resulting distribution can be described by

- normal distribution
- Gamma distribution

4-Stage Model: protein elongation

- lots of identical steps ($\sim$ 400 amino acids per protein)
- each step is exponentially distributed

The resulting distribution can be described by

- normal distribution
- Gamma distribution

Problem: hard to obtain explicit formula depending on the model parameters

4-Stage Model: protein elongation

The resulting distribution can be described by

- normal distribution
- Gamma distribution

Problem: hard to obtain explicit formula depending on the model parameters

Approximation: deterministic protein elongation τ_3

4-Stage Model:

4-Stage model with deterministic elongation

$$\mathrm{var}_{\mathrm{D}}(P) \quad = \quad \mathbb{E}(P)\left[1+\frac{\lambda_3}{\mu_2+\mu_4}+\frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\Lambda+\mu_2)(\Lambda+\mu_4)}\right].$$

4-Stage Model:

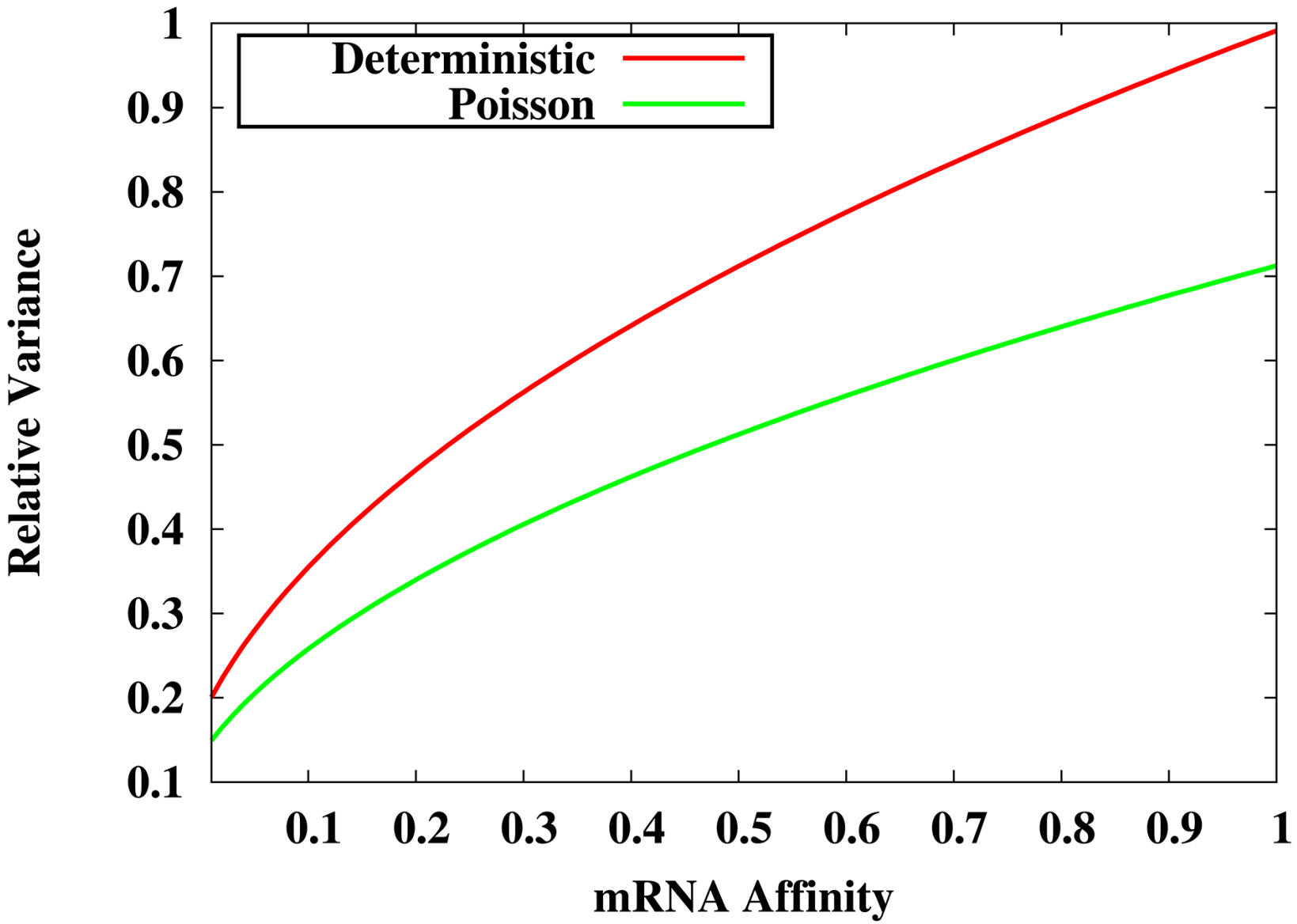
4-Stage model with deterministic elongation

$$\mathrm{var}_{\mathrm{D}}(P) \quad = \quad \mathbb{E}(P)\left[1+\frac{\lambda_3}{\mu_2+\mu_4}+\frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\Lambda+\mu_2)(\Lambda+\mu_4)}\right].$$

4-Stage model with exponential elongation

$$\mathrm{var}_{\mathrm{E}}(P)=\mathbb{E}(P)\left[1+\frac{\lambda_3\mu_3(\mu_2+\mu_3+\mu_4)}{(\mu_2+\mu_3)(\mu_2+\mu_4)(\mu_3+\mu_4)}+\right.$$
$$\left.\frac{\lambda_2\lambda_3(1-\delta_+)\mu_3\mu_4^2}{(\Lambda+\mu_2)(\mu_4^2-\mu_3^2)}\left(\frac{\Lambda+\mu_2+\mu_3}{\mu_3(\mu_2+\mu_3)(\Lambda+\mu_3)}-\frac{\Lambda+\mu_2+\mu_4}{\mu_4(\mu_2+\mu_4)(\Lambda+\mu_4)}\right)\right].$$

Deterministic vs Exponential



Conclusions

Few results:

- explicit formula depending on the model parameters for specific assumptions;
- counter-intuitive: $\mathrm{var}_{\mathrm{DET}}(P) > \mathrm{var}_{\mathrm{EXP}}(P)$.

Conclusions

Few results:

- explicit formula depending on the model parameters for specific assumptions;
- counter-intuitive: $\mathrm{var}_{\mathrm{DET}}(P) > \mathrm{var}_{\mathrm{EXP}}(P)$.

Biological consequences:

- the estimated variance could have been underestimated;
- deterministic protein elongation as upper-bound for protein variance;
- possibility to compute (numerically) more precise protein variance with realistic assumptions.