Stochastic model for gene expression in prokaryotes

Emanuele LEONCINI

INRIA Rocquencourt - INRA Jouy-en-Josas

"Modélisation pour l'Evolution du Vivant" – CMAP

November 21st 2012

Central role of protein production

- Proteins are the core of biologic processes: *enzymes*, DNA duplication machinery, ...
- $\sim$ 50% of the bacteria dry weight

Central role of protein production

- Proteins are the core of biologic processes: *enzymes*, DNA duplication machinery, ...
- $\sim$ 50% of the bacteria dry weight

A highly consuming process:

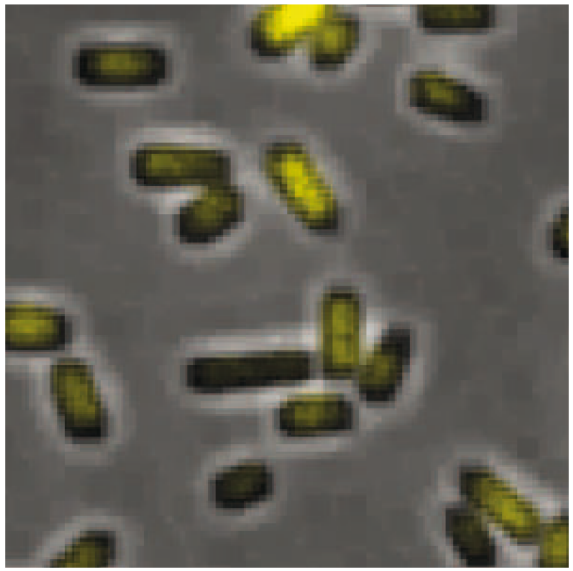
- more than 85% of cell resources
- $\sim$ 3.5 millions of proteins in each cell
- $\sim$ 2000 types of proteins produced at any time at any growth condition (volume growth)
- proteins ranging from few dozens up to 10^5

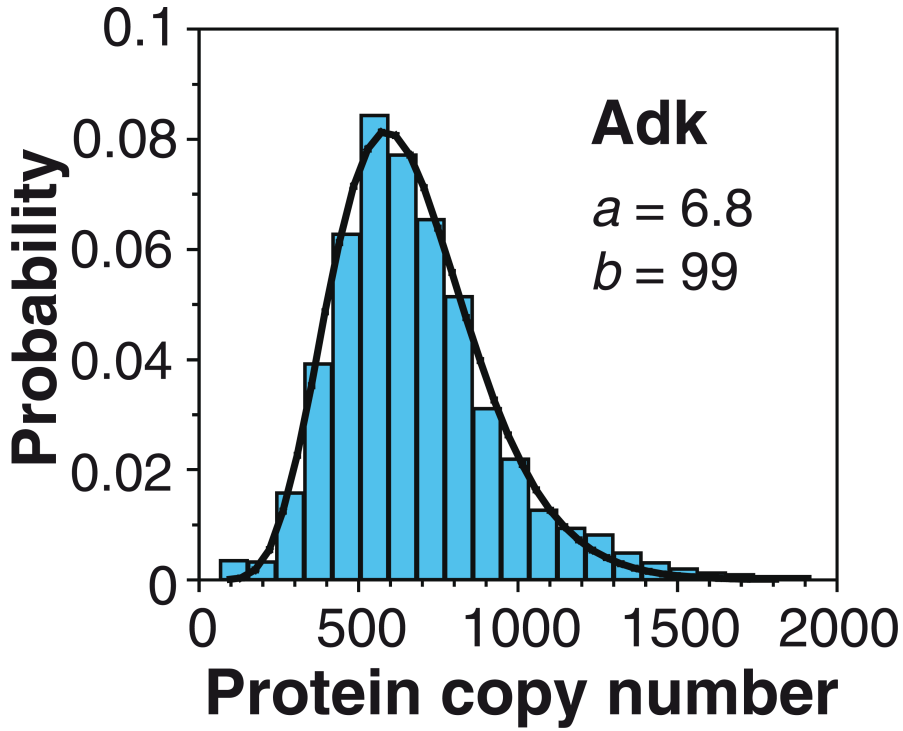
Stochasticity in protein production

- bacterial *cytoplasm*: disordered medium
- main cellular motility mechanism: diffusion in a stiff medium
- protein production steps require the encounter of cellular components
- experiments: direct proof of fluctuations in protein levels

Stochasticity in protein production: experimental viewpoint

ADk cytoplasm protein

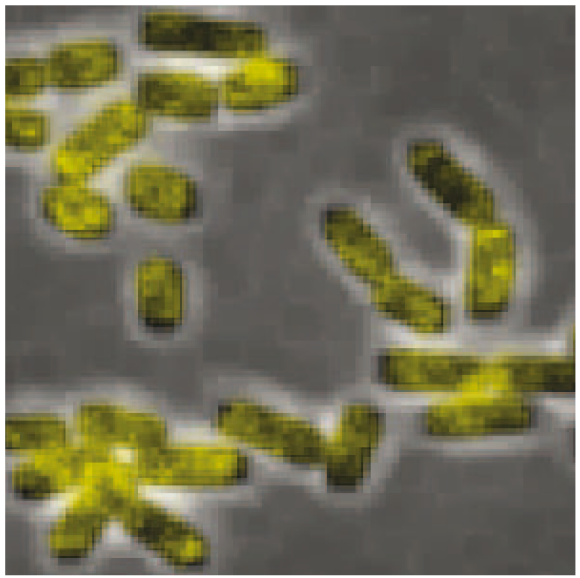


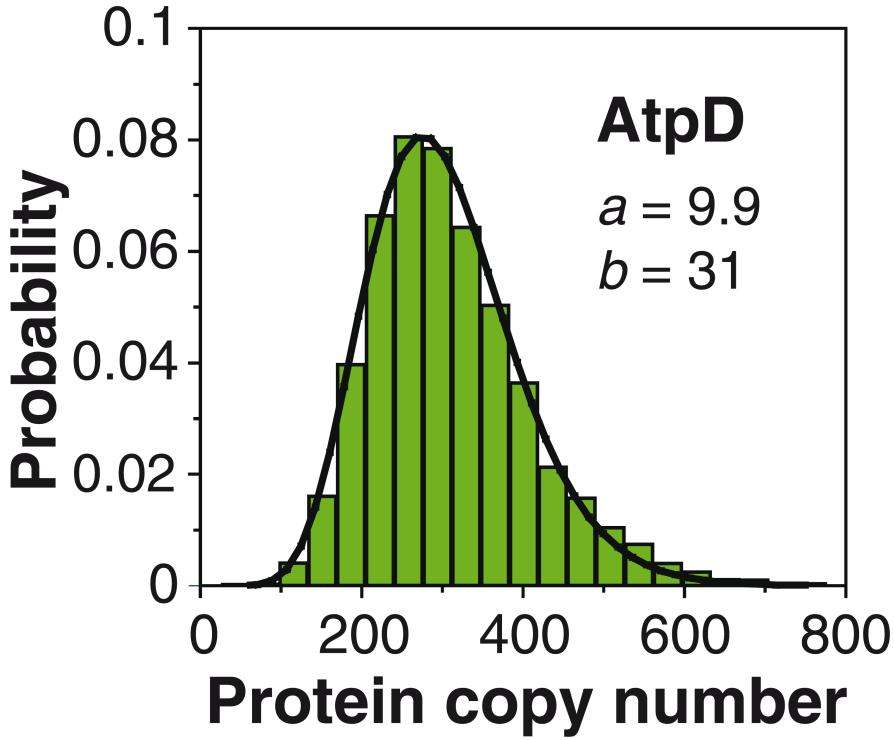


Taniguchi Y. *et alii*, Quantifying *E. Coli* Proteome and Transcriptome, 2010.

Stochasticity in protein production: experimental viewpoint

AtpD membrane protein

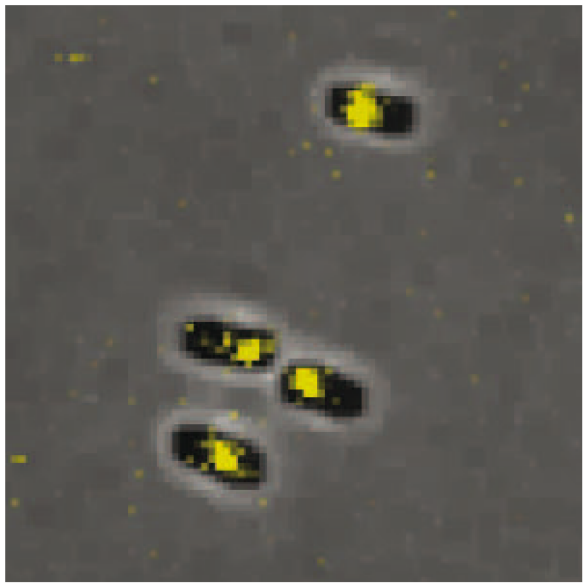


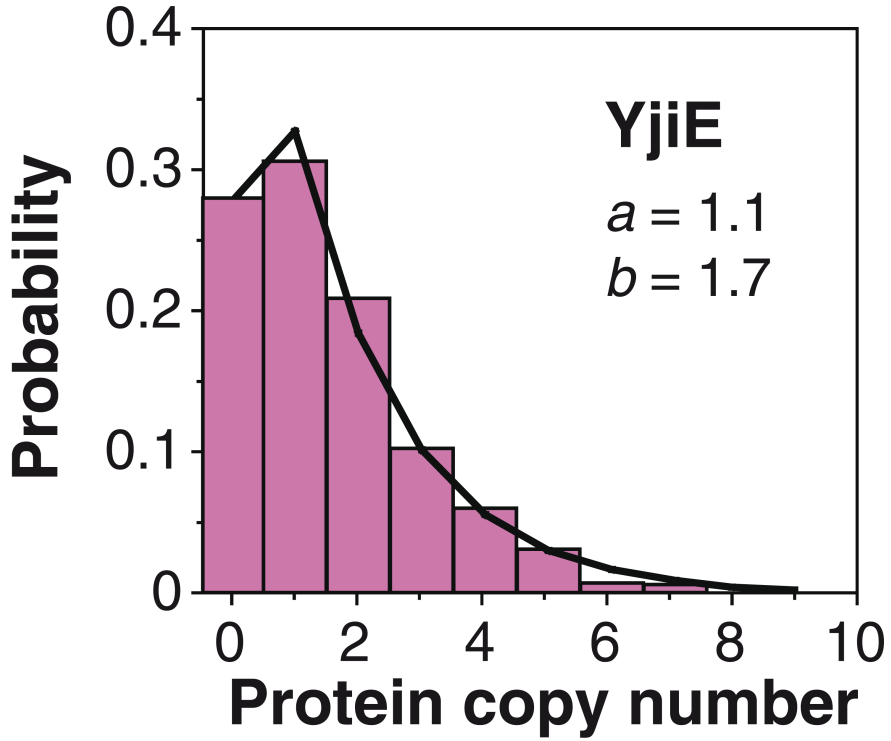


Taniguchi Y. *et alii*, Quantifying *E. Coli* Proteome and Transcriptome, 2010.

Stochasticity in protein production: experimental viewpoint

YjiE DNA-binding protein





Taniguchi Y. *et alii*, Quantifying *E. Coli* Proteome and Transcriptome, 2010.

Questions:

- How can simple organisms deal with fluctuations?
- What is the nature of this stochasticity?
- How can we characterize it?

Questions:

- How can simple organisms deal with fluctuations?
- What is the nature of this stochasticity?
- How can we characterize it?

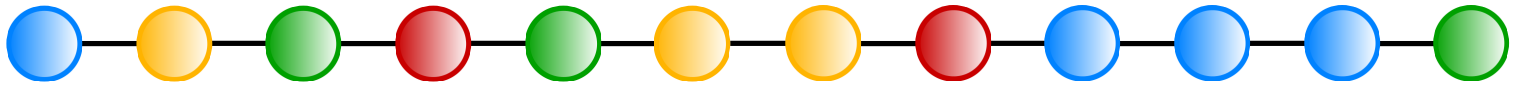
Objective: find an appropriate description of the protein production

Biology

What is a protein ?

A protein is a chain of elementary bricks, the *amino acids*, and is characterized by

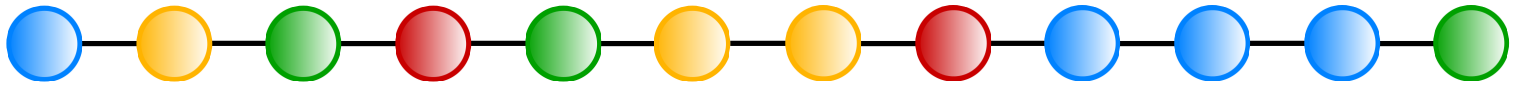
- the order of the amino acids



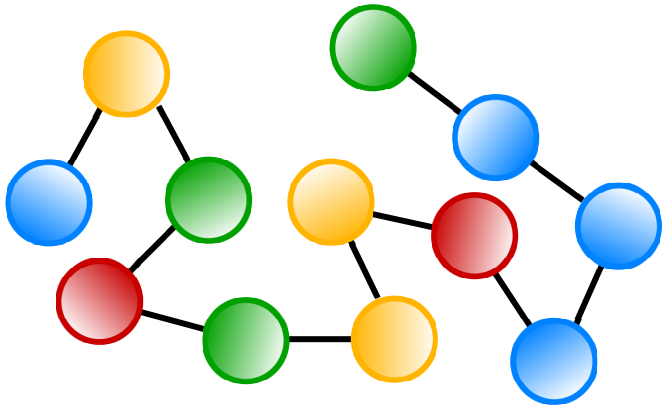
What is a protein ?

A protein is a chain of elementary bricks, the *amino acids*, and is characterized by

- the order of the amino acids

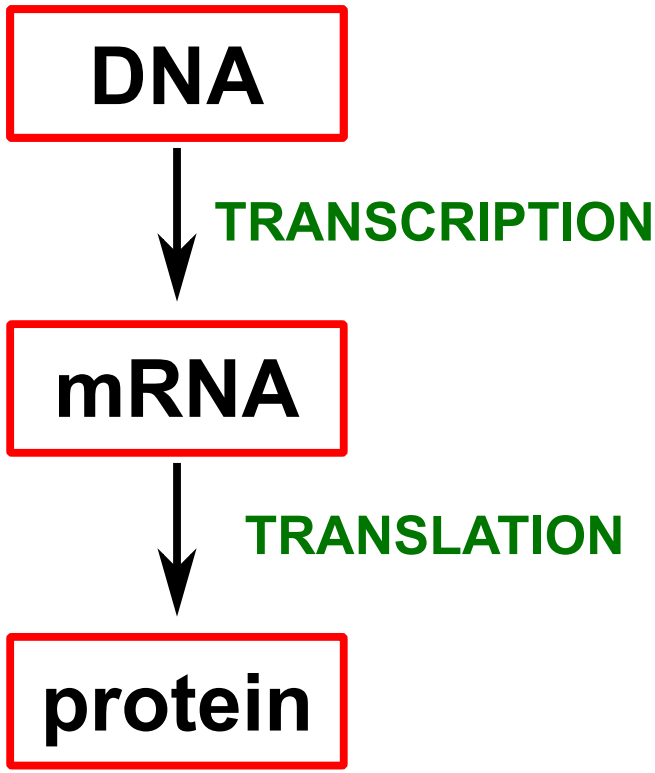


- the 3D conformation



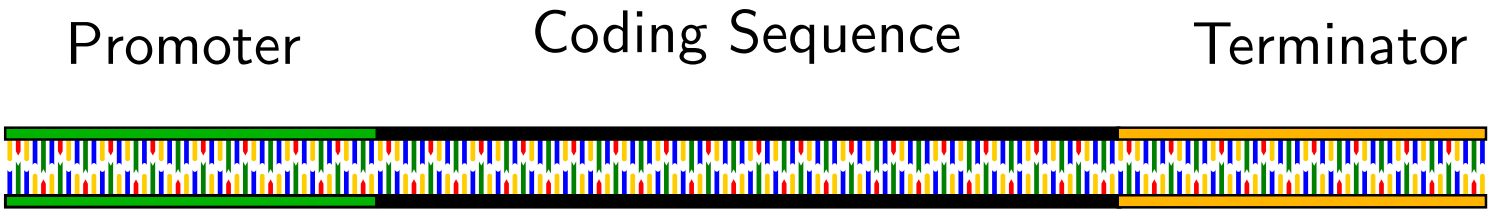
both determining the protein function

Central Dogma of molecular biology

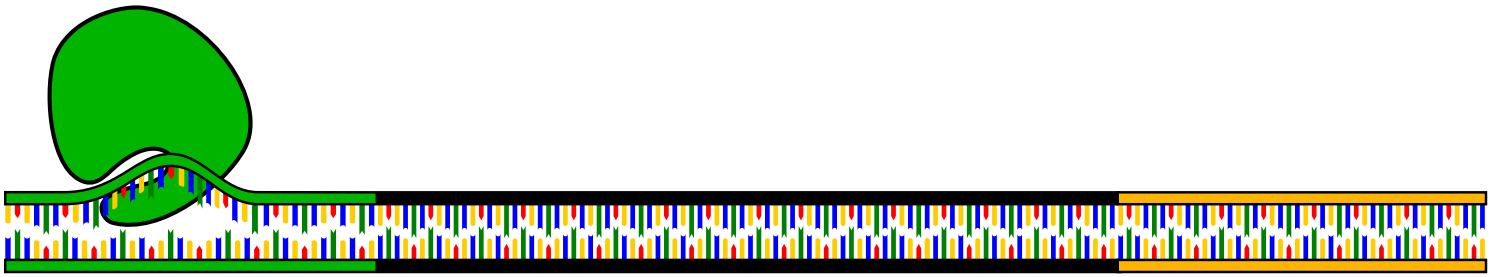


DNA

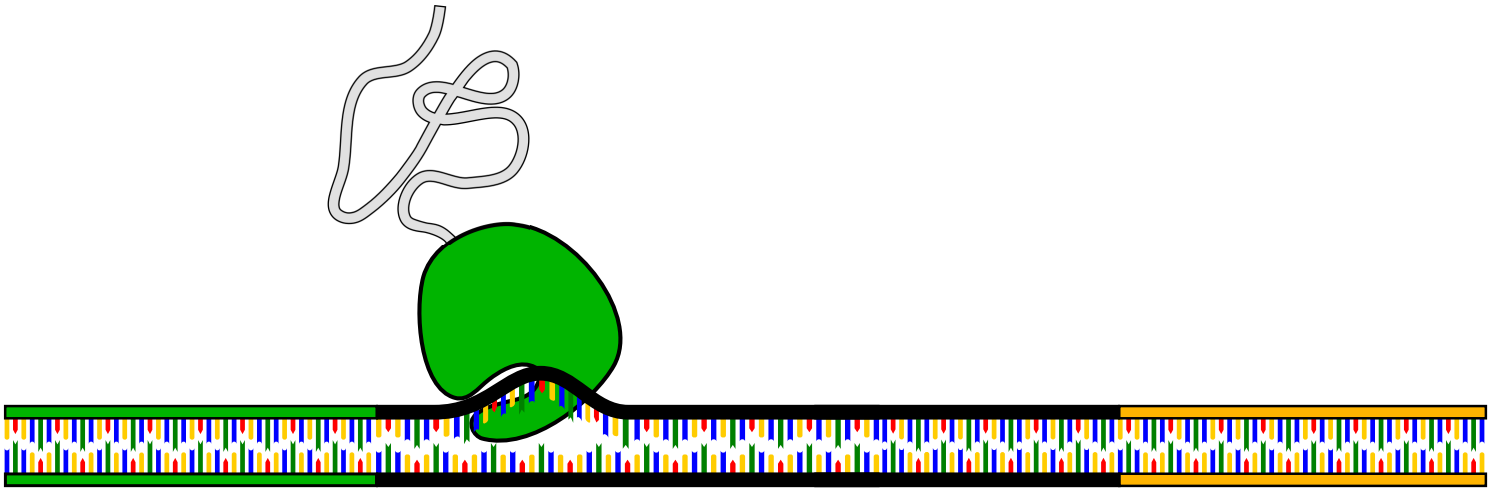
Gene: portion of DNA encoding for a specific protein.



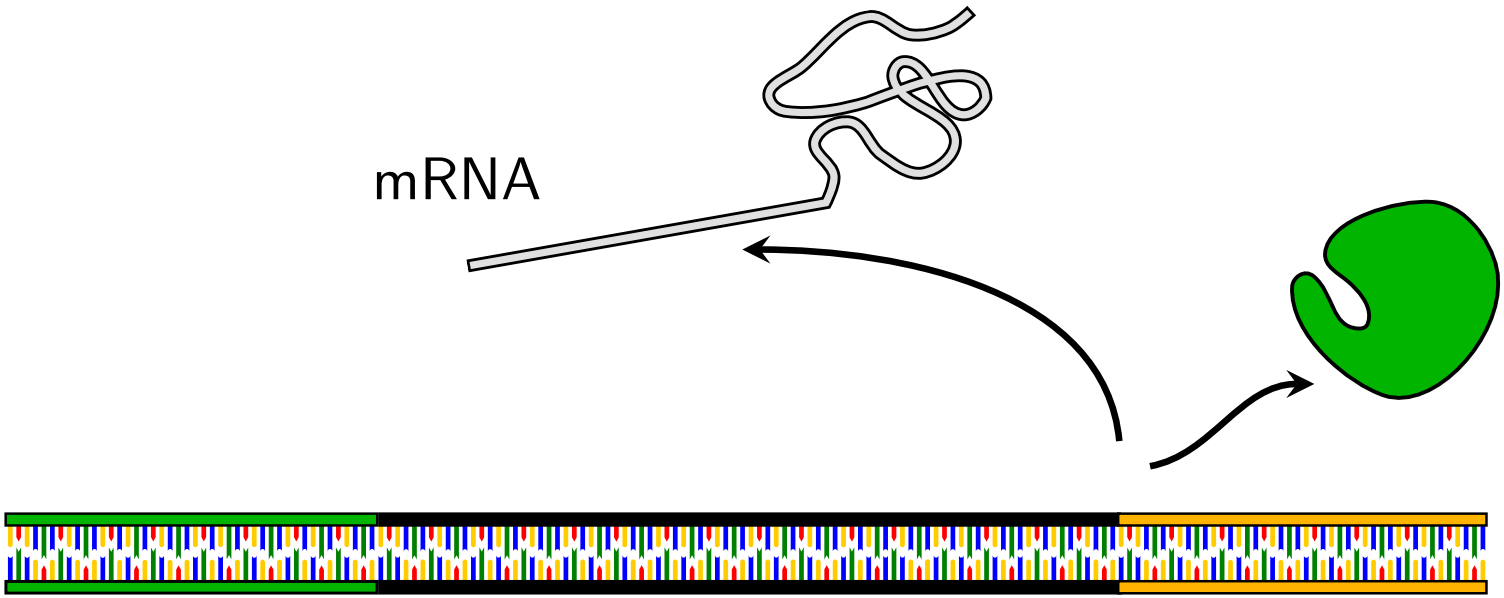
Transcription: initiation



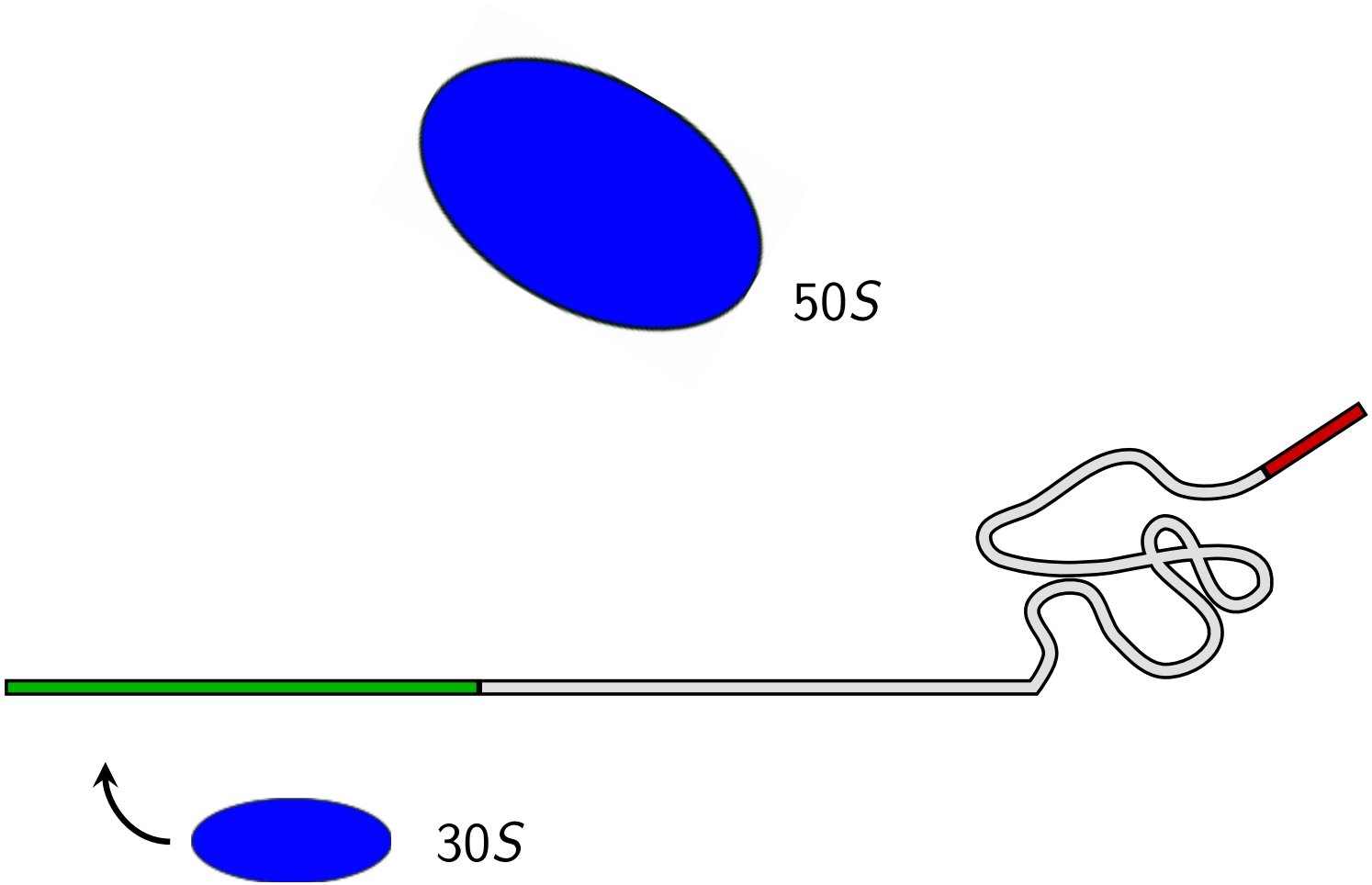
Transcription: mRNA elongation



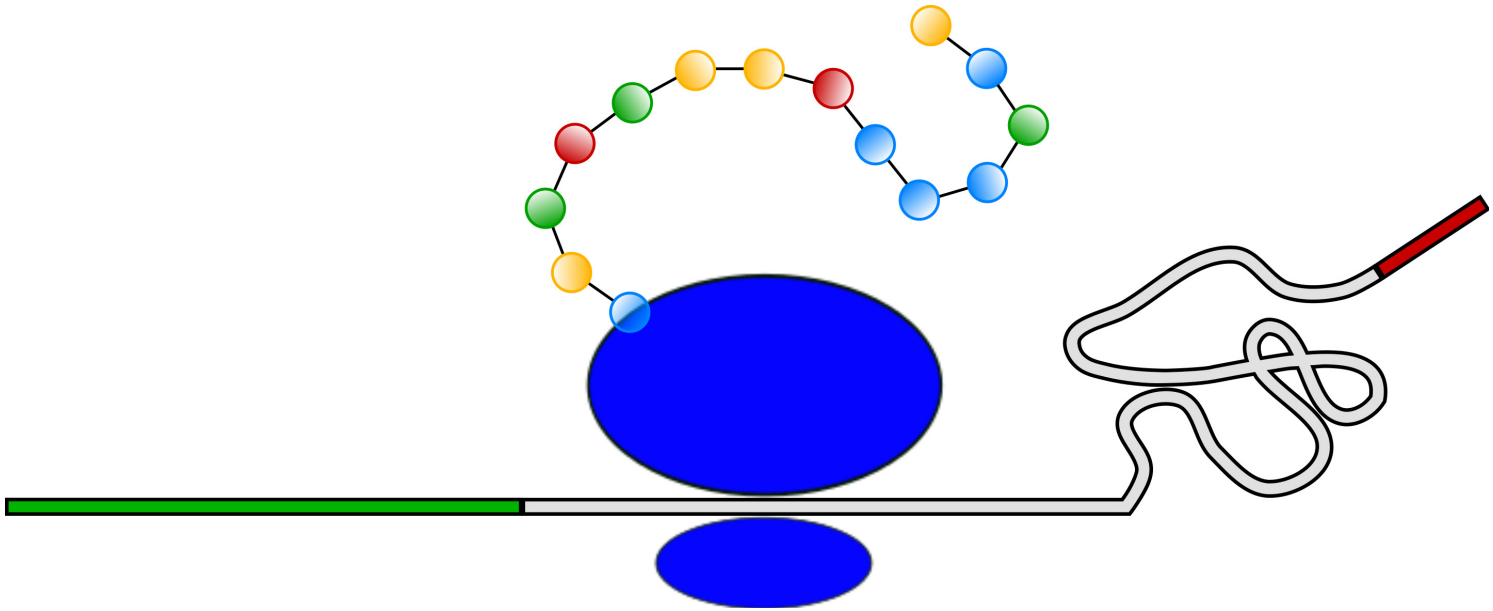
Transcription: termination



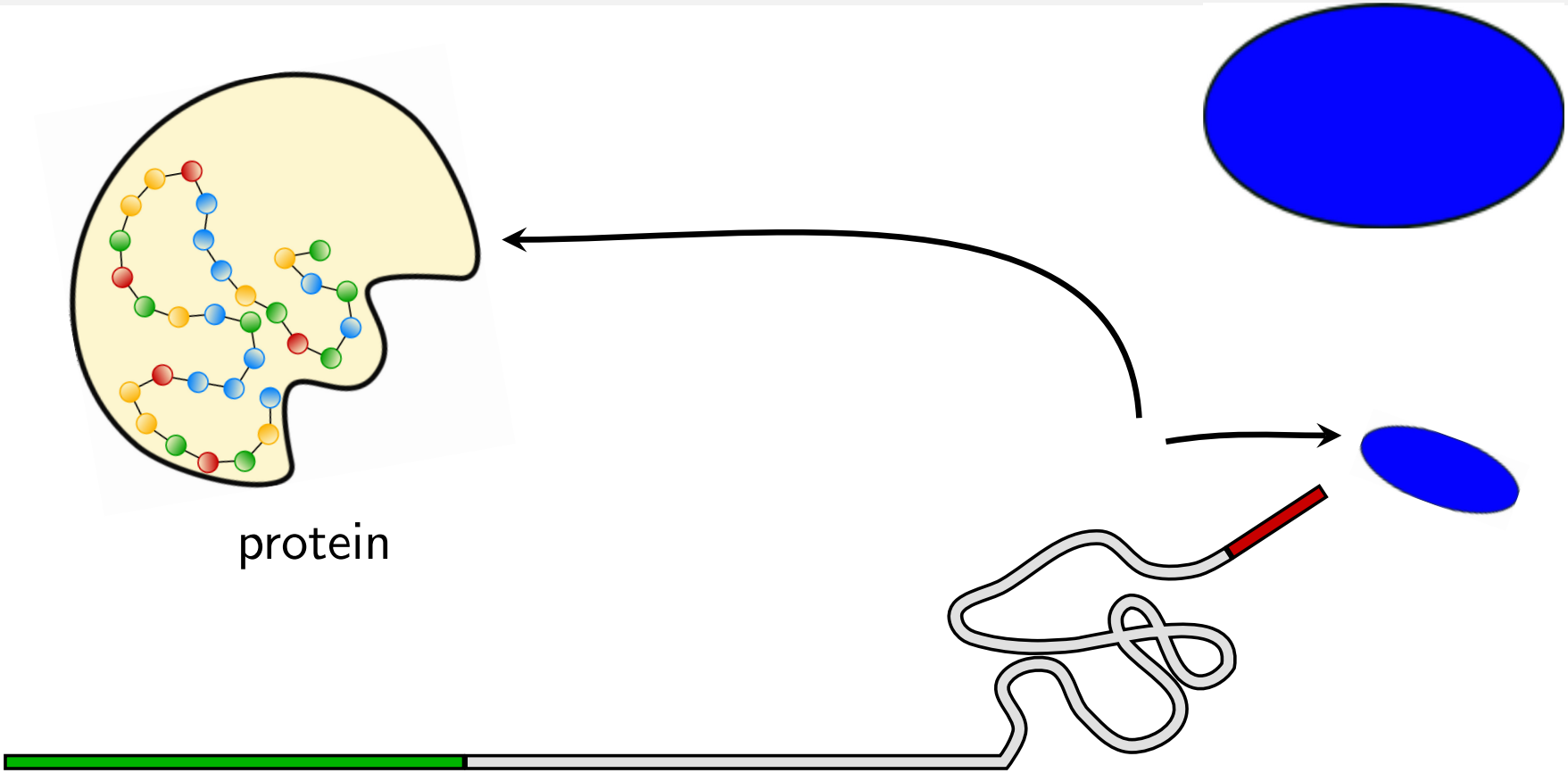
Translation: initiation



Translation: protein elongation

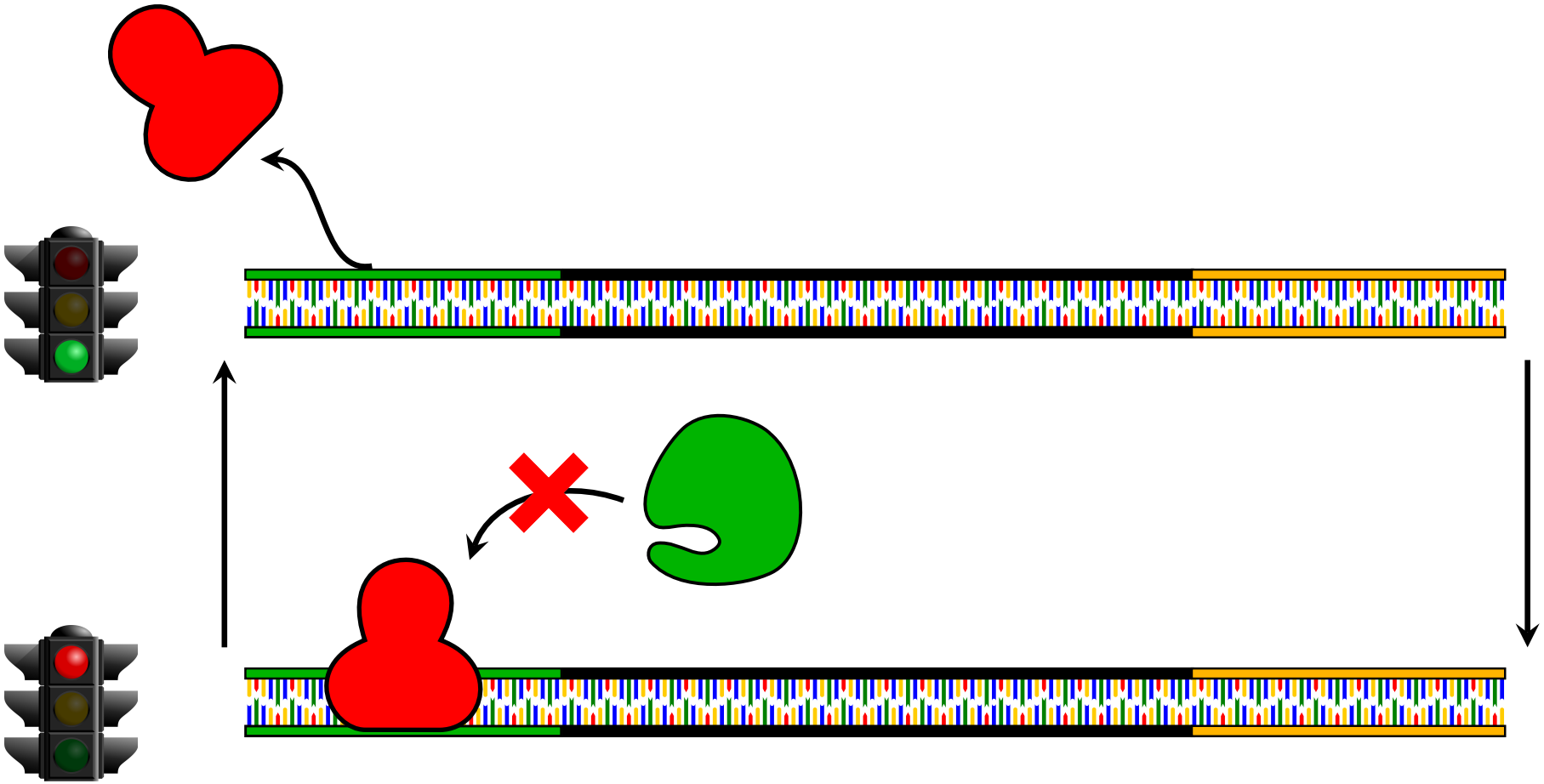


Translation: termination



Gene activation

Two states of gene: **active** and **inactive**.



Model

Classic Models

First stochastic models late 70s (Rigney 1977, Berg 1978)

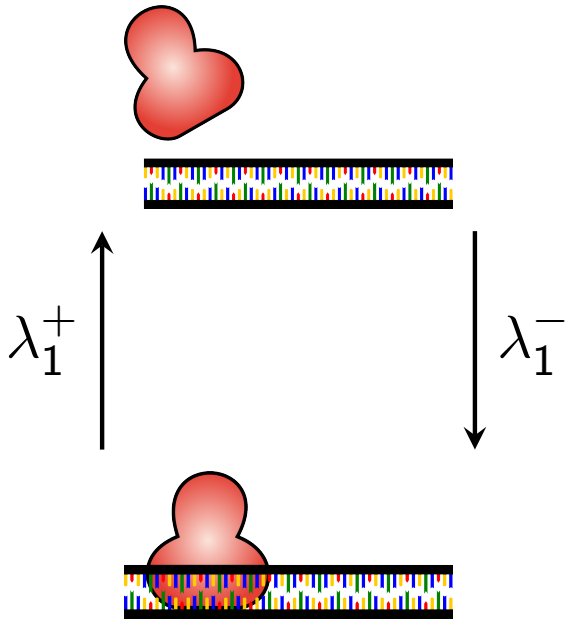
Fundamental assumption: each step has exponentially distributed duration

Markovian description of the protein production.

Classic Models

Activation

$$Y(t) \in \{0, 1\}$$

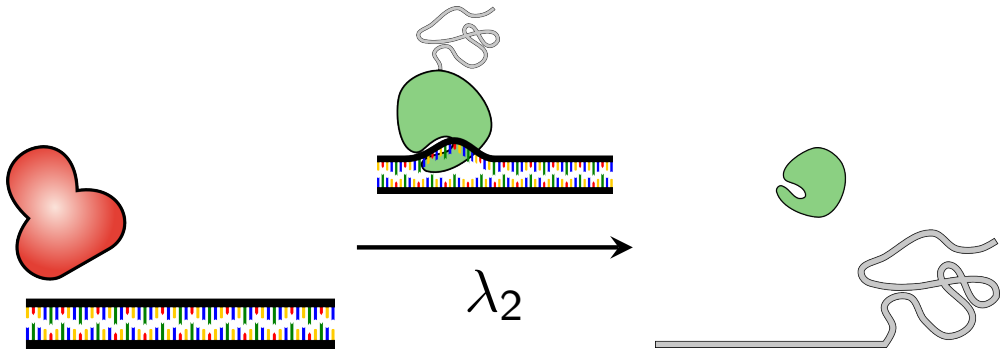


Paulsson J., Models of stochastic gene expression, 2005.

Classic Models

Transcription

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



Paulsson J., Models of stochastic gene expression, 2005.

Classic Models

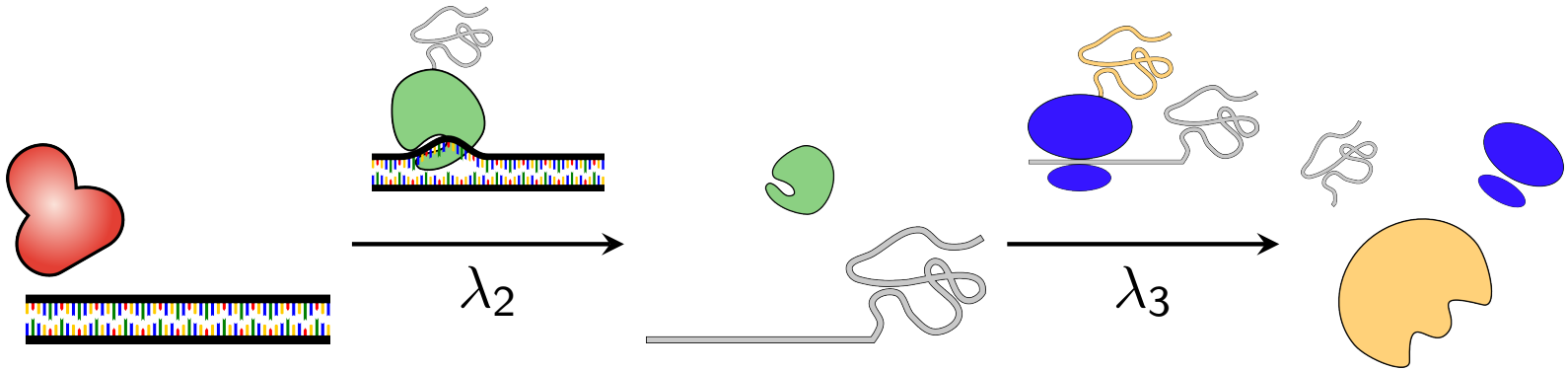
Transcription

Translation

$Y(t) \in \{0, 1\}$

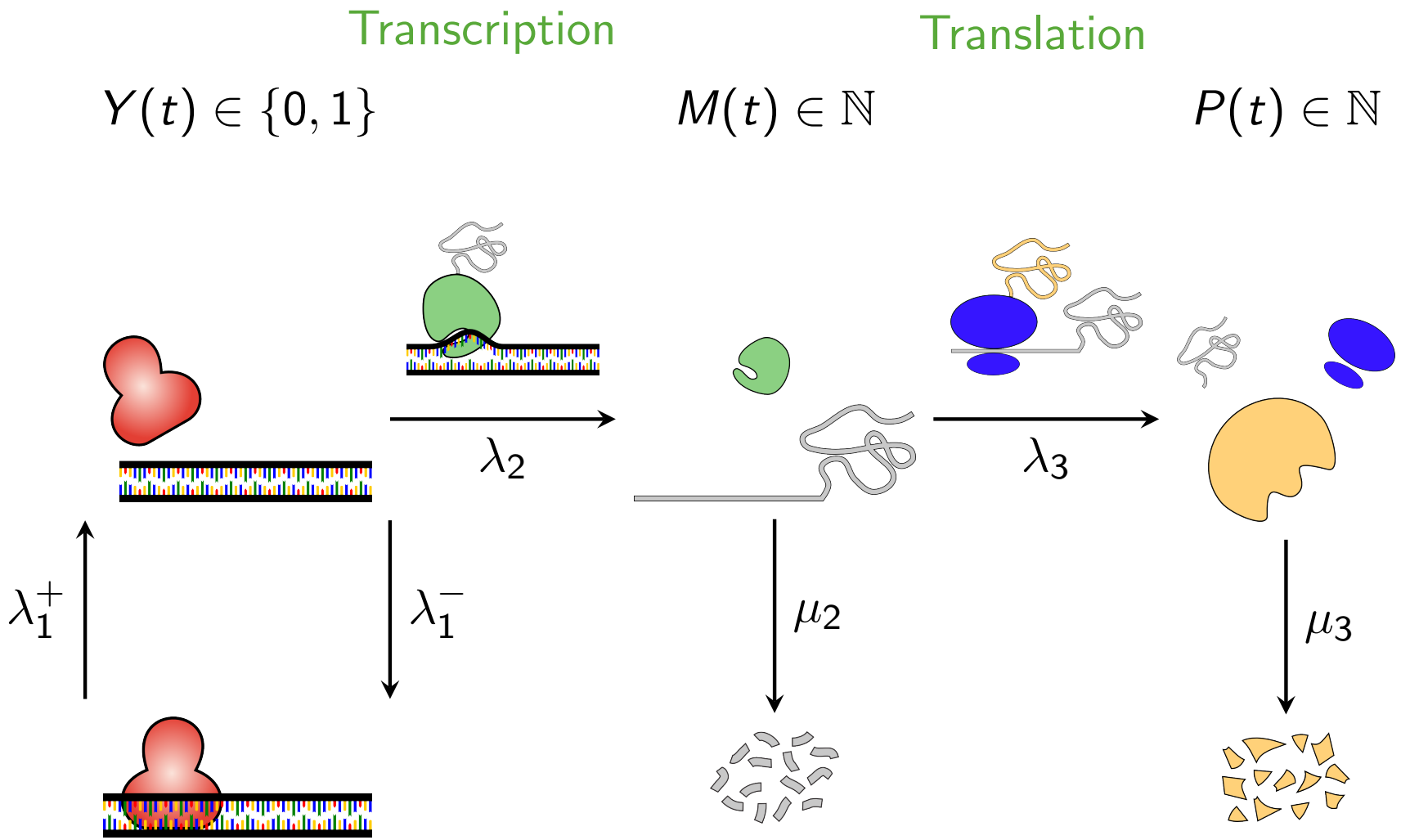
$M(t) \in \mathbb{N}$

$P(t) \in \mathbb{N}$



Paulsson J., Models of stochastic gene expression, 2005.

Classic Models



Paulsson J., Models of stochastic gene expression, 2005.

Results of classic models

Tools:

- *Markov chains;*
- *Fokker-Plank equations;*

Results of classic models

Tools:

- *Markov chains;*
- *Fokker-Plank equations;*

Results:

- *simple analytic close formulas of mean and variance of protein number, i.e. see Paulsson (2005)*

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

- *first quantitative characterisation of protein fluctuations.*

Why this modelisation is not satisfying ?

Two processes are not exponential

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

Why this modelisation is not satisfying ?

Two processes are not exponential

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

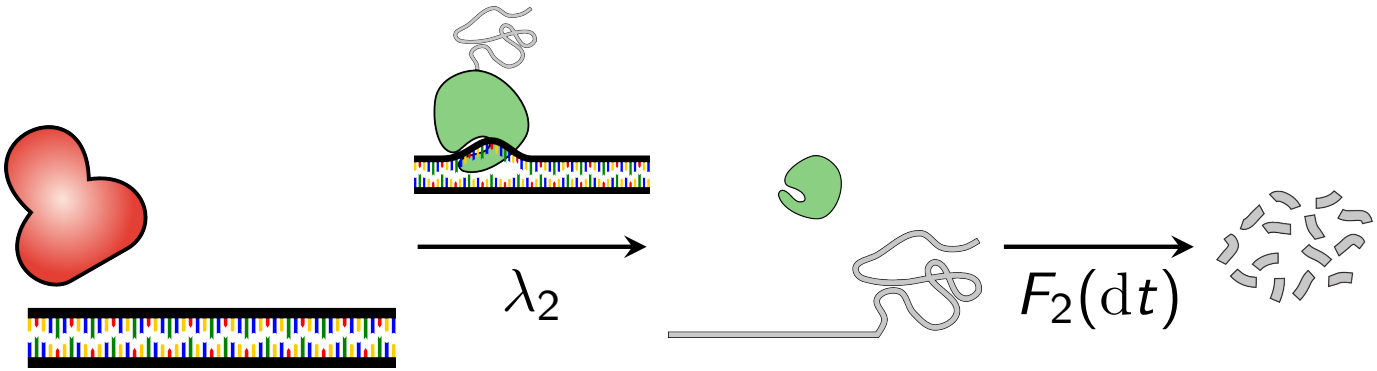
Why this modelisation is not satisfying ?

Two processes are not exponential

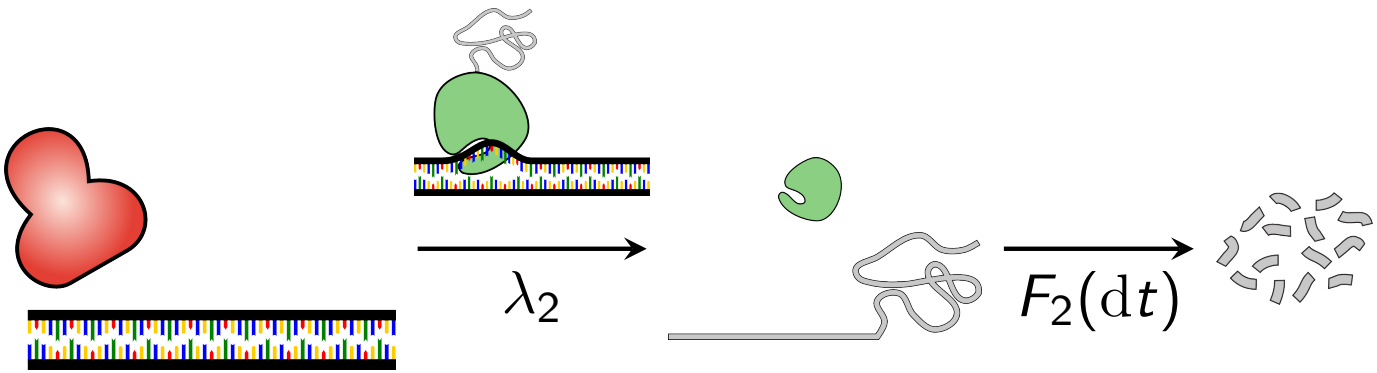
- *protein/mRNA elongation*
  - *exponential assumption for protein elongation* $\Rightarrow$ *incorrect*
  - *real elongation process: large number of exponential steps*
  - *non-exponential elongation distribution cannot be included in classic models*
- *protein degradation*

MPPP: toward a new description of gene expression

General distributions: introducing MPPP



General distributions: introducing MPPP

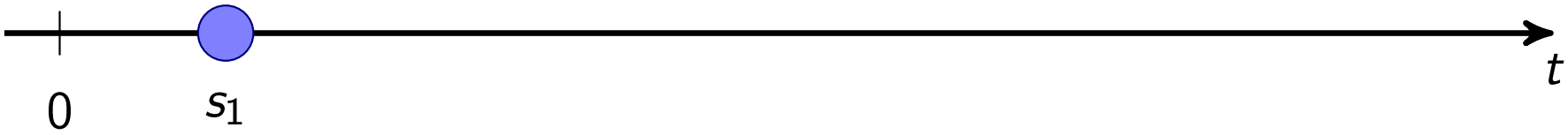


Assumptions:

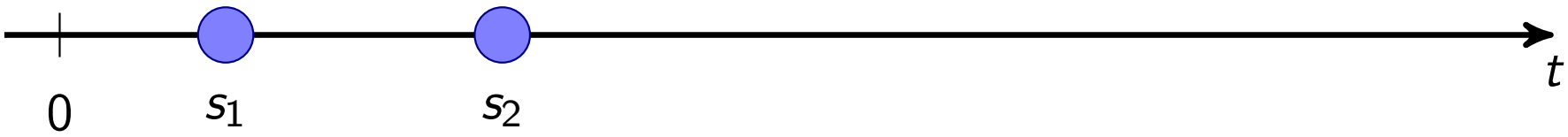
- mRNA births follow a Poisson process of parameter λ_2
- mRNA lifetimes σ_2 have distribution $F_2(\mathrm{d}t)$

mRNA dynamics are a $M/G/\infty$ queue

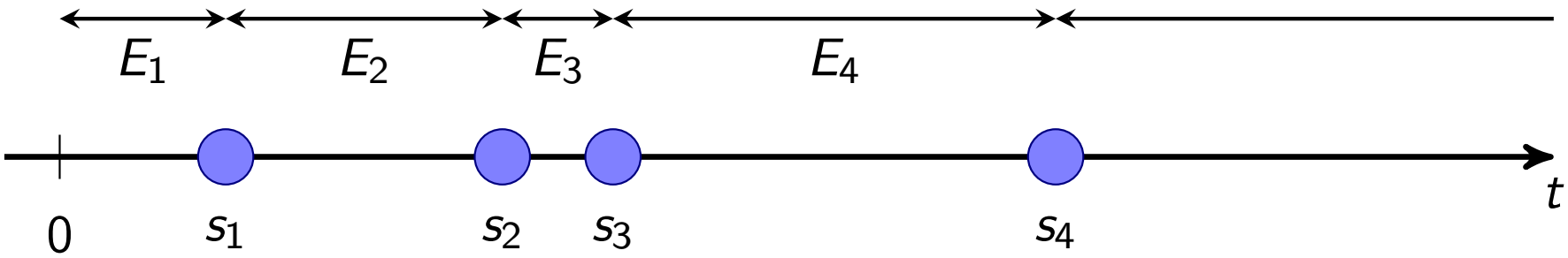
mRNA births with active gene



mRNA births with active gene



mRNA births with active gene



E_i exponential random variables of parameter λ_2.

mRNA births up to time t

$$\mathcal{M}^{(b)}(t) = \sum_{n=1}^{+\infty} \mathbb{1}_{\{s_n \leq t\}}$$

Poisson process as random measure

$$\mathcal{M}^{(b)}(\mathrm{d}s) = \sum_{n=1}^{+\infty} \delta_{s_n}(\mathrm{d}s)$$

Poisson process as random measure

$$\mathcal{M}^{(b)}(\mathrm{d}s) = \sum_{n=1}^{+\infty} \delta_{s_n}(\mathrm{d}s)$$

In particular

$$\mathcal{M}^{(b)}(f) \overset{\text{def}}{=} \int_0^{+\infty} f(t)\mathcal{N}(\mathrm{d}t) = \sum_{n=1}^{+\infty} f(s_n)$$

Example

mRNA births

$$\mathcal{M}^{(b)}(t) = \mathcal{M}^{(b)}\left(\mathbb{1}_{\{s \leq t\}}\right)$$

Laplace transform of a Poisson point process

If $\mathcal{N}$ is a Poisson point process of intensity λ

$$\mathcal{L}_{\mathcal{N}}(f) = \mathbb{E}\left[e^{-\mathcal{N}(f)}\right] = \exp\left(-\lambda \int_0^{+\infty} \left(1 - e^{-f(x)}\right) \mathrm{d}x\right) \tag{1}$$

Laplace transform of a Poisson point process

If $\mathcal{N}$ is a Poisson point process of intensity λ

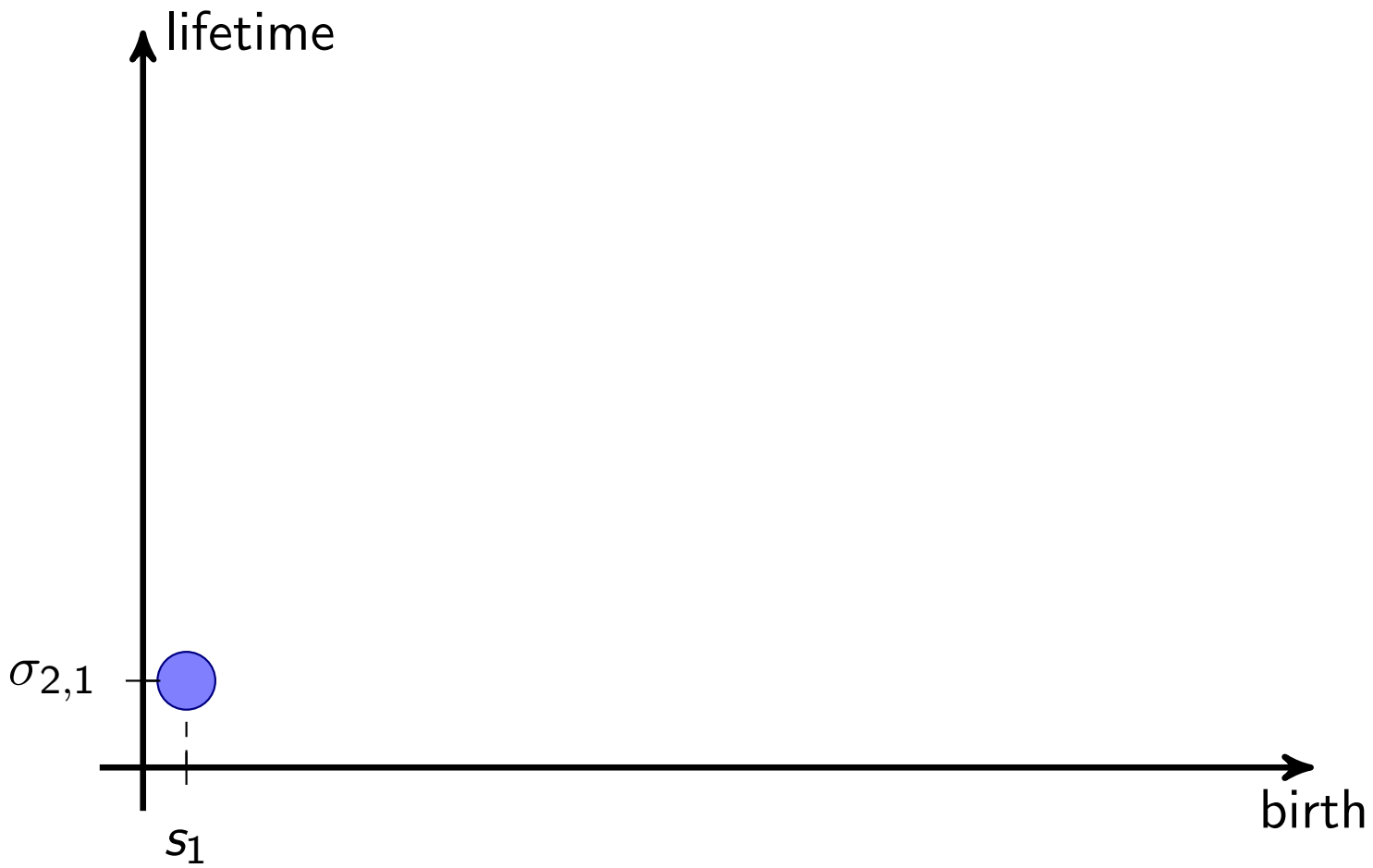
$$\mathcal{L}_{\mathcal{N}}(f) = \mathbb{E}\left[e^{-\mathcal{N}(f)}\right] = \exp\left(-\lambda \int_0^{+\infty} \left(1 - e^{-f(x)}\right) \mathrm{d}x\right) \qquad (1)$$

Special case: generating function of a Poisson process

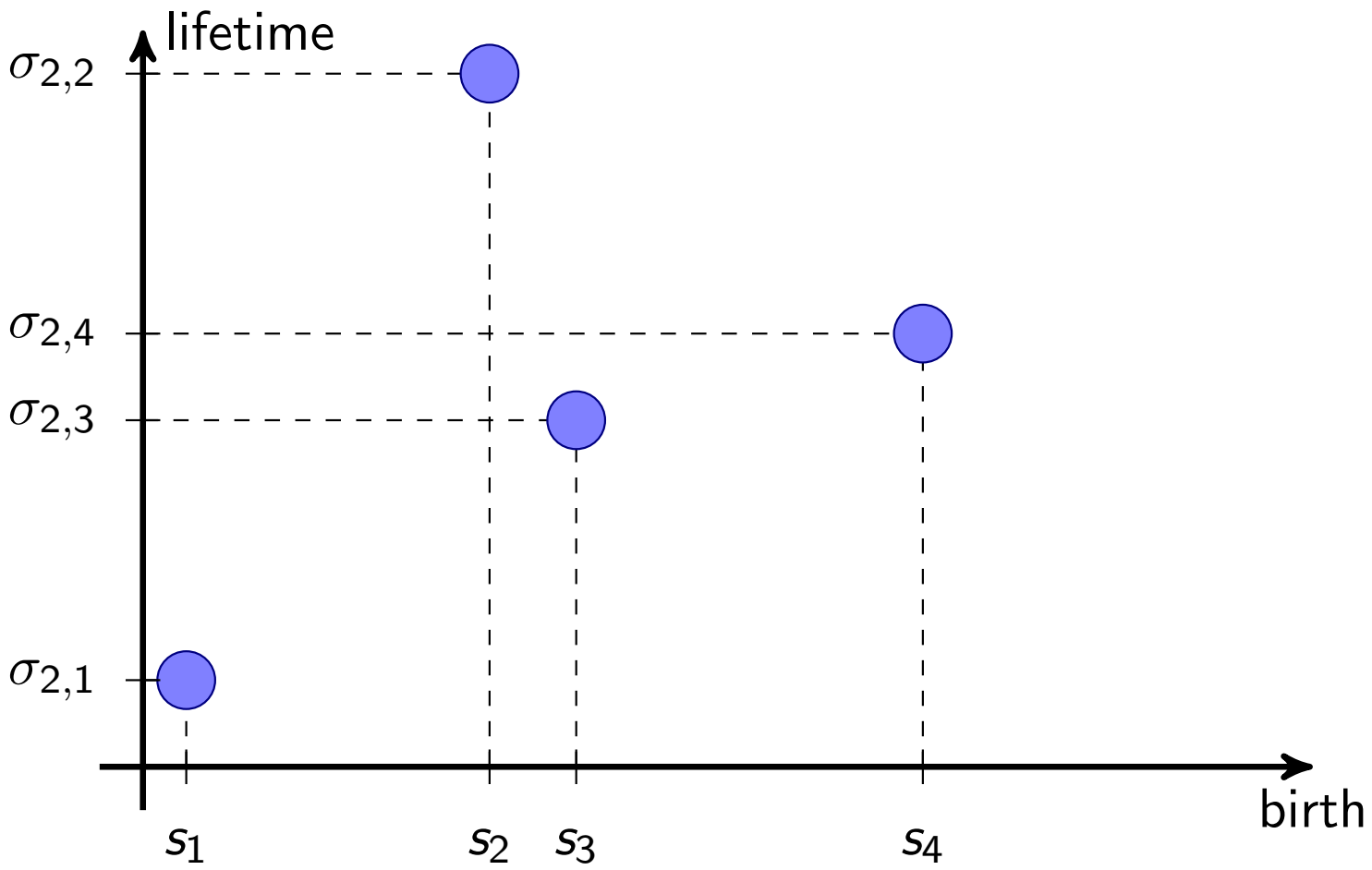
$$\mathcal{L}_{\mathcal{N}}(f) = \mathbb{E}\left[u^{\mathcal{N}(t)}\right] = e^{-\lambda t(1-u)} \qquad u \in]0,1],$$

with $f(x) = \ln\left(\frac{1}{u}\right) \mathbb{1}_{\{]0,t]\}}(x)$.

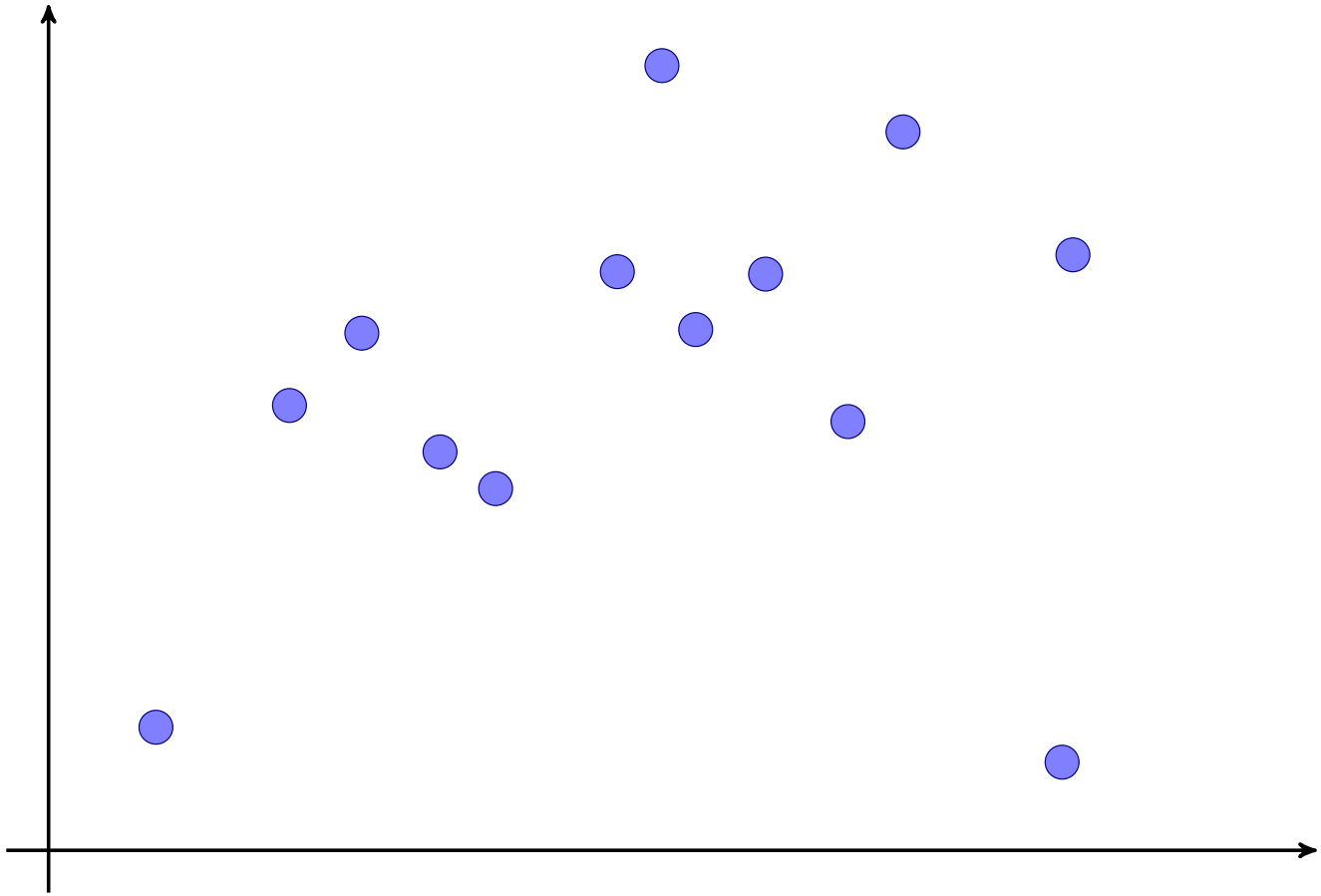
Construction of a Poisson process in $\mathbb{R}_+^2$



Construction of a Poisson process in $\mathbb{R}_+^2$



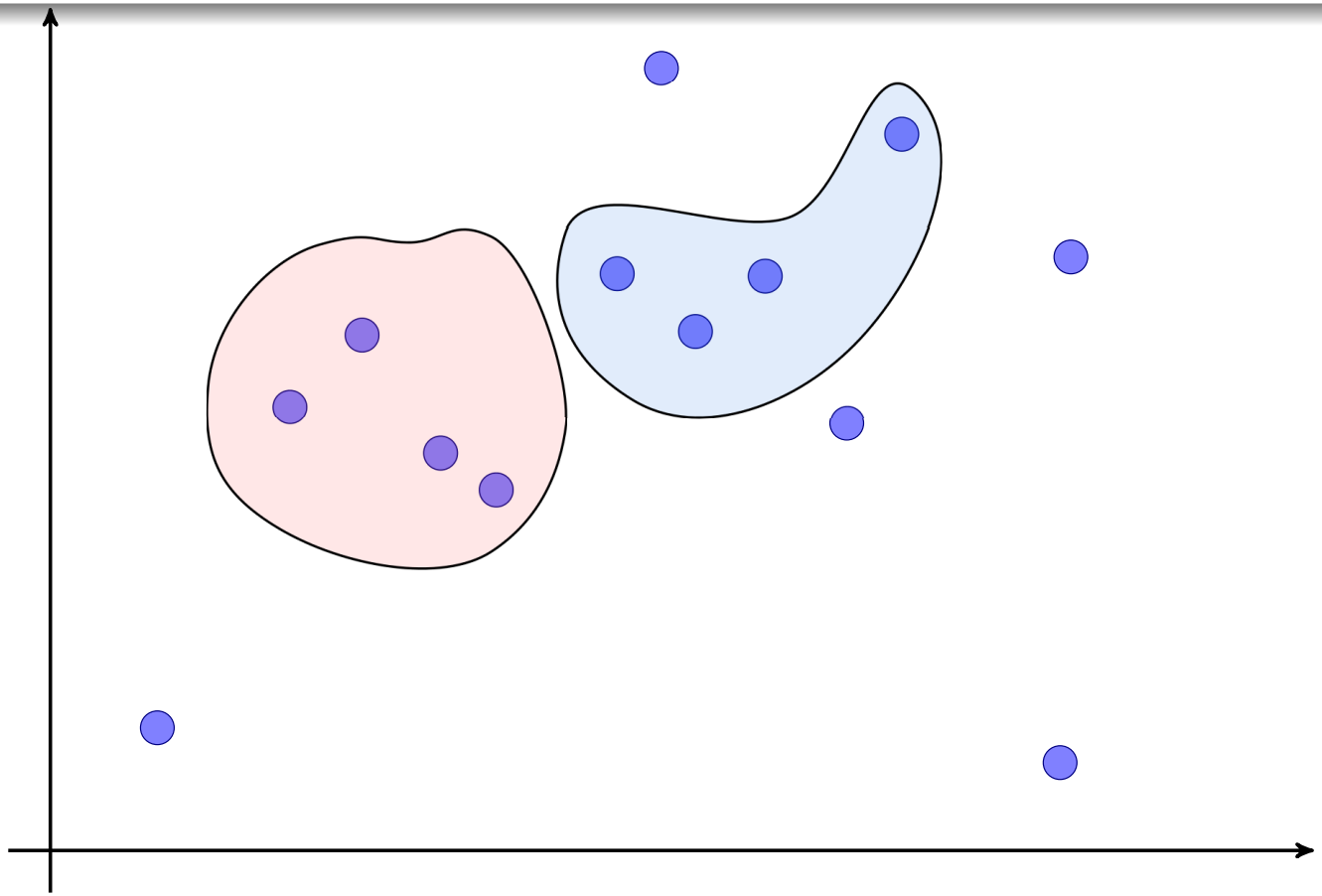
Construction of a Poisson process in $\mathbb{R}_+^2$



Construction of a Poisson process in $\mathbb{R}^2_+$

Property:

For any disjoint sets $A_1, \ldots, A_n$, the random variables $\mathcal{N}(A_1), \ldots, \mathcal{N}(A_n)$ are independent (Poisson point process).



Marked Poisson Point Process

Ingredients: $\mathcal{N} = (s_n)$ *Poisson point process of intensity* $\lambda_2 \, \mathrm{d}x$
$(\sigma_{2,n})$ *i.i.d. with distribution* $F_2(\mathrm{d}y)$

MPPP: $$\mathcal{N}_{\lambda_2} \overset{\text{def}}{=} \sum \delta_{(s_n, \sigma_{2,n})}$$

Mark: $(\sigma_{2,n})$

Marked Poisson Point Process

Ingredients: $\mathcal{N} = (s_n)$ *Poisson point process of intensity* $\lambda_2 \, \mathrm{d}x$
$(\sigma_{2,n})$ *i.i.d. with distribution* $F_2(\mathrm{d}y)$

MPPP: $\mathcal{N}_{\lambda_2} \overset{\text{def}}{=} \sum \delta_{(s_n, \sigma_{2,n})}$

Mark: $(\sigma_{2,n})$

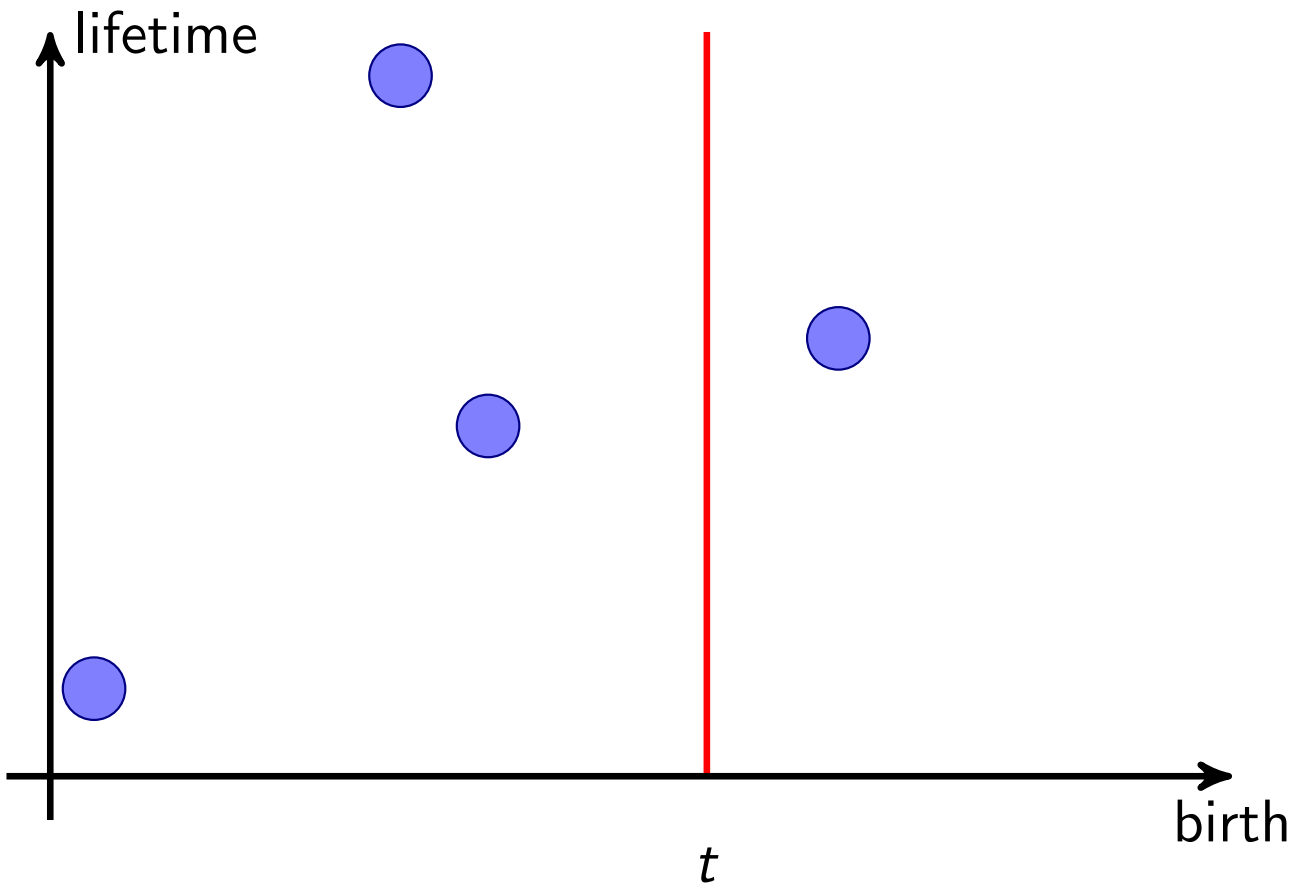
The Laplace transform of a Marked Poisson point process is

$$\mathcal{L}_{\mathcal{N}_{\lambda_2}}(f) = \mathbb{E}\left[e^{-\mathcal{N}(f)}\right] = \exp\left(-\int \left(1 - e^{-f(x,y)}\right) \lambda_2 \, \mathrm{d}x F_2(\mathrm{d}y)\right)$$

where $\mathcal{N}_{\lambda_2}(f) = \sum_n f\left(s_n, \sigma_{2,n}\right)$.

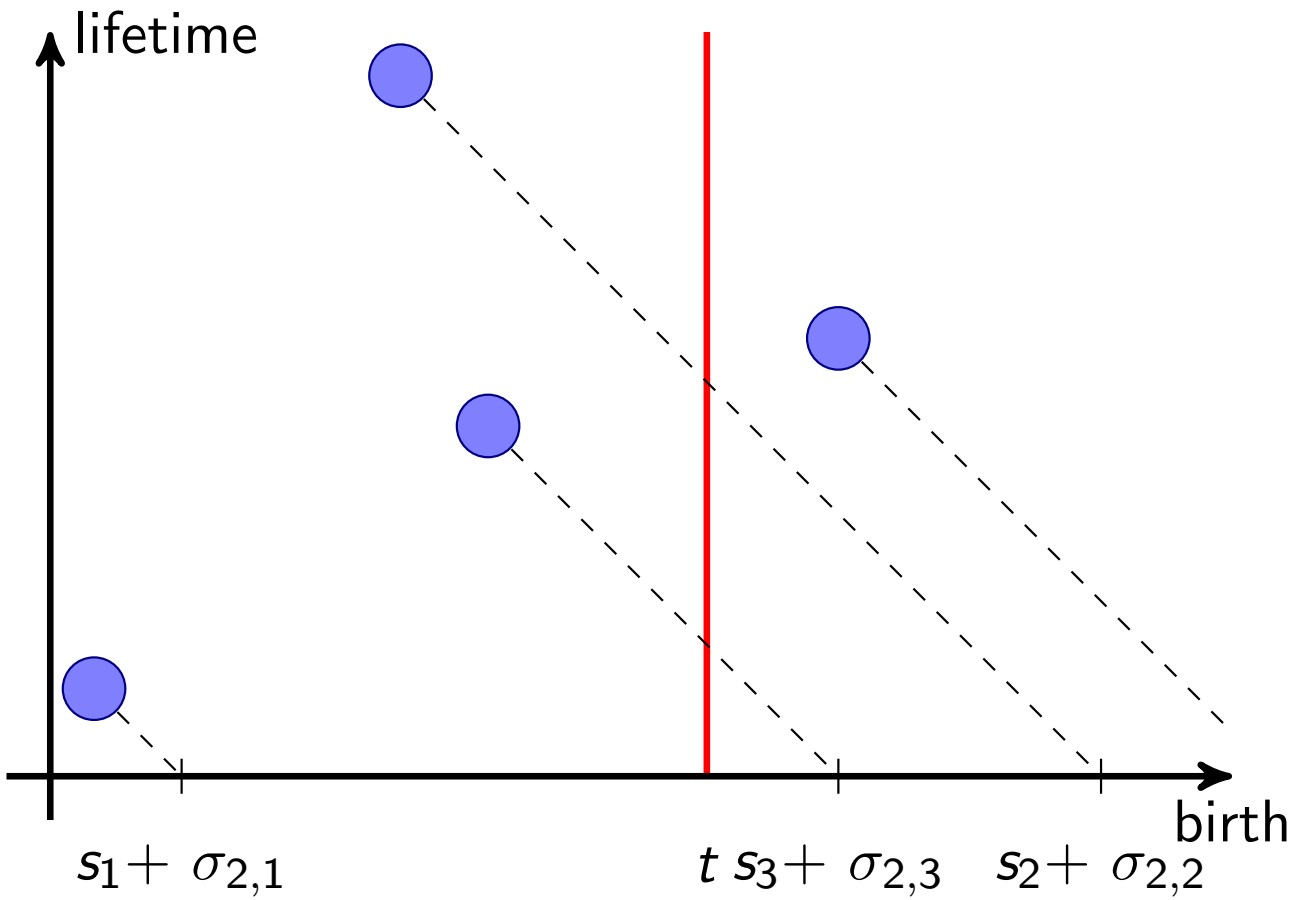
mRNA

How many mRNAs at time t ?



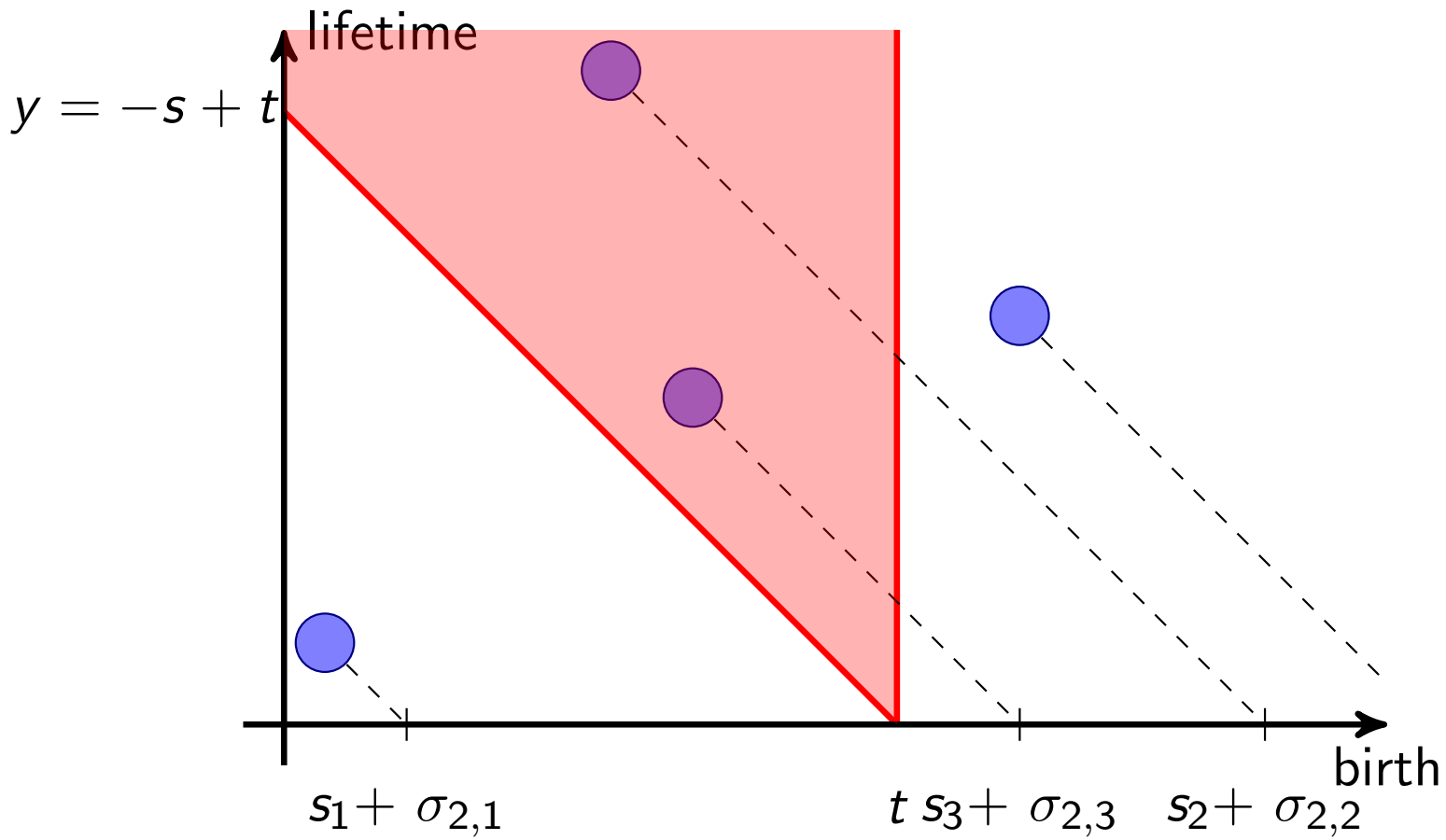
mRNA

How many mRNAs at time t ?



mRNA

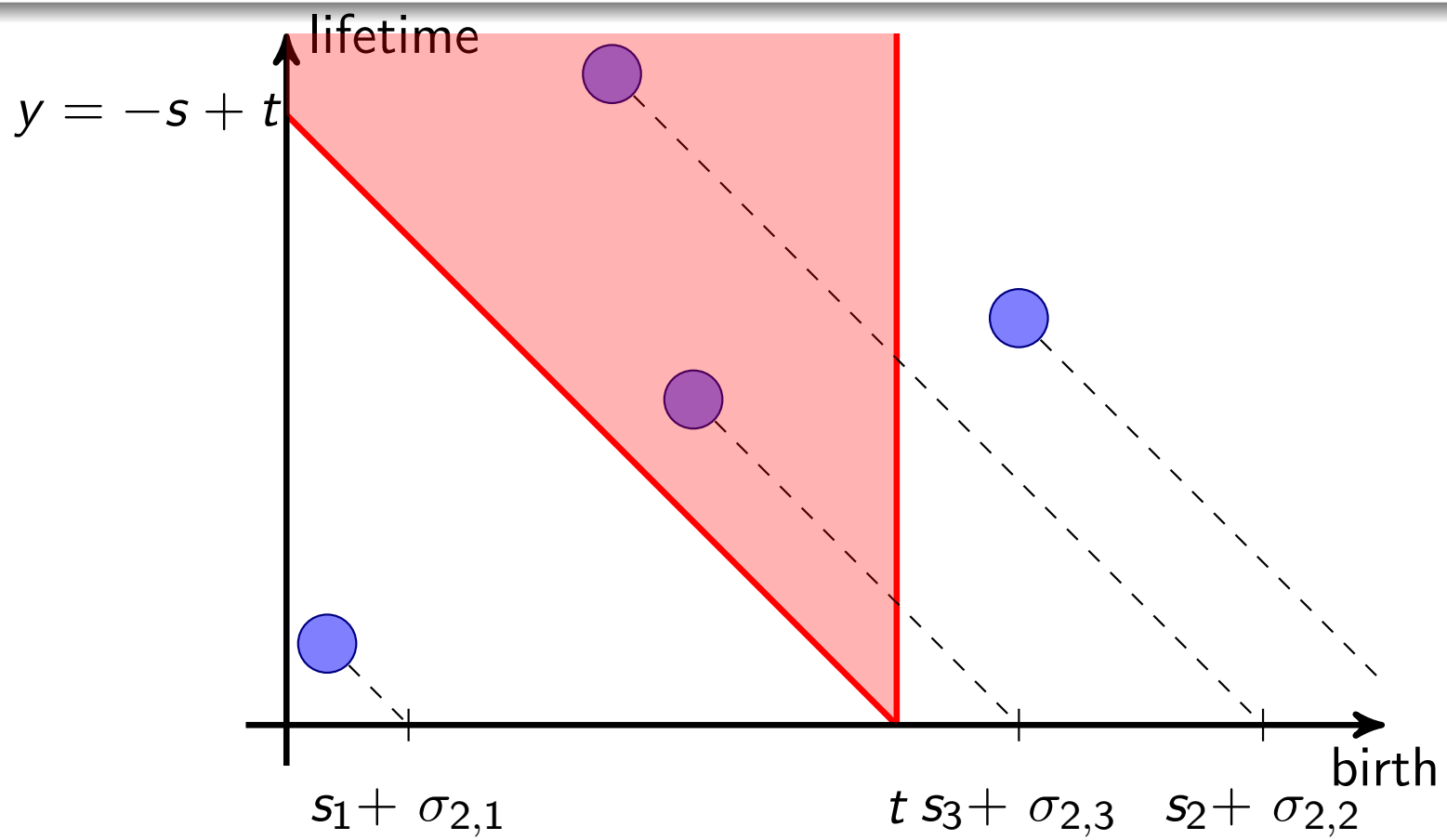
How many mRNAs at time t ?



mRNA

How many mRNAs at time t ?

$$M_t = \mathcal{N}(A) = \sum_n \mathbb{1}_{\{0 \leq s_n \leq t \leq s_n + \sigma_{2,n}\}}$$



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

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

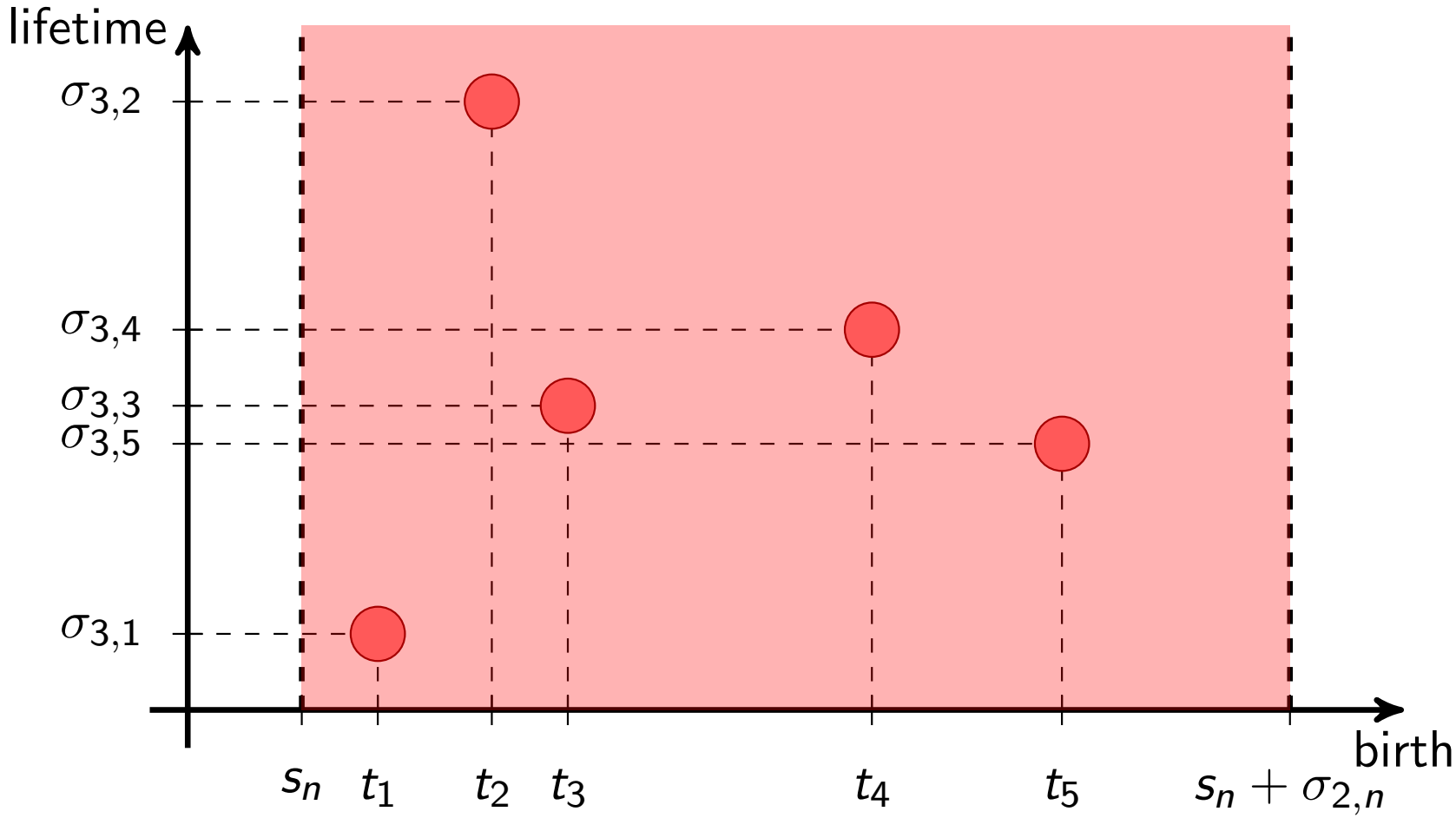
$$\mathrm{var}(M) = \mathbb{E}[M] + 2\lambda_2^2 \delta_+ (1-\delta_+) \cdot \cdot \int_0^{+\infty} \int_{-u}^{0} e^{-\Lambda v} (1 - F_2(u))(1 - F_2(u+v)) \,\mathrm{d}u \,\mathrm{d}v$$

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

Proteins

(t_k) *Poisson process of intensity* $\lambda_3 \, \mathrm{d}x$
$(\sigma_{3,k})$ *i.i.d. with distribution* $F_3(\mathrm{d}y)$

Proteins



(t_k) *Poisson process of intensity* $\lambda_3 \, \mathrm{d}x$
$(\sigma_{3,k})$ *i.i.d. with distribution* $F_3(\mathrm{d}y)$

Proteins

Proteins:

MPPP: $\mathcal{P} = \left(s_n, \sigma_{2,n}, \mathcal{N}_{\lambda_3}^{s_n}\right)$

Mark: $\left(\sigma_{2,n}, \mathcal{N}_{\lambda_3}^{s_n}\right)$ *and* $\mathcal{N}_{\lambda_3}^{s_n} = (t_k, \sigma_{3,k})$

Proteins at equilibrium:

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

Proteins: general results

Proposition

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

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

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

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

Strategy:

- description of proteins P at equilibrium via MPPP;

Strategy:

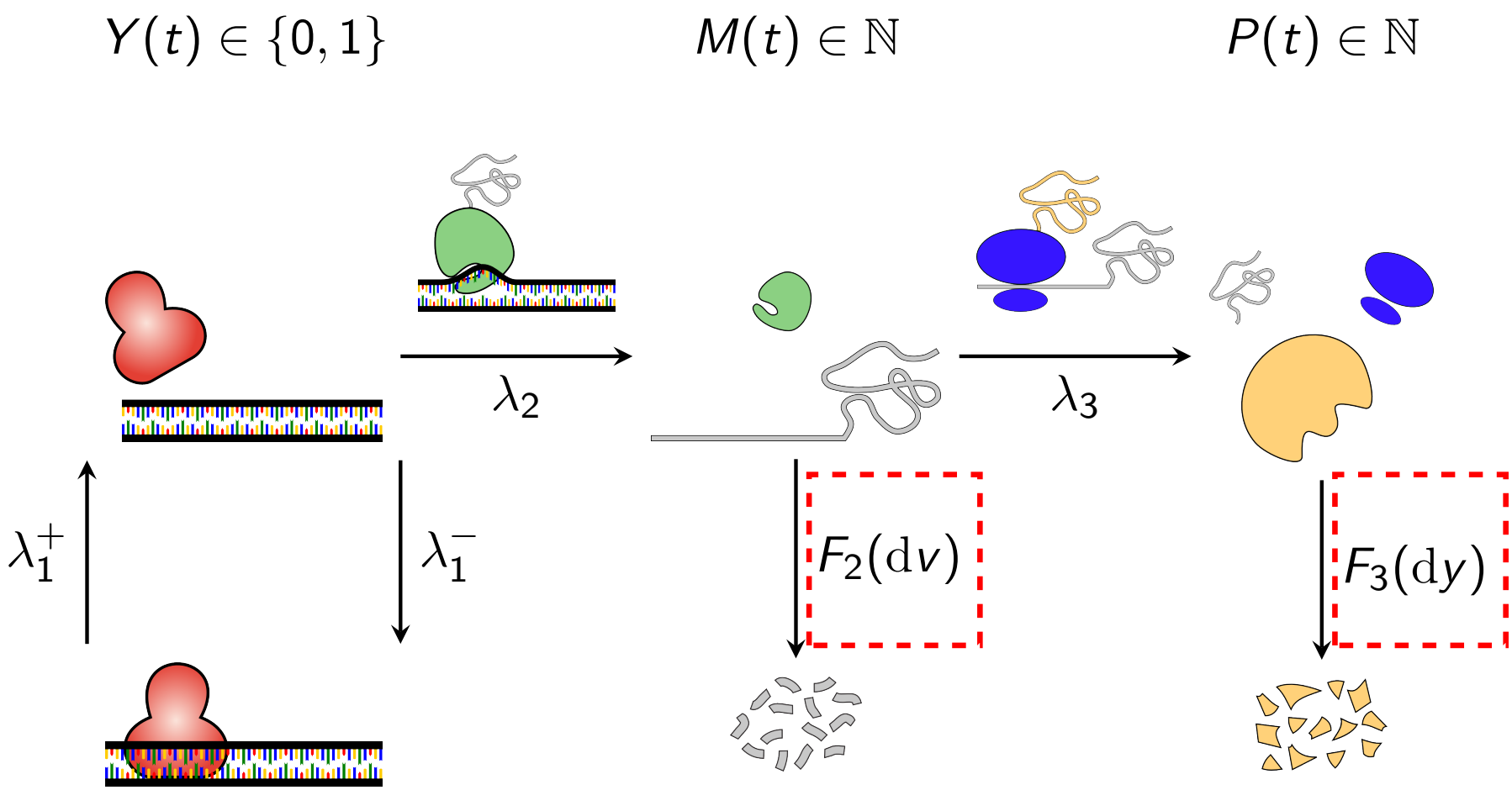
- description of proteins P at equilibrium via MPPP;
- Laplace transform of P;

Strategy:

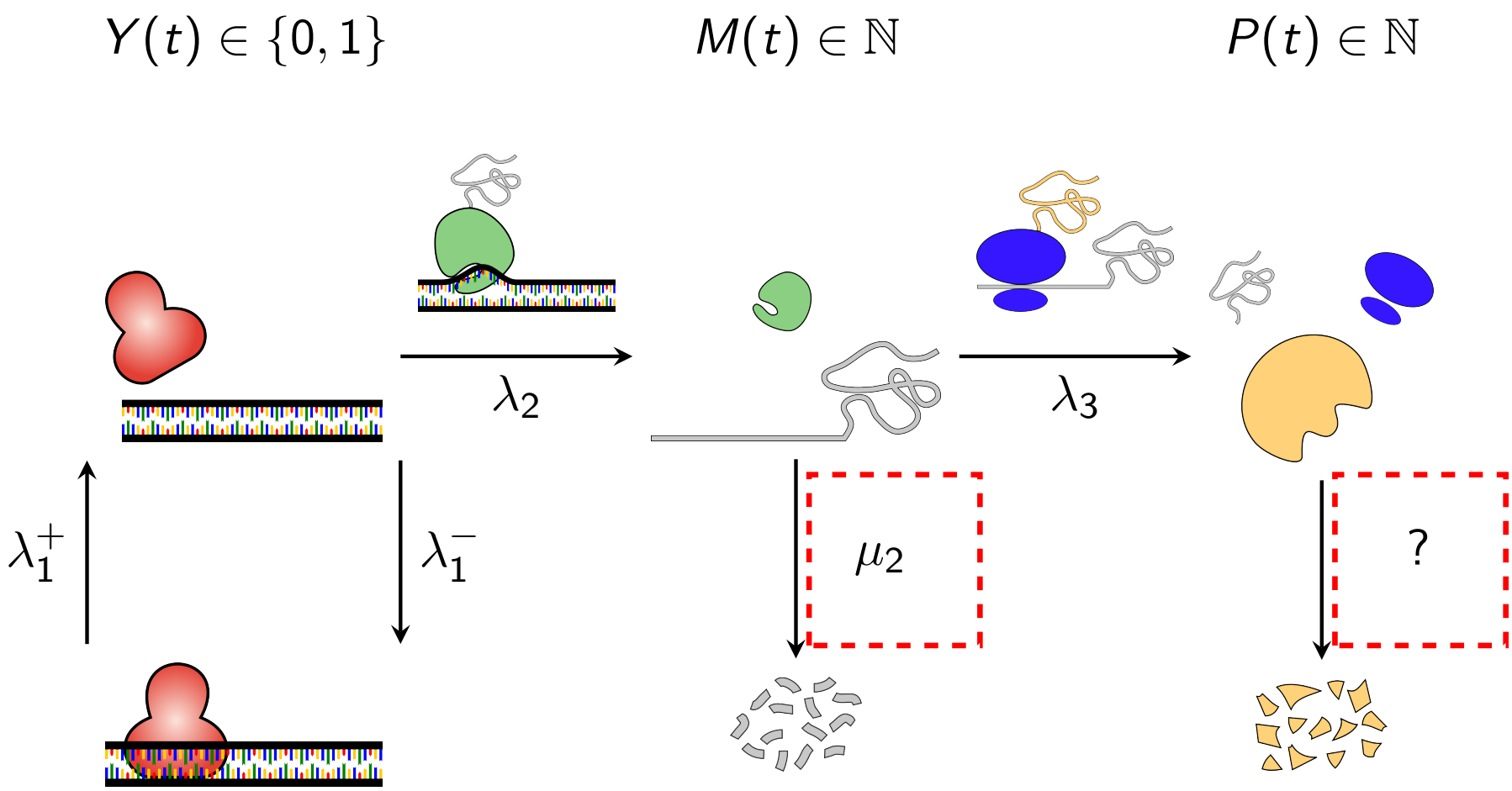
- description of proteins P at equilibrium via MPPP;
- Laplace transform of P;
- protein mean and variance formula for any distributions F_2, F_3.

MPPP Applications & Model Extensions

MPPP 3-Stage Model



MPPP 3-Stage Model



Protein degradation

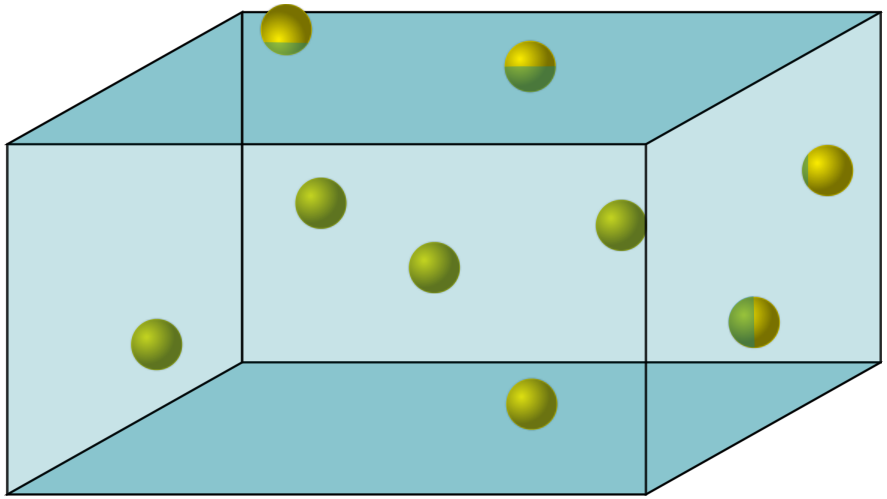
Two main mechanisms:

Proteolysis

Protein dilution

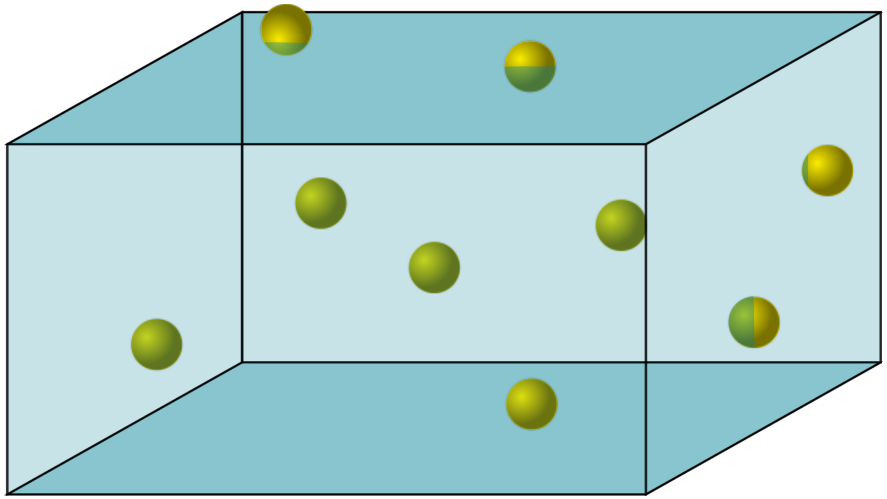
Protein dilution

- deterministic phenomenon
- continuous leaking of proteins



Protein dilution

- deterministic phenomenon
- continuous leaking of proteins



Result: the MPPP approach allows to consider this more accurate description and to compute statistics

Summary 3-Stage MPPP Model

- (MPPP) appropriate mathematical tool to describe gene expression;

Summary 3-Stage MPPP Model

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;

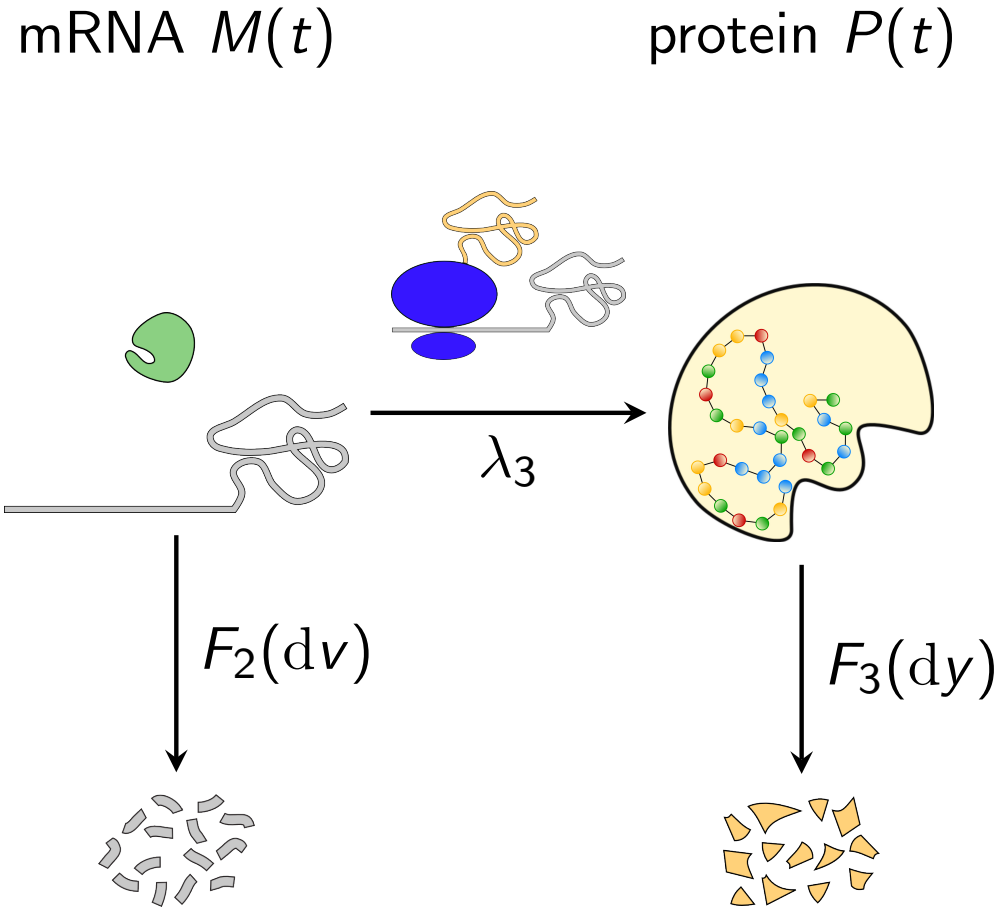
Summary 3-Stage MPPP Model

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- explicit formula depending on model parameters for specific assumptions, *i.e.* classic models as special case;

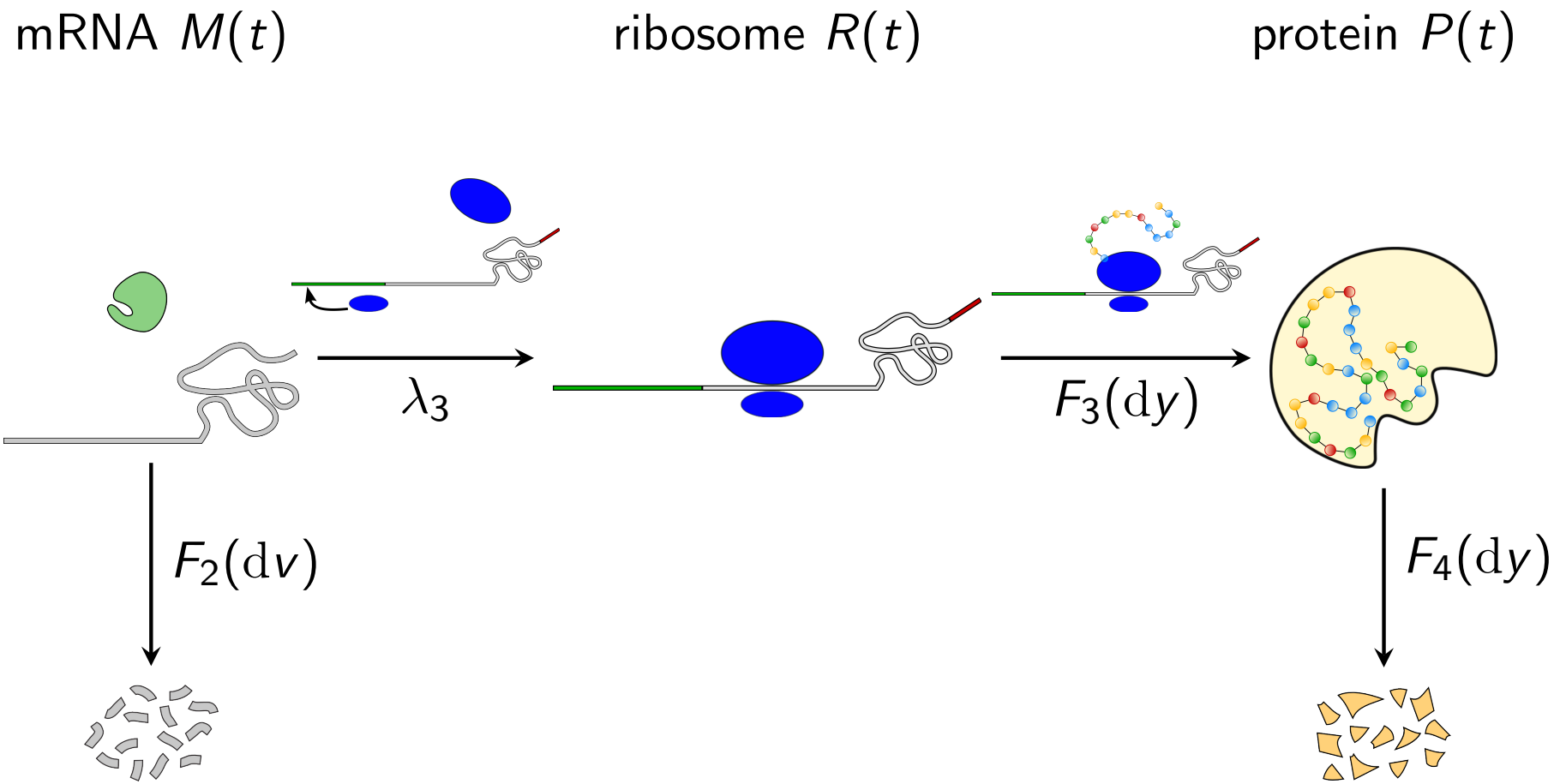
Summary 3-Stage MPPP Model

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- explicit formula depending on model parameters for specific assumptions, *i.e.* classic models as special case;
- possibility to describe realistic phenomena such as protein dilution.

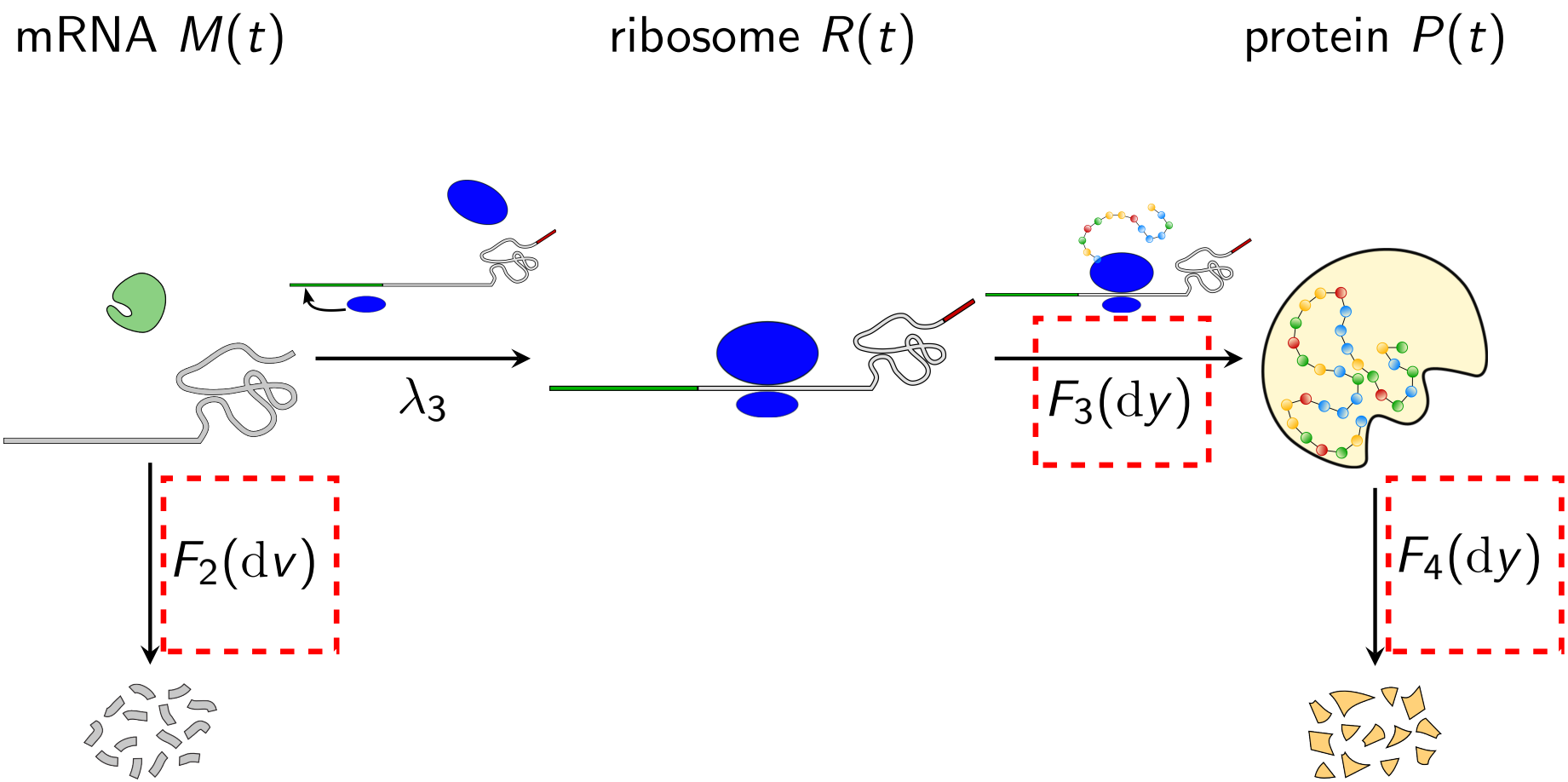
4-Stage Model:



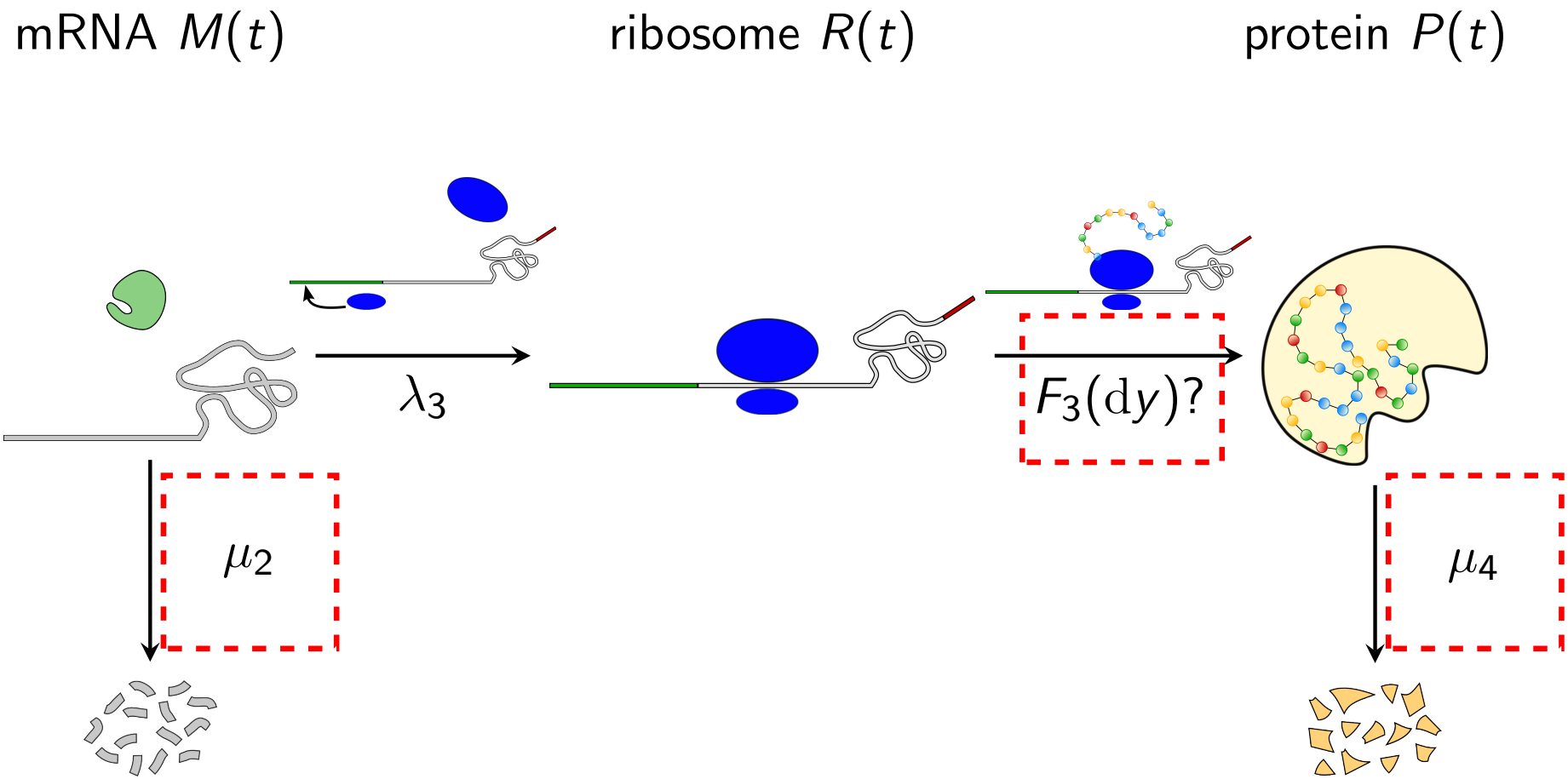
4-Stage Model:



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:

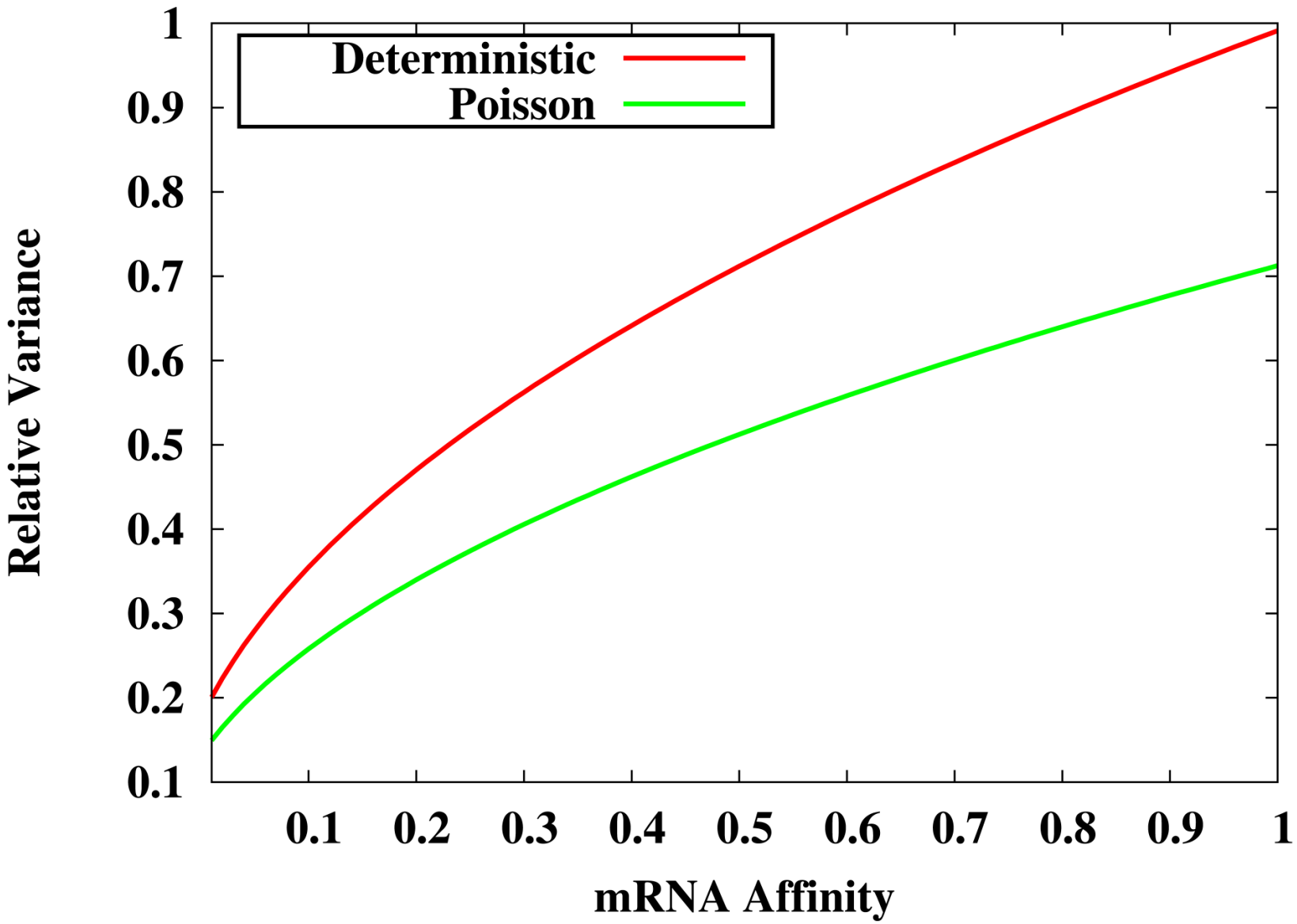
4-Stage model with deterministic elongation

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

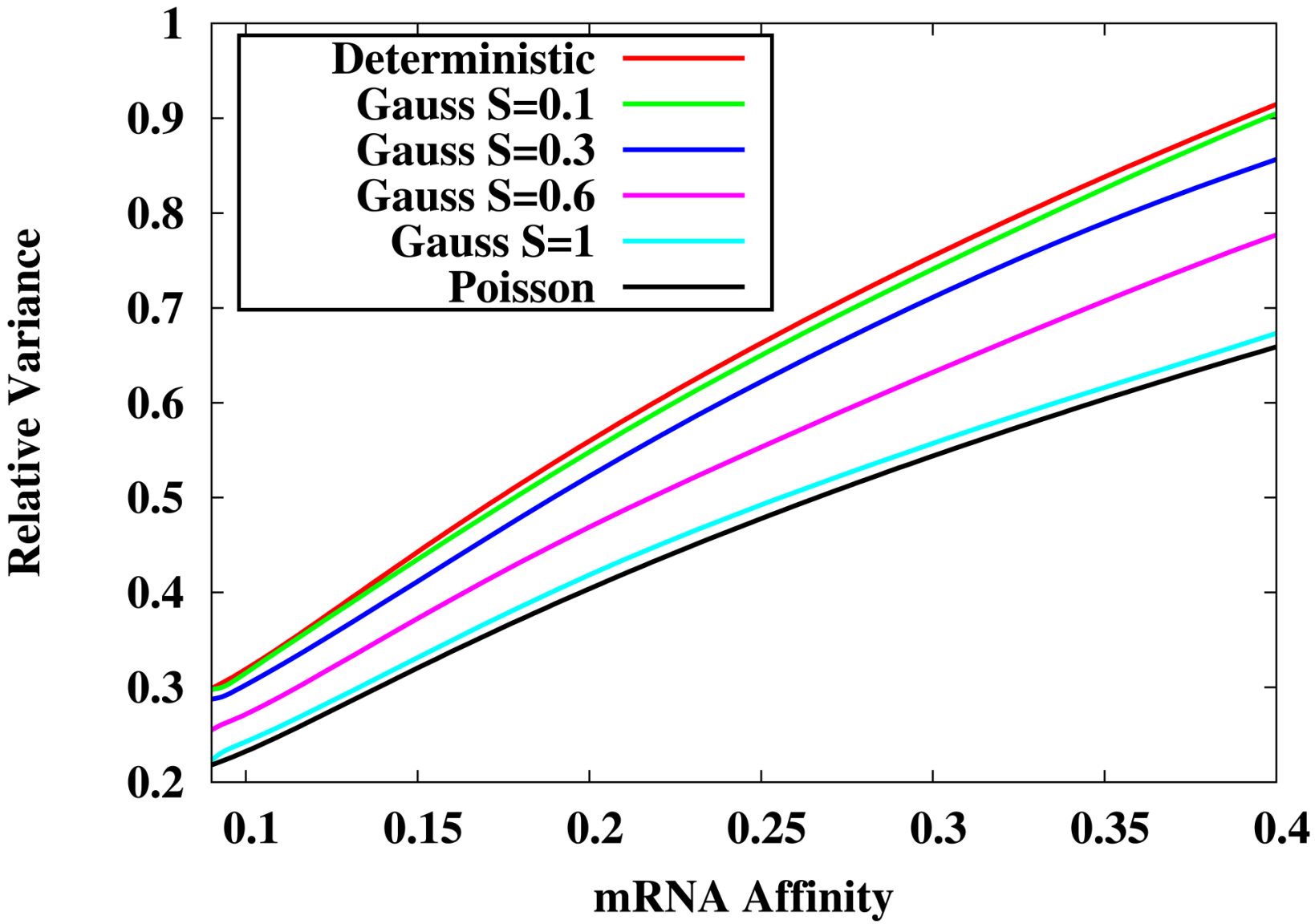
4-Stage model with exponential elongation

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

Deterministic vs Exponential



Deterministic vs Exponential



Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- averages independent of the chosen distribution;

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- averages independent of the chosen distribution;
- explicit formula depending on the model parameters for specific assumptions;

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- averages independent of the chosen distribution;
- explicit formula depending on the model parameters for specific assumptions;
- appropriate modeling of protein volume dilution;

Conclusions

- (MPPP) appropriate mathematical tool to describe gene expression;
- analytic formula for protein mean and variance for any distribution of protein/mRNA degradation;
- averages independent of the chosen distribution;
- explicit formula depending on the model parameters for specific assumptions;
- appropriate modeling of protein volume dilution;
- counter-intuitive: $\mathrm{var}_{\mathrm{DET}}(P) > \mathrm{var}_{\mathrm{EXP}}(P)$.

Conclusions

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.

Future work: multi-protein production

- polymerases: shared by all genes
  ribosomes: shared by all mRNAs;

Future work: multi-protein production

- polymerases: shared by all genes
  ribosomes: shared by all mRNAs;
- competition for polymerases/ribosomes;

Future work: multi-protein production

- polymerases: shared by all genes
  ribosomes: shared by all mRNAs;
- competition for polymerases/ribosomes;
- math: mean field approximation / stochastic averaging;

Future work: multi-protein production

- polymerases: shared by all genes
  ribosomes: shared by all mRNAs;
- competition for polymerases/ribosomes;
- math: mean field approximation / stochastic averaging;
- simulations.

Thanks.