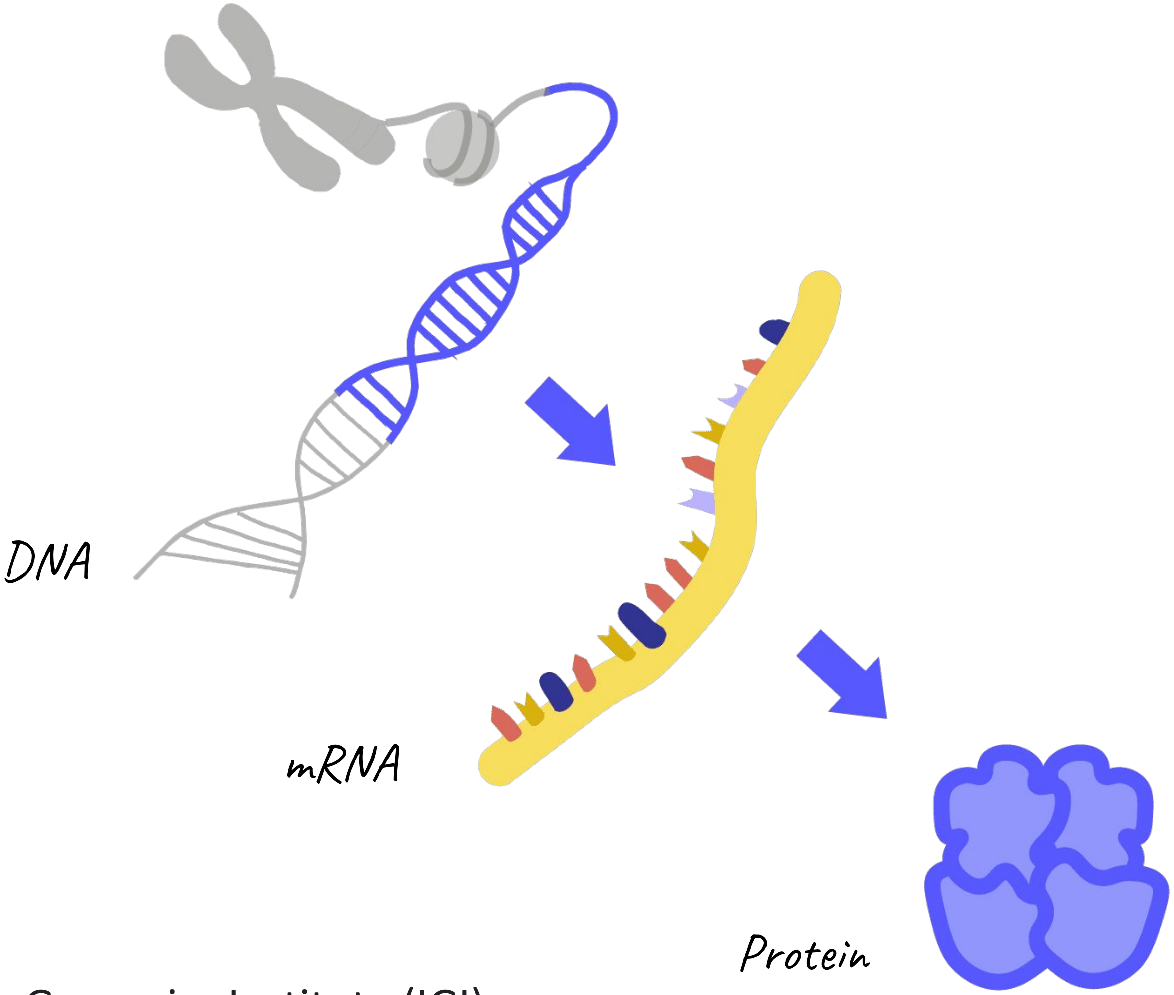
Stochastic Gene Expression Models

Physical Models of Living Systems, A.Y. 2022/2023

29/03/2023

Paolo Zinesi
paolo.zinesi@studenti.unipd.it
https://github.com/PaoloZinesi

Gene Expression



Innovative Genomics Institute (IGI)

Gene Expression

Experimental facts:

- Typical mRNA half-lives are in the range of 1-30 min
- Typical protein half-lives are in the range of some hours
- On average there are 1-30 mRNA molecules per cell
- On average there are 10^2-10^4 proteins for each mRNA
- Gene expression is stochastic, because the chemical reactions are randomly timed and the number of mRNA/proteins per cell is small

→ Stochastic gene expression models

Gene Expression

Parameters:

ν_0 : transcription rate

d_0 : mRNA degradation (or death rate)

ν_1 : translation rate

d_1 : protein degradation (or death rate)

$a = \nu_0/d_1$: mRNAs transcripted in a protein lifetime

$b = \nu_1/d_0$: proteins translated in an mRNA lifetime

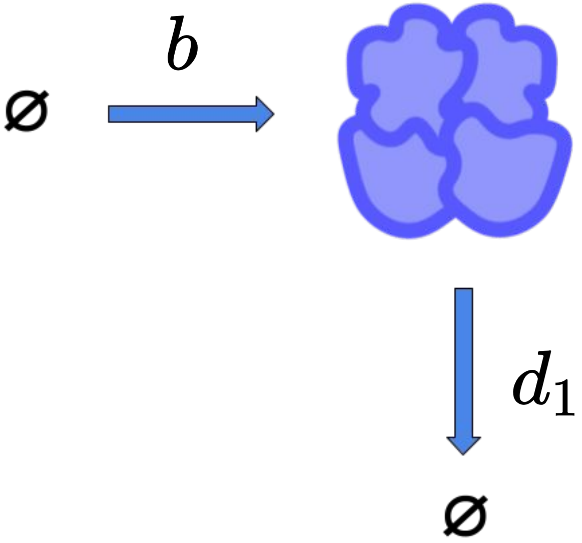
$\gamma = d_0/d_1$: mRNAs degraded in a protein lifetime

One Stage Model

- ➔ First approach to model the dynamics of the protein number
  - ◆ Birth-death process on the number of proteins
  - ◆ Neutral Theory approach

$$
\begin{aligned}
n &= \text{number of proteins} \\
m &= \text{number of mRNAs}
\end{aligned}
$$

$$
\begin{cases}
b_n \approx b = \langle m \rangle \, \nu_1 = \dfrac{\nu_0 \, \nu_1}{d_0} \\
d_n = d_1 \cdot n
\end{cases}
$$



One Stage Model

$$\boxed{\dot{P}_n = b_{n-1}P_{n-1} + d_{n+1}P_{n+1} - (b_n + d_n)P_n}$$

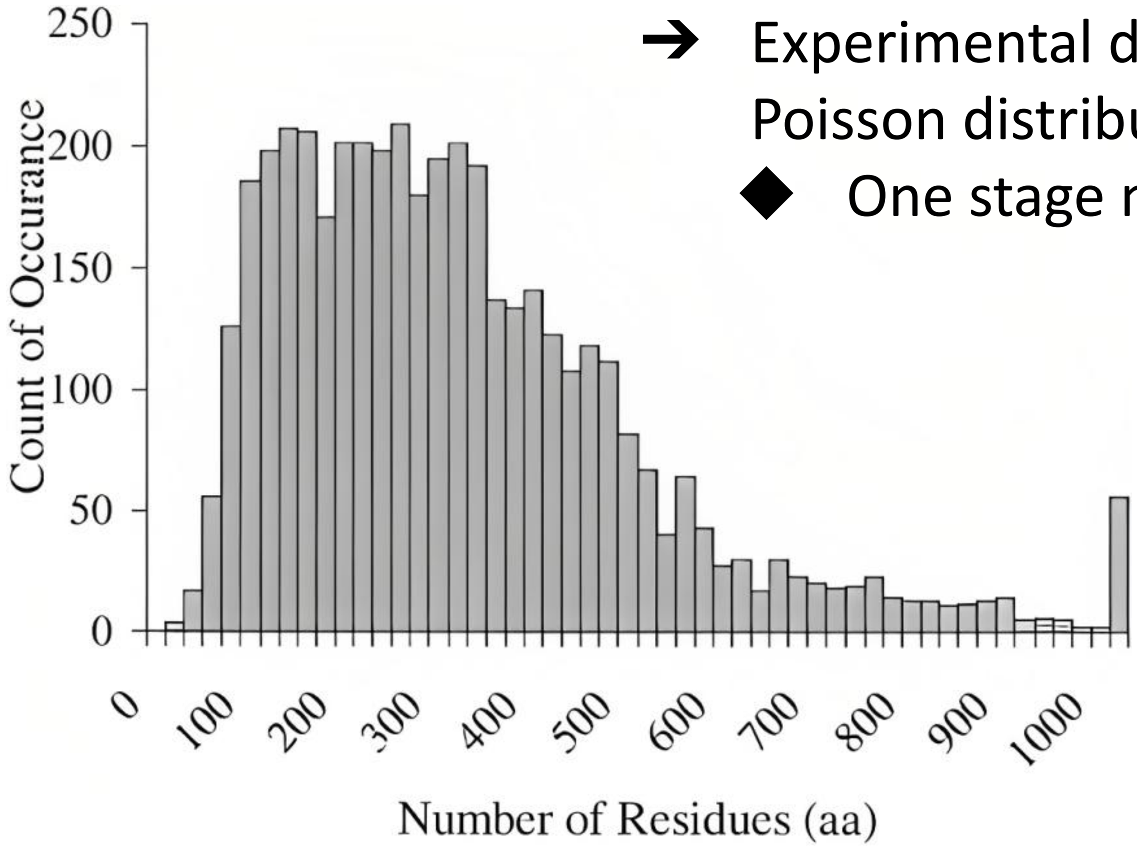
$P_n = P_n(t)$: protein number distribution

$$F(x,t) = \sum_{k=0}^{\infty} x^k P_k(t) = \exp\left(\frac{b}{d_1}(1 - e^{-d_1 t})(x-1)\right)$$

$$P_n(t) = \frac{\mu^n}{n!}e^{-\mu}, \text{ with } \mu = \frac{b}{d_1}(1 - e^{-d_1 t})$$

Poissonian !

One Stage Model



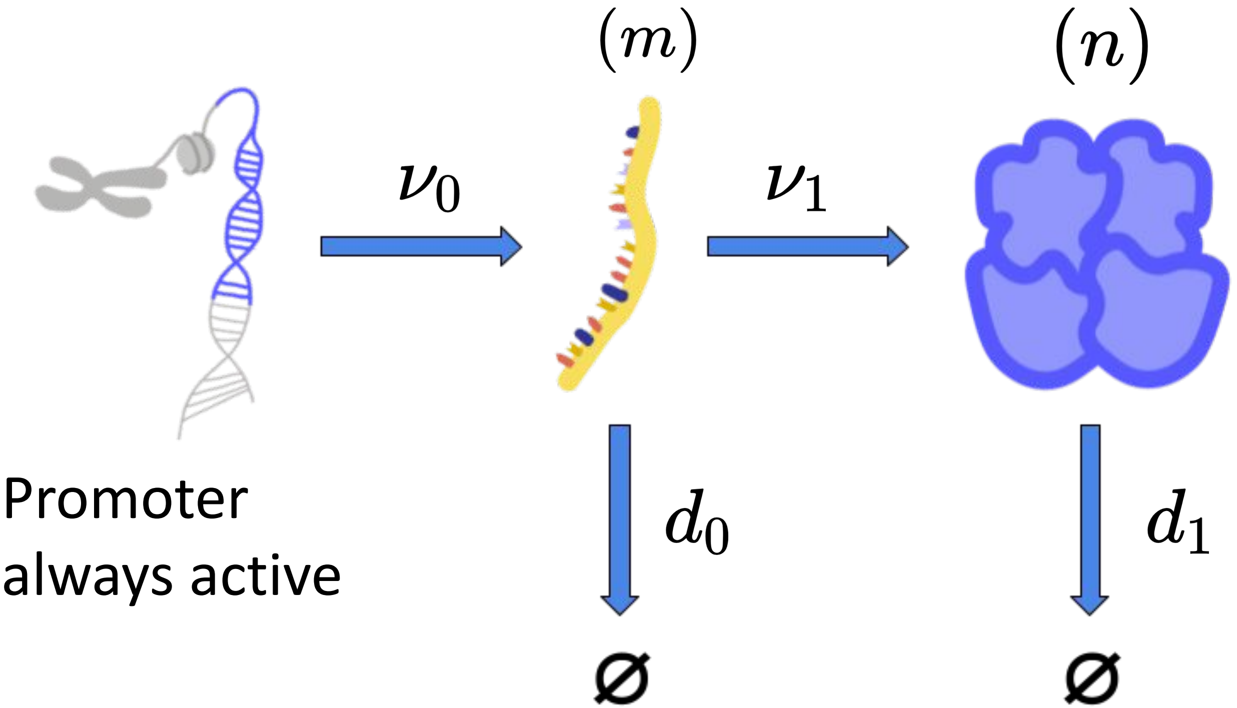
➔ Experimental data do not follow a Poisson distribution

- One stage model is not accurate

VALAFAR, H., PRESTEGARD, J.H. and VALAFAR, F. (2002), Datamining Protein Structure Databanks for Crystallization Patterns of Proteins. Annals of the New York Academy of Sciences, 980: 13-22.

Two Stage Model

- → The mRNA plays a fundamental role in gene expression
  - ◆ mRNA controls the protein bursts
  - ◆ mRNA and protein dynamics are coupled



Promoter
always active

Two Stage Model - MF

Mean Field:

$$\begin{cases} \dot{m} = \nu_0 - d_0 \cdot m \\ \dot{n} = \nu_1 \cdot m - d_1 \cdot n \end{cases} \Rightarrow \begin{cases} m^\star = \dfrac{\nu_0}{d_0} \\ n^\star = \dfrac{\nu_0\, \nu_1}{d_0\, d_1} \end{cases}$$

➔ Experimental data do not agree with the mean field model

- ◆ Protein number fluctuates a lot and follows the MF predictions only on average
- ◆ Protein number is small and it cannot be treated as a continuous variable

Two Stage Model - Stochastic

$$\boxed{\begin{aligned}\dot{P}_{m,n} = {} & \nu_0(P_{m-1,n} - P_{m,n}) + \\ & + \nu_1 m(P_{m,n-1} - P_{m,n}) + \\ & + d_0[(m+1)P_{m+1,n} - m\,P_{m,n}] + \\ & + d_1[(n+1)P_{m,n+1} - n\,P_{m,n}]\end{aligned}}$$

$\downarrow$ $\gamma \gg 1$ limit

$$F(z, z', t) \simeq F(z,t) = \left[\frac{1 - b(z-1)e^{-d_1 t}}{1 + b - bz}\right]^a$$

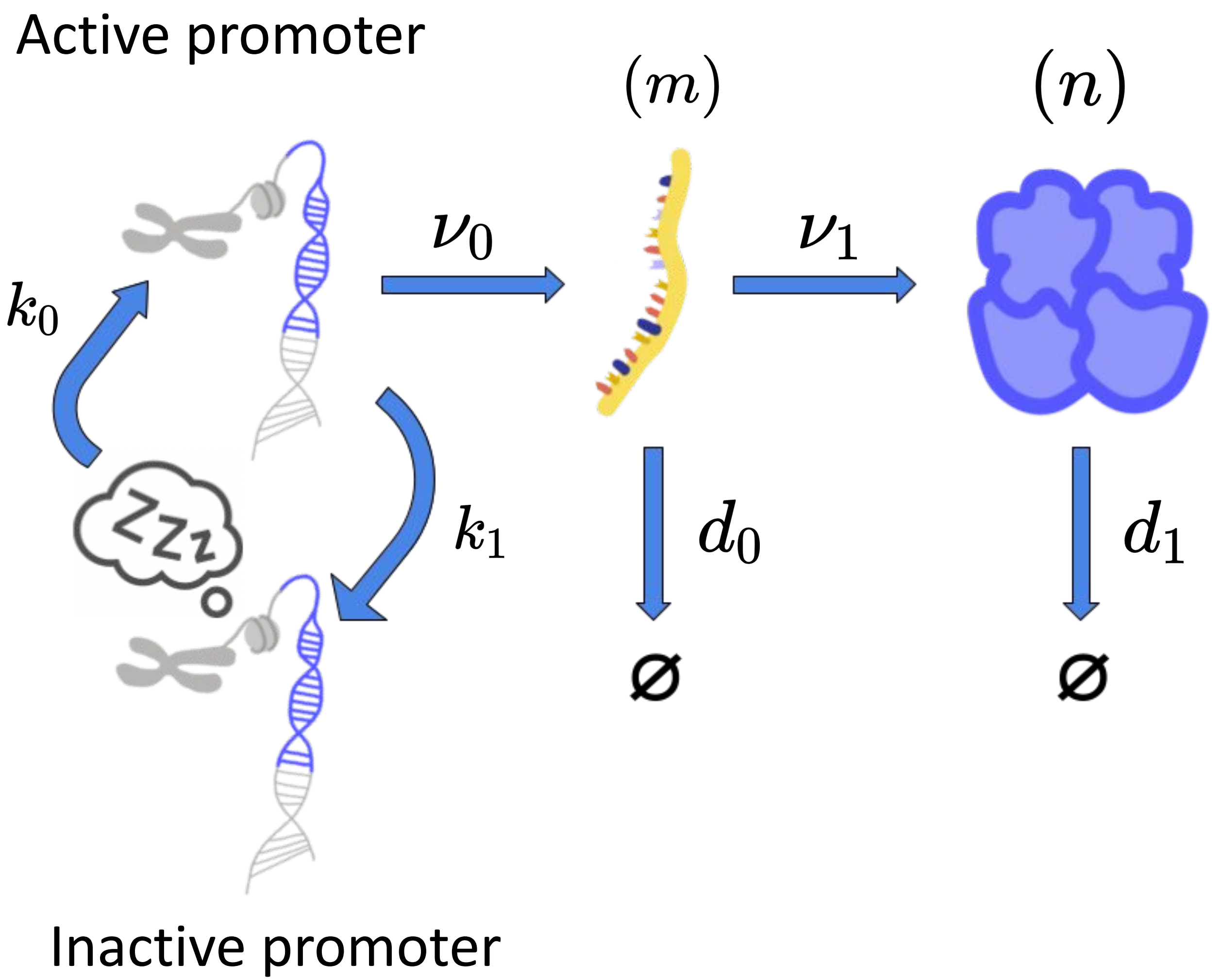
Two Stage Model - Stochastic

$$P_n = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)} \left(\frac{b}{1+b}\right)^n \left(1 - \frac{b}{1+b}\right)^a$$

$P_n \simeq P_{0,n}$: protein number stationary distribution

- ➔ **In agreement with experimental data!**
- ➔ The mode of the stationary distribution is at zero mRNAs (even if the mean number of mRNAs is greater than zero)
- ➔ This stationary distribution can be obtained more easily by explicitly modeling the process of <u>protein bursts</u>

Three Stage Model



Three Stage Model

- ➔ More general model, but it tends to the two stage model in the limit of fast-switching active/inactive states

$\kappa_0 = k_0/d_1$: DNA activations in a protein lifetime

$\kappa_1 = k_1/d_1$: DNA deactivations in a protein lifetime

$$P_n \to \frac{\Gamma(\beta+n)}{\Gamma(n+1)\Gamma(\beta)}\left(\frac{b}{1+b}\right)^n\left(1-\frac{b}{1+b}\right)^\beta$$

when $\kappa_0, \kappa_1 \gg 1$ but κ_0/κ_1 is fixed

$\beta = \beta(a, \kappa_0, \kappa_1)$

Gillespie Algorithm

- Each chemical reaction in a well-stirred environment can be completely characterized by the quantities:
  - The elements **concentration** $x_i = x_i(t)$
  - A **state-change vector** $\vec{v}_j$
  - A **propensity** (or rate) $a_j(\vec{x})$

- Example:

$$j : (m, n) \xrightarrow{\nu_1} (m, n+1)$$

$$\vec{x} = (m, n)$$

$$\vec{v}_j = (0, +1)$$

$$a_j(\vec{x}) = m\,\nu_1$$

Gillespie Algorithm

- ➔ Probability that the **j**-th reaction occurs after a time **τ**:

$$p(\tau, j \,|\, \vec{x}, t) = a_j(\vec{x}) \exp\left(-a_0(\vec{x})\tau\right)$$

$$\text{with } a_0(\vec{x}) = \sum_j a_j(\vec{x})$$

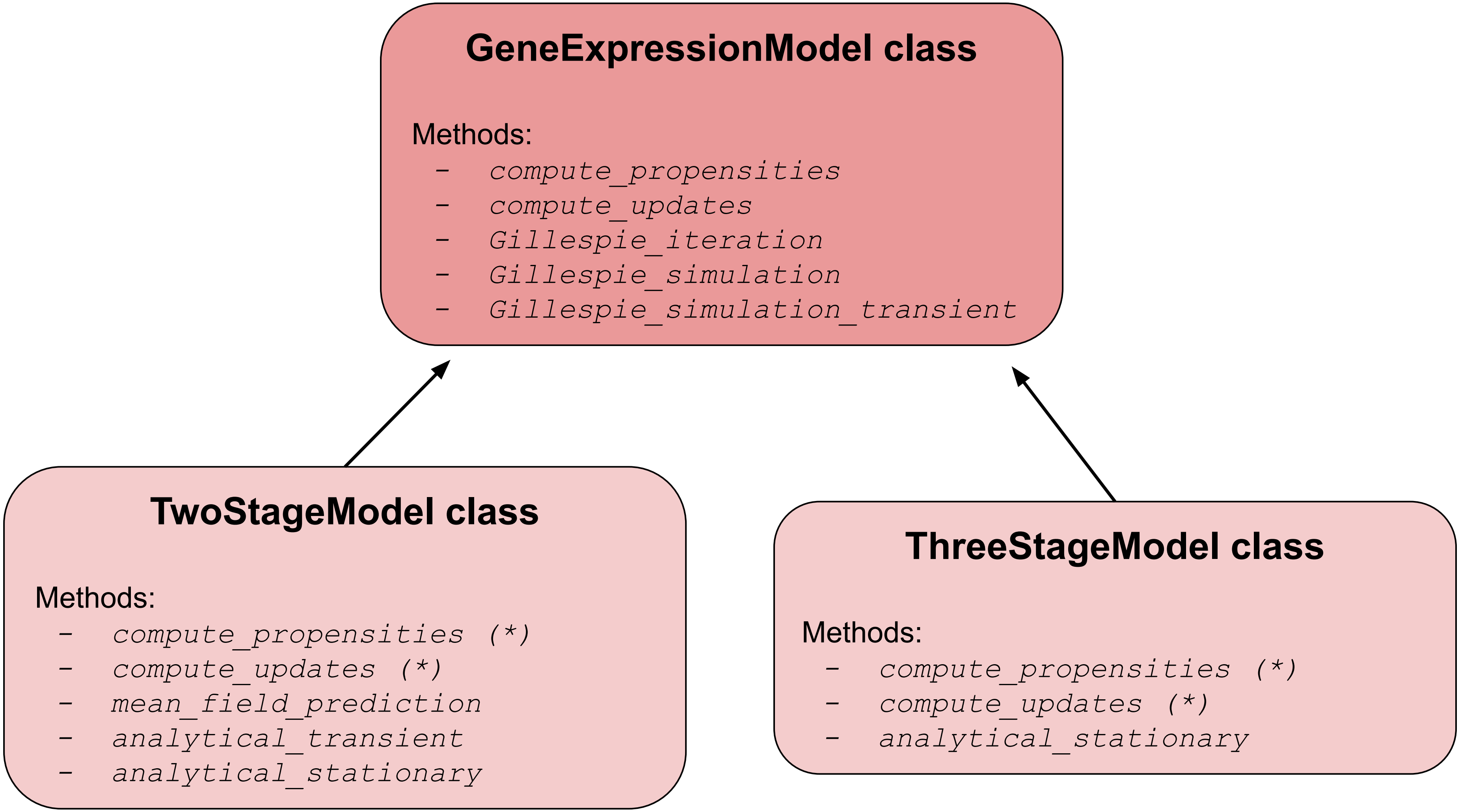
- ➔ Idea of the algorithm:
  - ◆ Extract the next reaction time **τ** from an exponential distribution
  - ◆ Choose reaction **j** with probability $a_j(\vec{x})/a_0(\vec{x})$

Gillespie Algorithm

Algorithm:

```
1. Initialize $\vec{x} = \vec{x}(t_0)$
2. Evaluate $a_j(\vec{x})$ and $a_0(\vec{x}) = \sum_j a_j(\vec{x})$
3. $\tau \sim \mathrm{Exp}(\tau \,|\, \lambda = a_0(\vec{x}))$
4. Extract $j$ according to probability $p(j) = a_j(\vec{x})/a_0(\vec{x})$
5. Update $\vec{x} \leftarrow \vec{x} + \vec{v}_j$
6. Go to step 2
```

“Genexpr” Library



GeneExpressionModel class

Methods:

- *compute_propensities*
- *compute_updates*
- *Gillespie_iteration*
- *Gillespie_simulation*
- *Gillespie_simulation_transient*

TwoStageModel class

Methods:

- *compute_propensities (*)*
- *compute_updates (*)*
- *mean_field_prediction*
- *analytical_transient*
- *analytical_stationary*

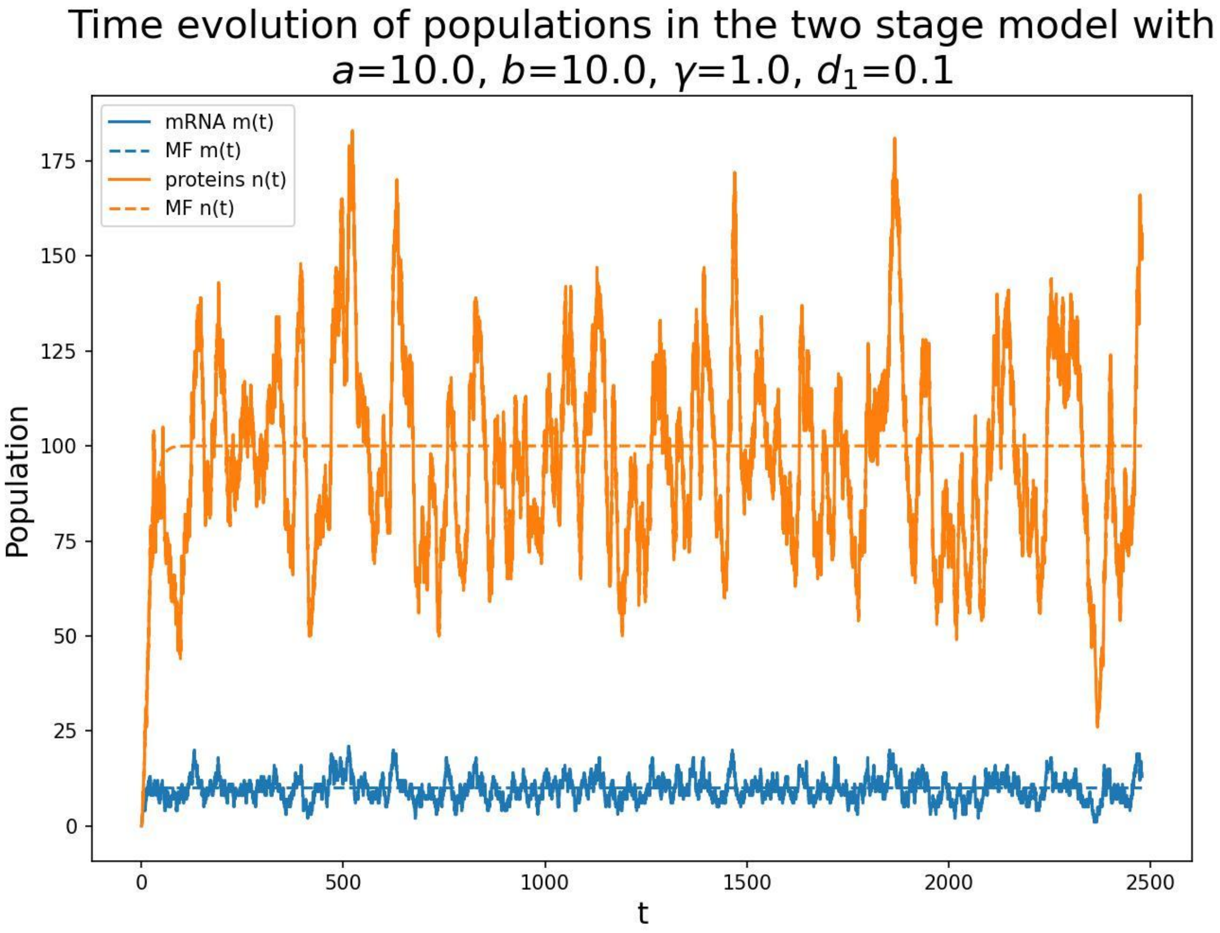
ThreeStageModel class

Methods:

- *compute_propensities (*)*
- *compute_updates (*)*
- *analytical_stationary*

(*) = overloaded

Simulation Results

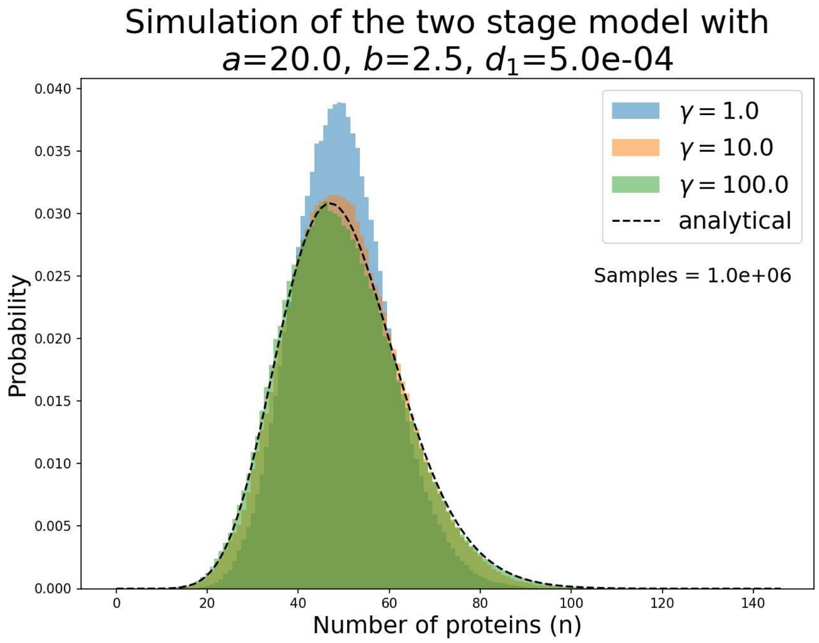


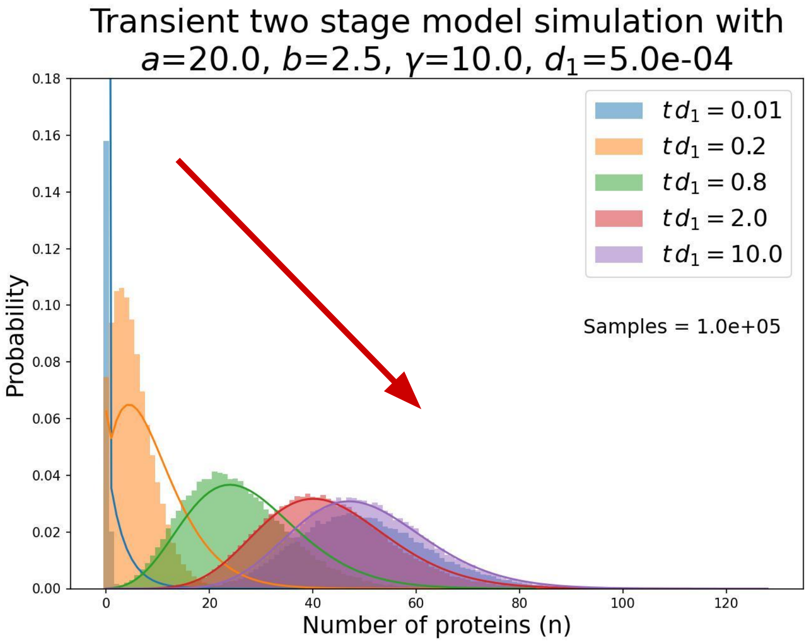
$$\langle n \rangle = a \cdot b = 100$$

$$\langle m \rangle = a / \gamma = 10$$

Simulation Results

UNIVERSITÀ DEGLI STUDI DI PADOVA

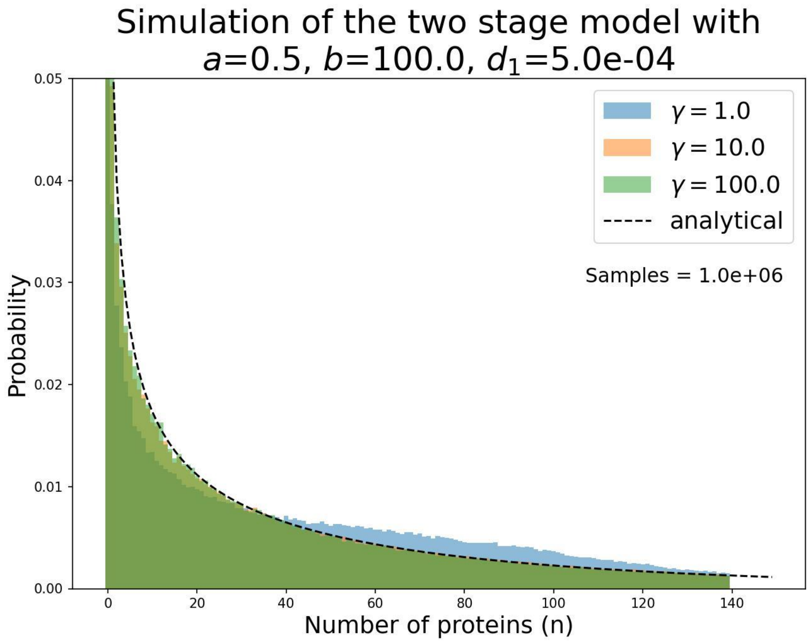

Simulation of the two stage model with a=20.0, b=2.5, d_1=5.0e-04


Transient two stage model simulation with a=20.0, b=2.5, γ=10.0, d_1=5.0e-04

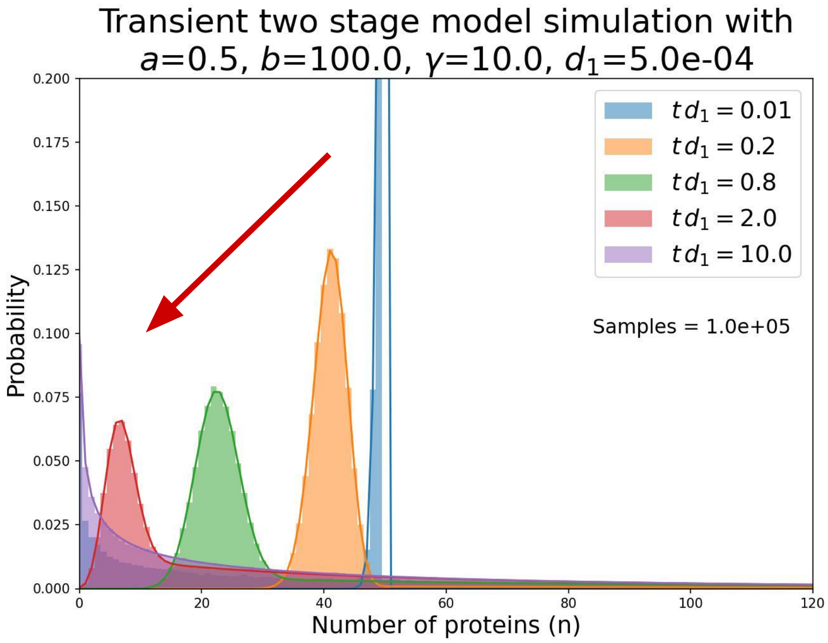
$$\langle m \rangle = a/\gamma = \{20, 2, 0.2\}$$

$$\langle n \rangle = a \cdot b = 50$$

Simulation Results



Simulation of the two stage model with a=0.5, b=100.0, d_1=5.0e-04

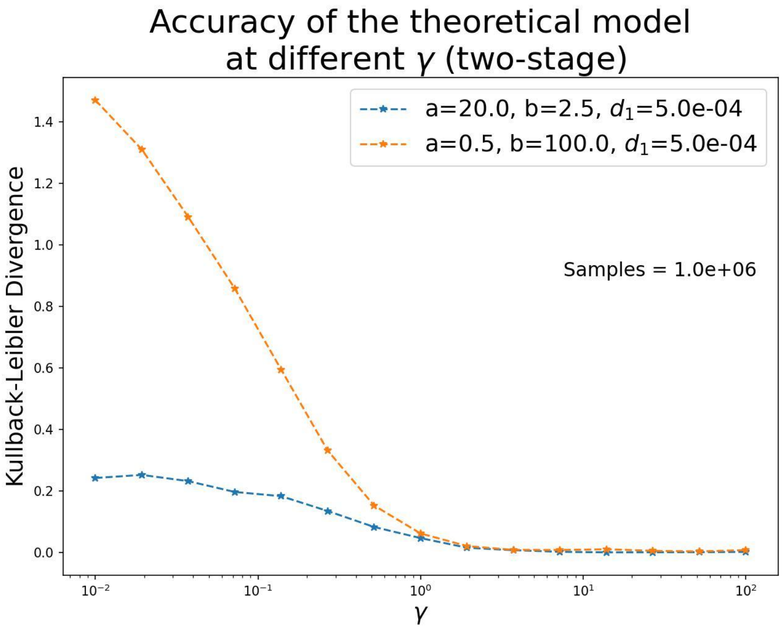


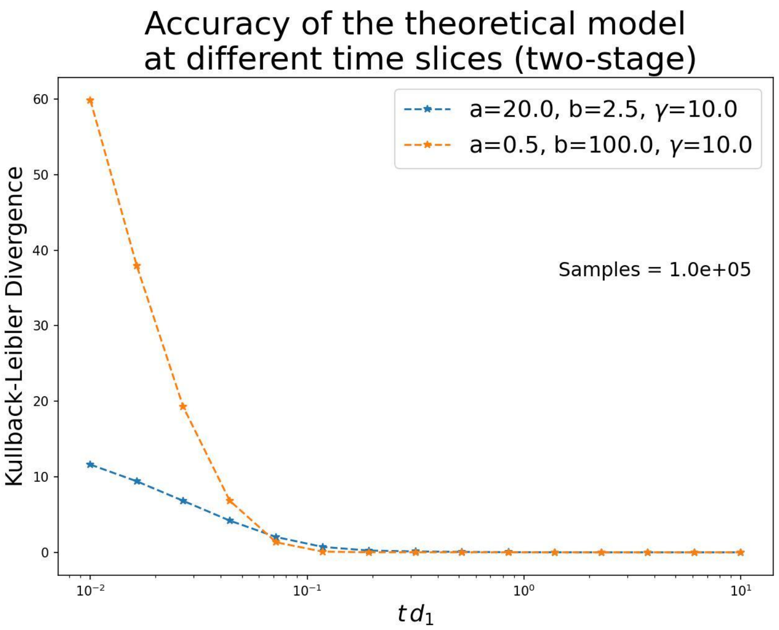
Transient two stage model simulation with a=0.5, b=100.0, γ=10.0, d_1=5.0e-04

$$\langle m \rangle = a/\gamma = \{0.5, 0.05, 0.005\}$$

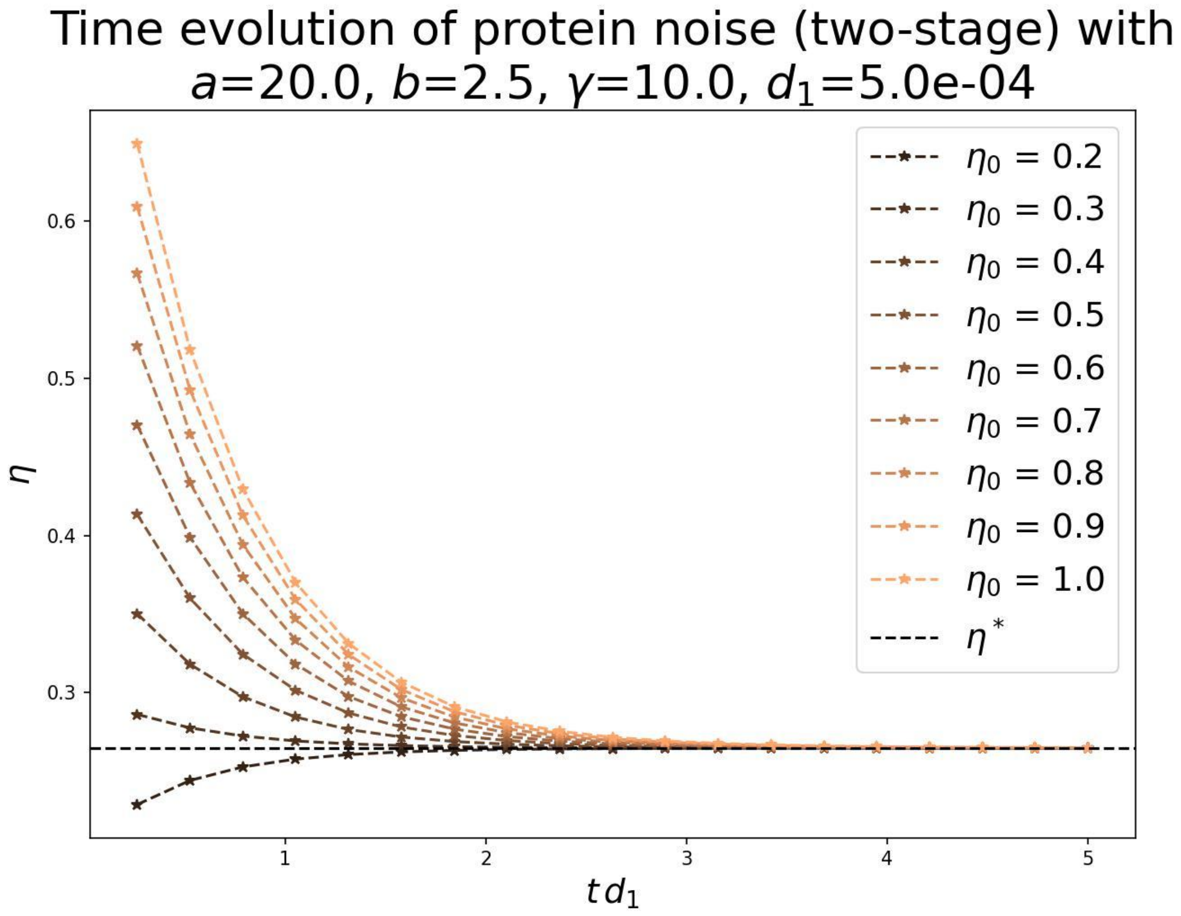
$$\langle n \rangle = a \cdot b = 50$$

Simulation Results

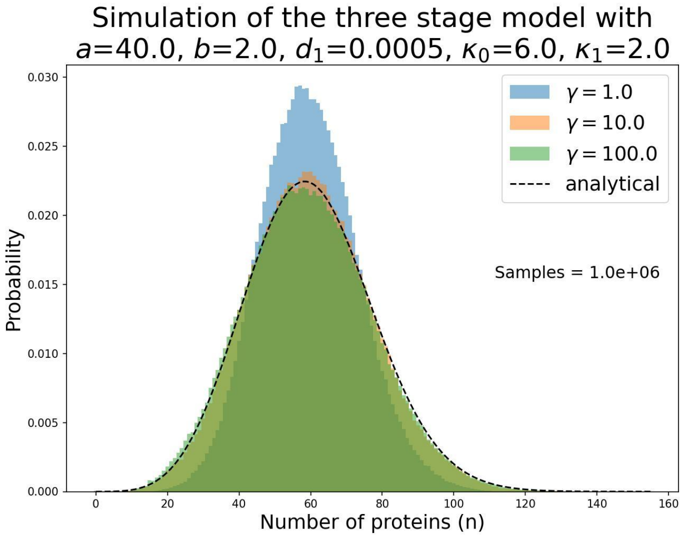


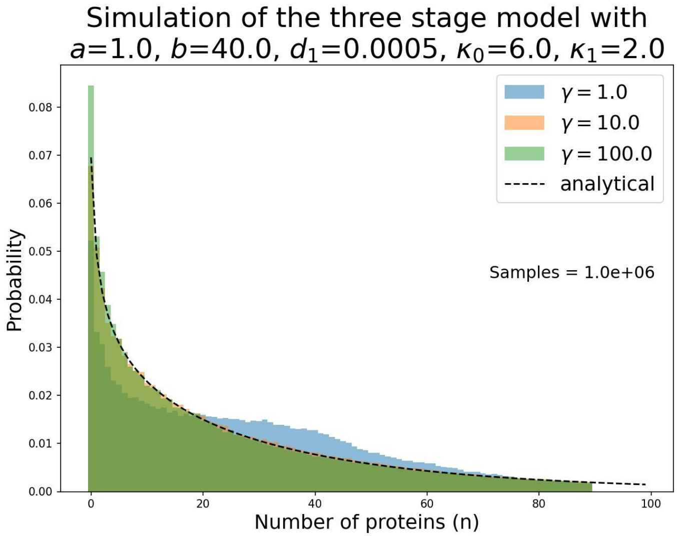


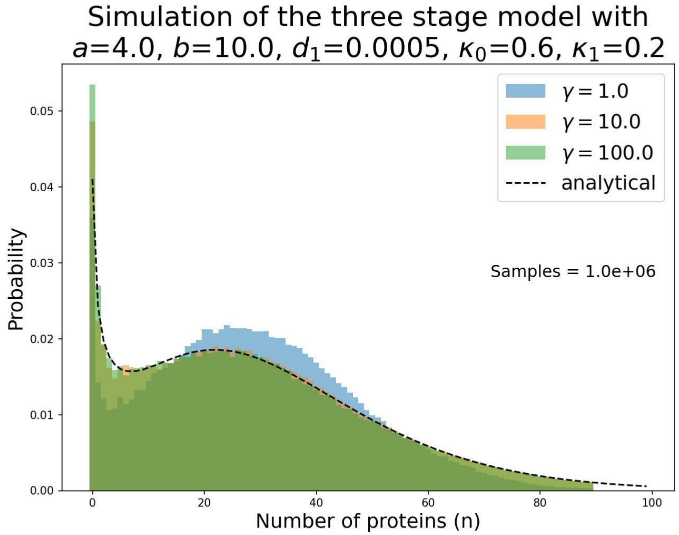
Simulation Results

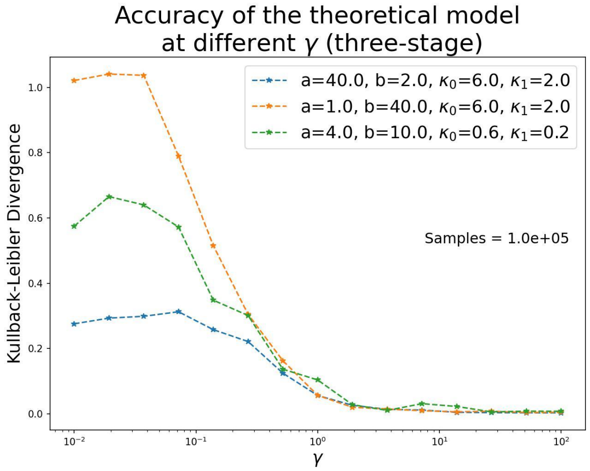


Simulation Results


Simulation of the three stage model with a=40.0, b=2.0, d_1=0.0005, κ_0=6.0, κ_1=2.0


Simulation of the three stage model with a=1.0, b=40.0, d_1=0.0005, κ_0=6.0, κ_1=2.0


Simulation of the three stage model with a=4.0, b=10.0, d_1=0.0005, κ_0=0.6, κ_1=0.2


Accuracy of the theoretical model at different γ (three-stage)

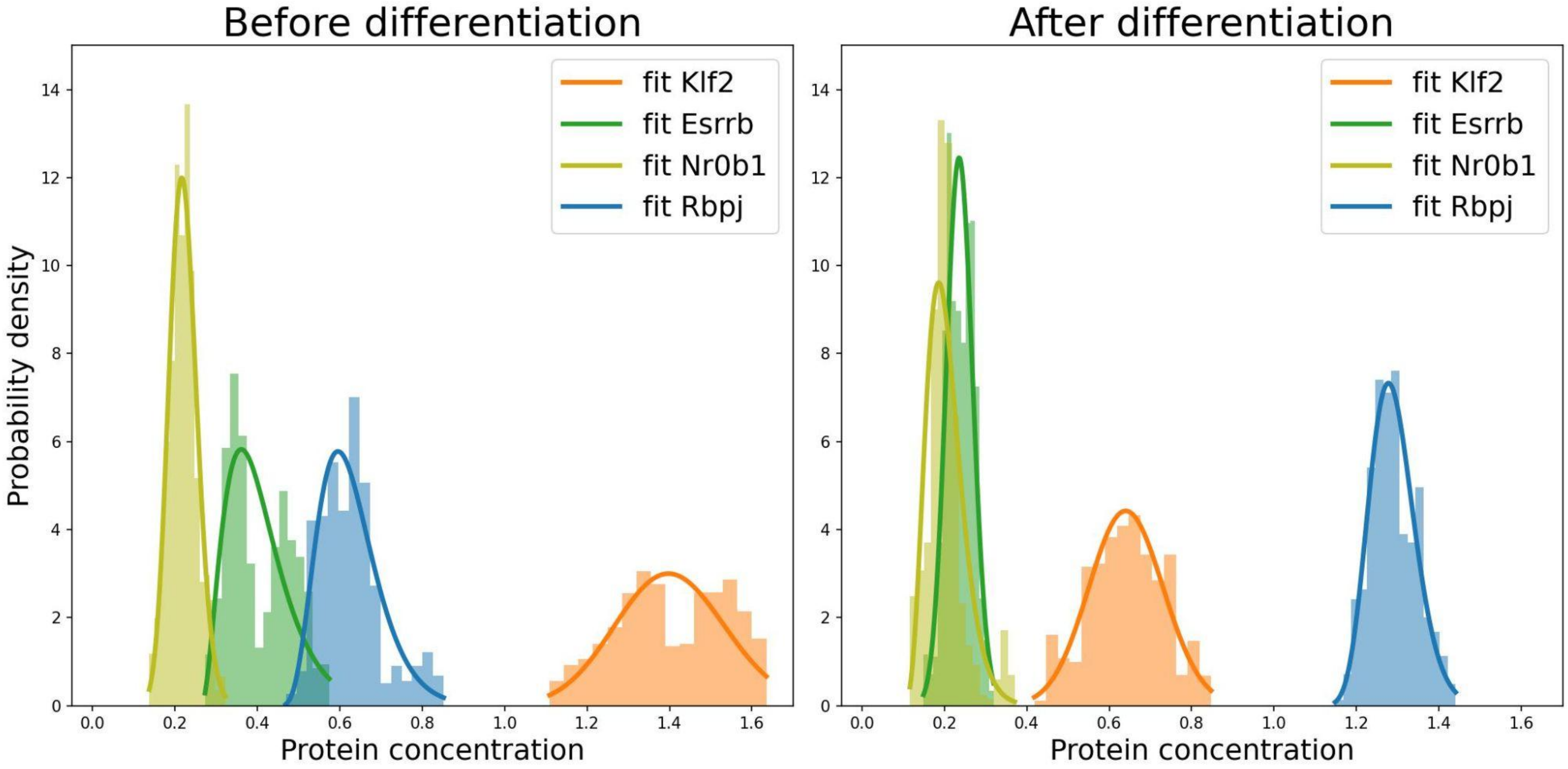
Application to real data

- A dataset containing the **concentrations of proteins produced by a stem cell** is analyzed
- The dataset captures the differentiation process of the cell
  - It contains **9547 time samples** of **24 proteins**

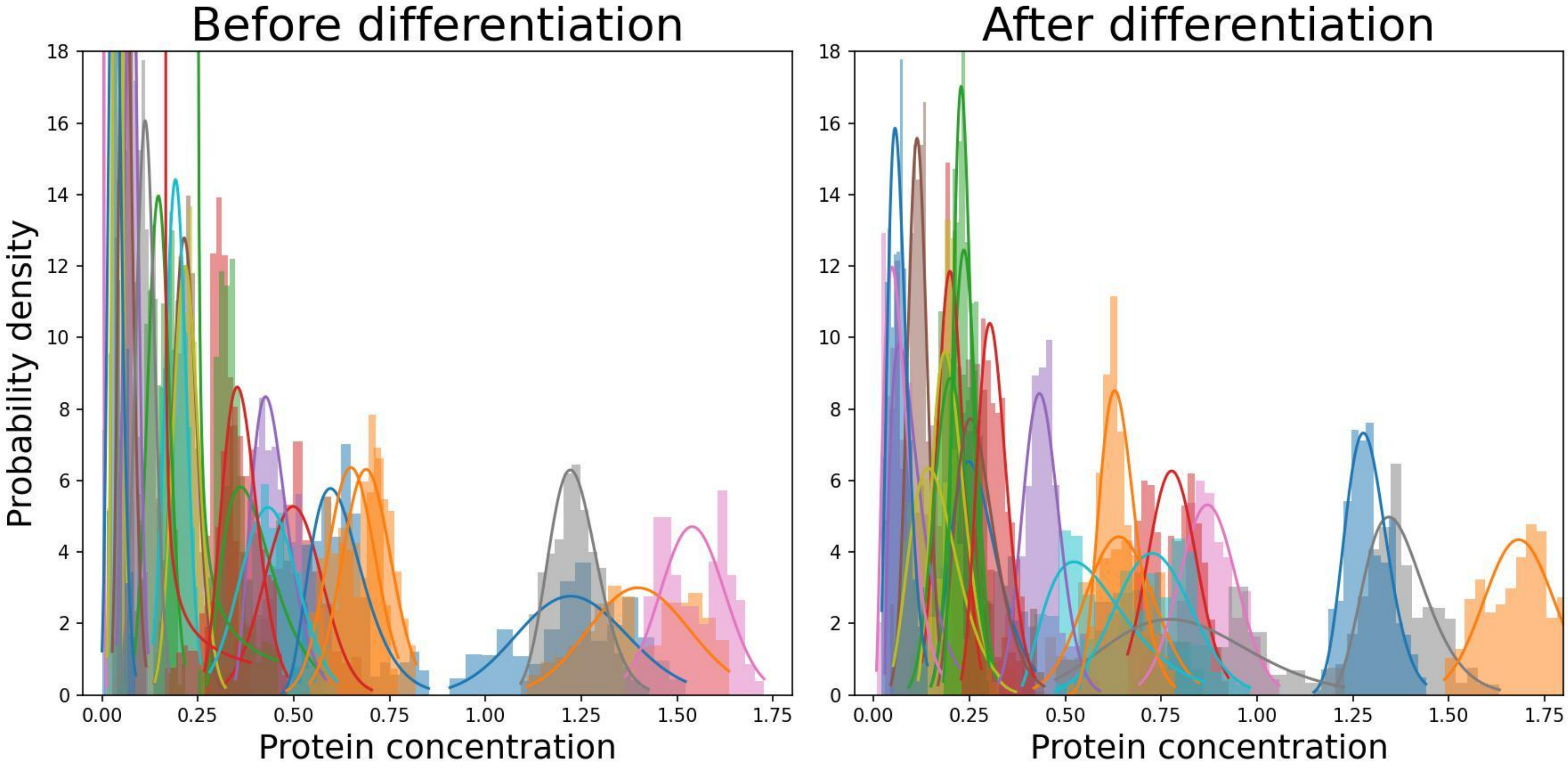
- Two dataset slices are considered:
  - Stationary distribution before differentiation, considering the first ⅙-th of the original dataset
  - Stationary distribution after differentiation, considering the last ⅙-th of the original dataset

- Protein distributions are fitted with a **Gamma distribution**

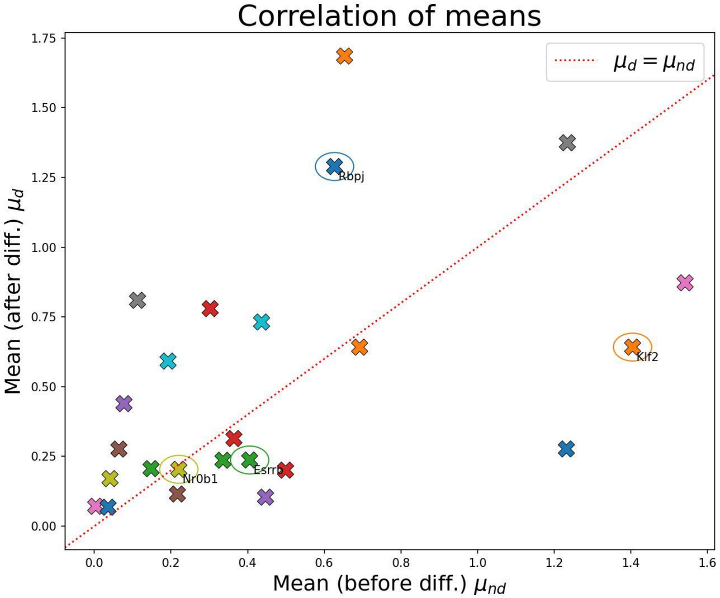
Application to real data

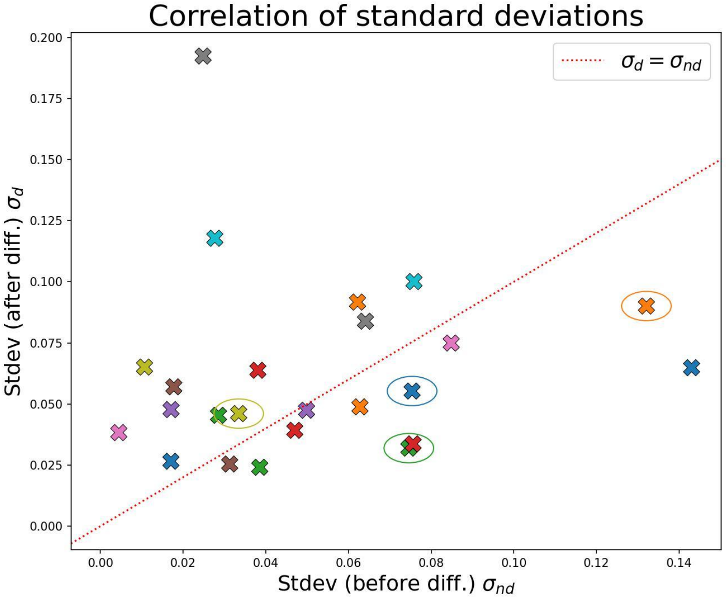


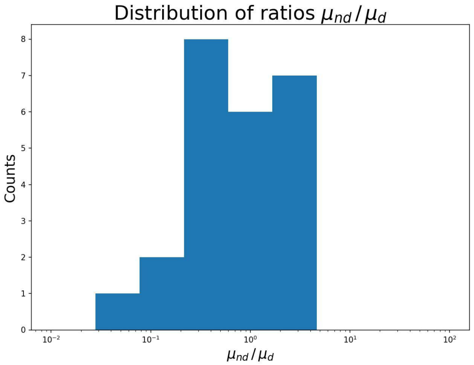
Application to real data

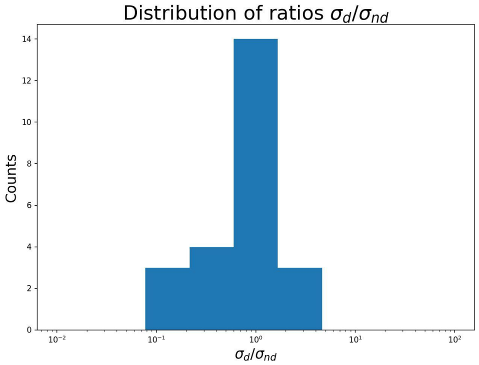


Application to real data

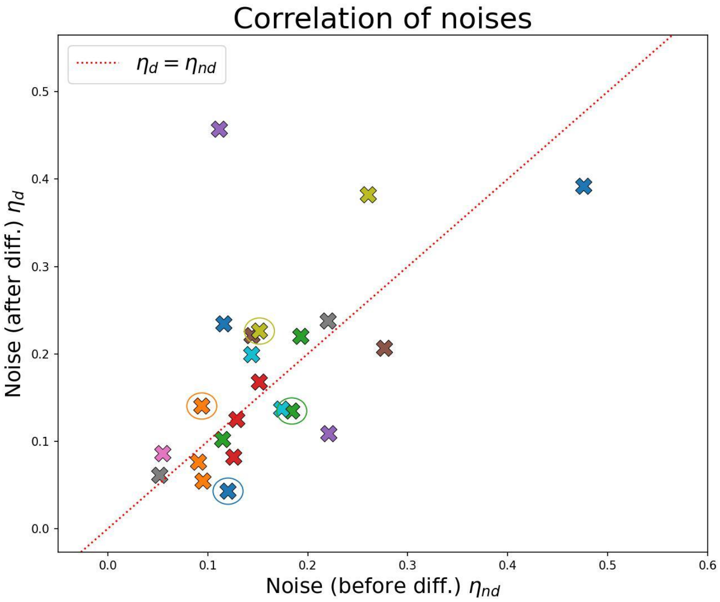

Correlation of means

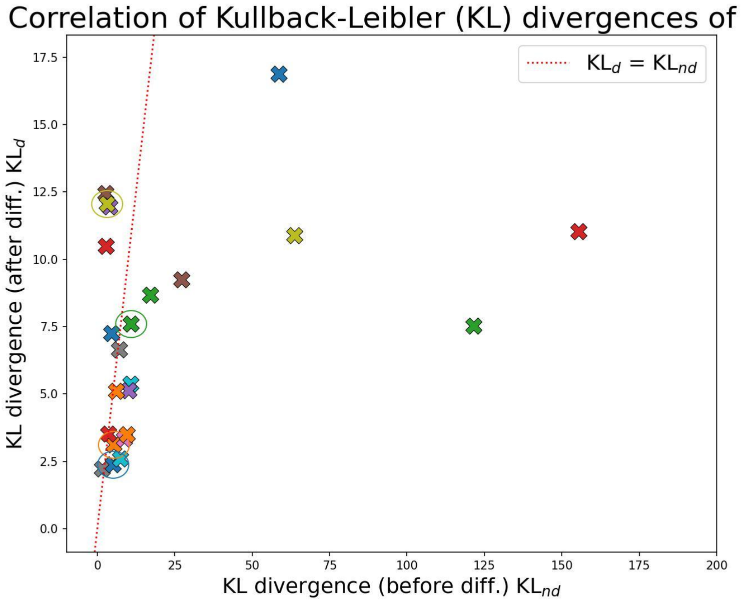

Correlation of standard deviations


Distribution of ratios μ_{nd} / μ_d


Distribution of ratios σ_d/σ_{nd}

Application to real data





$$\eta = \frac{\sigma}{\mu}$$

$$\begin{aligned}\mathrm{KL}(P||Q) &= D_{KL}(P||Q) \\ &= \mathbb{E}_{x\in\mathcal{X}}(\log\left(P(x)/Q(x)\right))\end{aligned}$$

Bimodality Index

- ➔ The proteins of non-differentiated cells show a **bimodal distribution**, which is not observed in the proteins of differentiated cells
- ➔ An index quantifying the bimodality of the protein distribution might be a useful indicator of the differentiation process

- ➔ Solution: the **Bimodality Index (BI)**
  - ◆ *Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Inform. 2009 Aug 5;7:199-216. doi: 10.4137/cin.s2846. PMID: 19718451; PMCID: PMC2730180.*

Bimodality Index

- ➔ The proposed Bimodality Index is estimated from the mixture of two Gaussian distributions with equal variance

$$y = p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2)$$

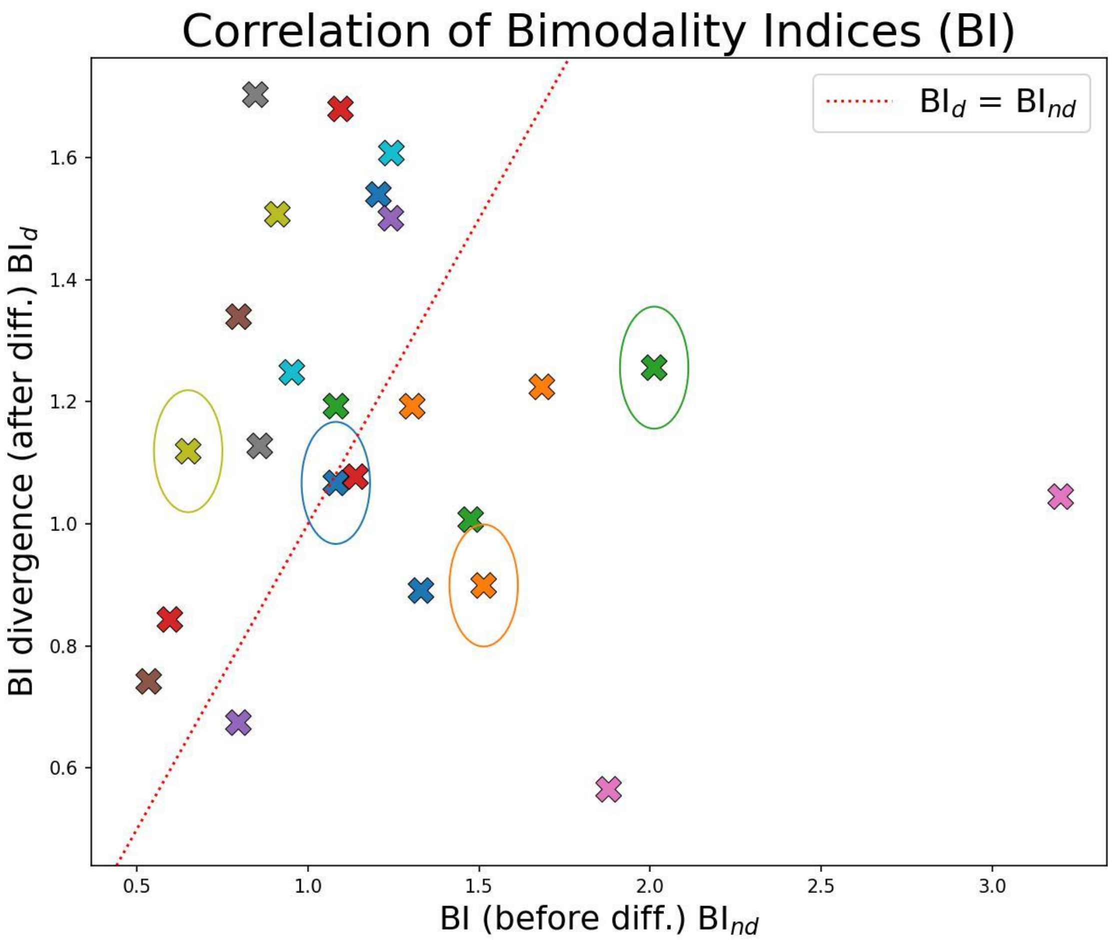
$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma} \; : \; \text{normalized distance}$$

$$\boxed{\mathrm{BI} = \sqrt{p(1-p)}\;\delta}$$

- ➔ Bimodality Index R package: https://cran.r-project.org/web/packages/BimodalIndex/BimodalIndex.pdf

Bimodality Index

BI = 1.1 cutoff is suggested in the original paper



Conclusions

- **mRNA is crucial in the protein production process**
  - It cannot be naïvely eliminated from the model
- **Stochastic models** are needed to properly describe the “bursts” of protein numbers
- Analytical solutions correctly describe the two- and three-stage models when protein lifetimes are way longer than mRNA lifetimes

- Protein concentrations of stem cells after differentiation follow a **Gamma** distribution.
- The **Bimodality Index** is a useful metrics to distinguish non-differentiated and differentiated cells.

References

- Shahrezaei, Vahid, and Peter S. Swain. "Analytical distributions for stochastic gene expression." *Proceedings of the National Academy of Sciences* 105.45 (2008): 17256-17261.
- Gillespie, Daniel T. "Exact stochastic simulation of coupled chemical reactions." *The journal of physical chemistry* 81.25 (1977): 2340-2361.
- Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Inform. 2009 Aug 5;7:199-216. doi: 10.4137/cin.s2846. PMID: 19718451; PMCID: PMC2730180.

Thank you

UNIVERSITAS · STUDII PADUANI · MCCXXII