Mathematical Modeling and Parameter Inference of Telomere Length Regulation in *Saccharomyces cerevisiae*

Viviana Gavilanes Guerrero[1,2]

Supervisors: Zhou Xu[1], Marie Doumic[2]

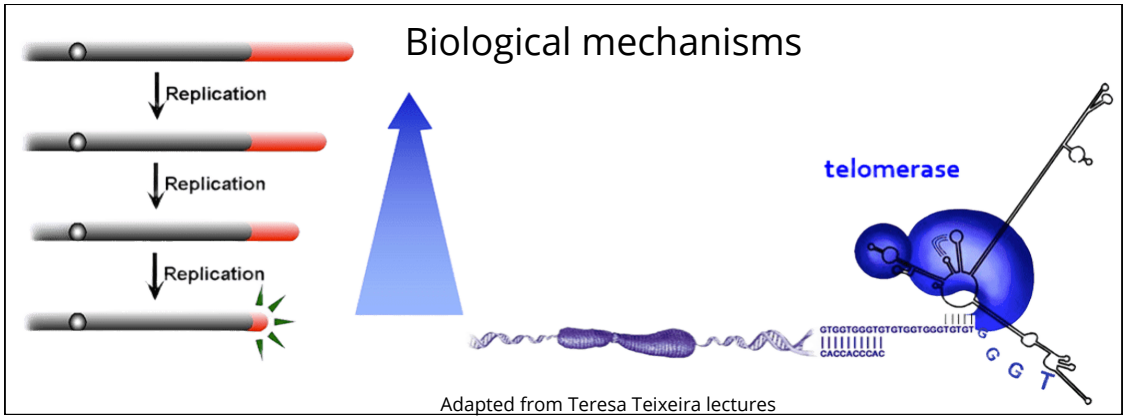
[1] Sorbonne Université, CNRS, Laboratory of Computational, Quantitative and Synthetic Biology
[2] Centre de Mathématiques Appliquées, École Polytechnique & Inria, Palaiseau

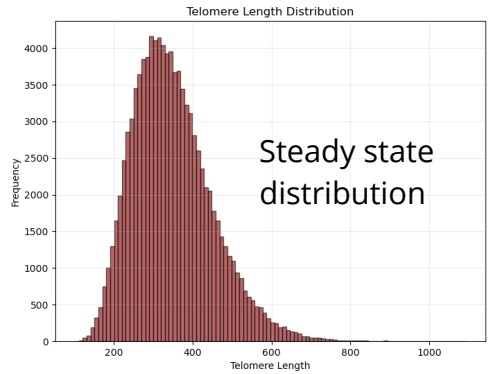
June 18th 2026

From telomere biology to observed length distributions

- **Telomeres**: chromosome ends, **genome integrity**, **shortening** at each division.
- No maintenance $\rightarrow$ **critical shortening**, **senescence**, **ageing**. In contrast, **cancer cells** bypass this via **telomerase**.
- Experiment: **steady-state length distribution**; goal: infer the **underlying mechanism**.

Modeling telomere dynamics





Inverse problem: parameter inference

Long-read Nanopore sequencing

Why is it useful for telomeres?

- Well suited to repetitive telomeric DNA.
- Precise access to full telomere-length distributions.

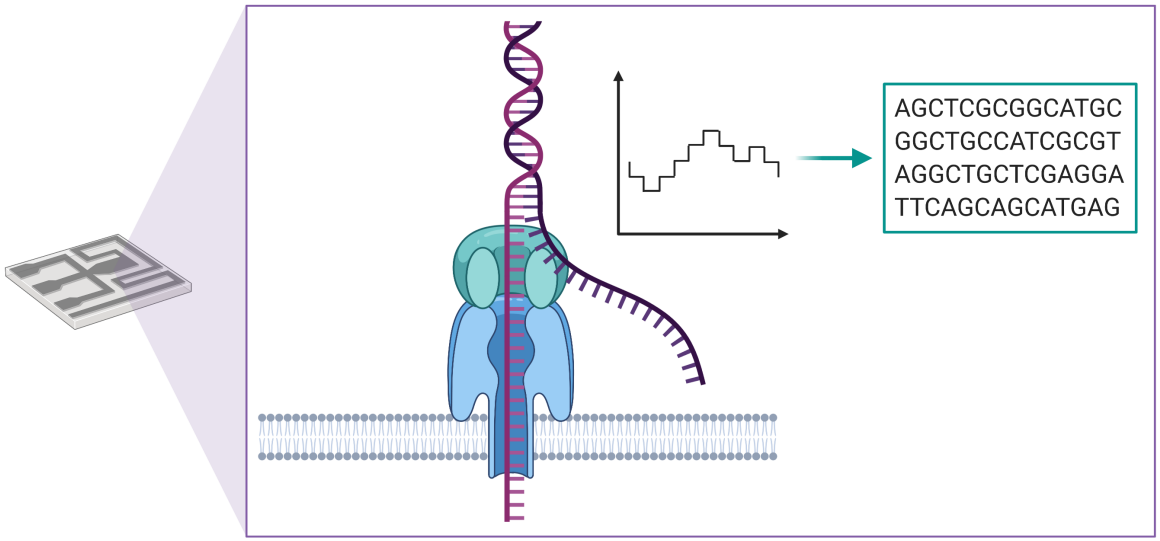
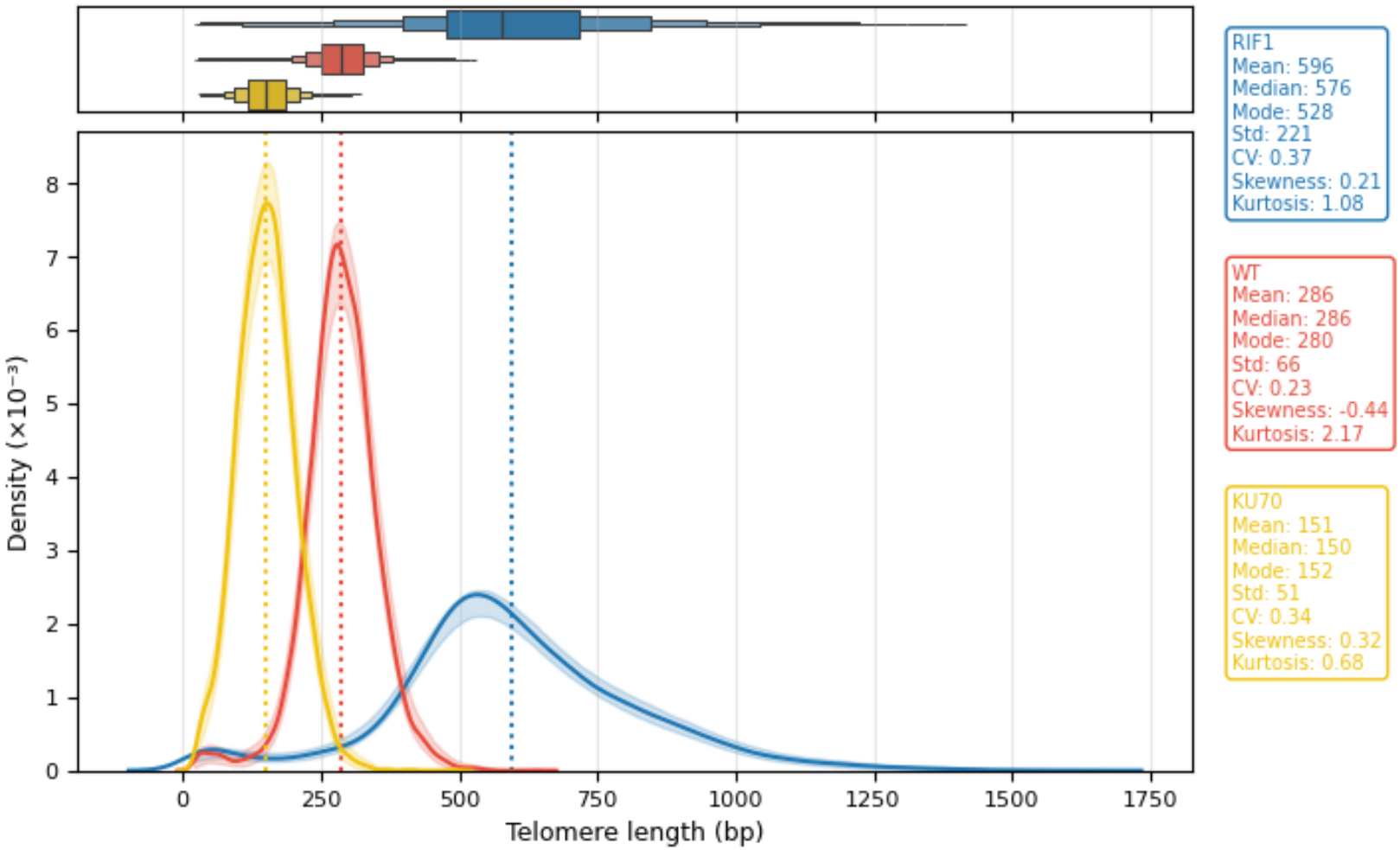


Image from Canada's Michael Smith Genome Sciences Centre

Data used in this work

Full telomere-length distributions from long-read Nanopore sequencing

- Garrido (2026) pipeline.
- Theulot (2025): additional mutants (*rif1*Δ, *yku70*Δ).



Stochastic telomere dynamics

- We follow the length of one telomere along a single lineage across successive cell divisions.

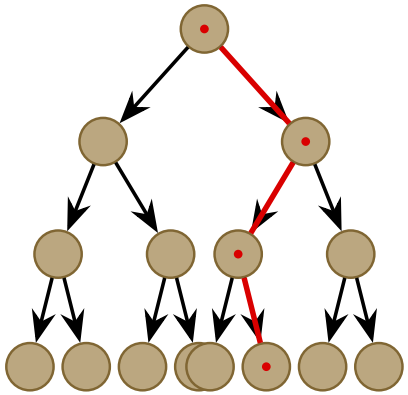
Mathematical representation (Xu, 2013):

$$L_{n+1} = \begin{cases} L_n - a, & \text{with probability } 1 - P(L_n), \\ L_n - a + b, & \text{with probability } P(L_n). \end{cases}$$

Equivalent formulation:

$$L_{n+1} = (L_n - a)_+ + C_n\, G_n,$$

- $a > 0$: constant shortening per division.
- $C_n \sim \text{Bernoulli}(f(L_n))$: elongation indicator.
- $G_n \sim \text{Geom}(p)$: elongation size, with $\mathbb{E}[G_n] = 1/p$.
- $f(L_n)$: elongation probability.



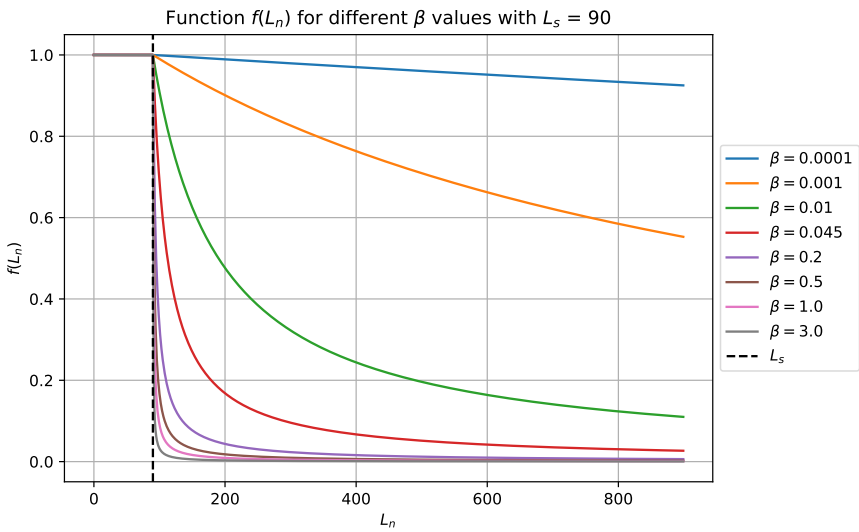
one tracked lineage

Red path: one lineage followed over divisions.

Length-dependent recruitment encodes the biology

$$f(L) = \begin{cases} 1, & \text{if } L \leq L_s, \\ \dfrac{1}{1 + \beta(L - L_s)}, & \text{if } L > L_s. \end{cases}$$

- L_s: threshold below which recruitment is maximal,
- β: rate of decay of recruitment above the threshold,



Interpretation

Very short telomeres have priority, while longer telomeres can still be elongated with decreasing probability (Texeira, 2004).

State space and transition mechanism

We model the telomere length by a time-homogeneous Markov chain $(L_n)_{n\geq 0}$ on $E = \mathbb{N}$, with update rule

$$L_{n+1} = (L_n - a)_+ + C_n G_n,$$

where $C_n \sim \mathcal{B}(f(L_n))$ and $G_n \sim \mathcal{G}(p)$.
The chain is ergodic in the sense proved by Eugène (2017): it admits a unique stationary distribution π

Transition kernel

For $i, j \in E$,

$$P(i,j) = \mathbb{P}(L_{n+1} = j \mid L_n = i) = \begin{cases} 1 - f(i), & j = (i-a)_+, \\ f(i)\, p(1-p)^{j-i+a-1}, & j \geq (i-a)_+ + 1, \\ 0, & \text{otherwise.} \end{cases}$$

Ergodic convergence

There exists a unique stationary distribution π and for all $i \in E$,

$$\|P^n(i,\cdot) - \pi\|_{\mathrm{TV}} \xrightarrow[n\to\infty]{} 0$$

Finite truncation used in practice

We work on $E_{L_{\max}} = \{0, \ldots, L_{\max}\}$ (with a truncated geometric tail) and compute π from

$$\pi = \pi P, \qquad \sum_{j \in E_{L_{\max}}} \pi_j = 1,$$

This is the equilibrium object used for parameter inference: $\pi = \pi(\theta)$.

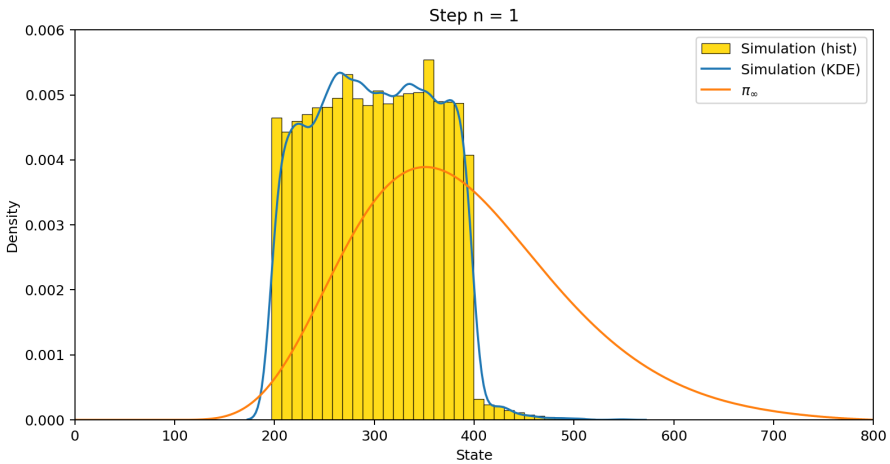
Convergence rate (finite state space)

On $E_{L_{\max}}$, irreducible + aperiodic $\Rightarrow P^n(i, \cdot) \to \pi$ with an geometric rate. If λ_2 is the eigenvalue with the second-largest modulus, then there exists $C > 0$ such that

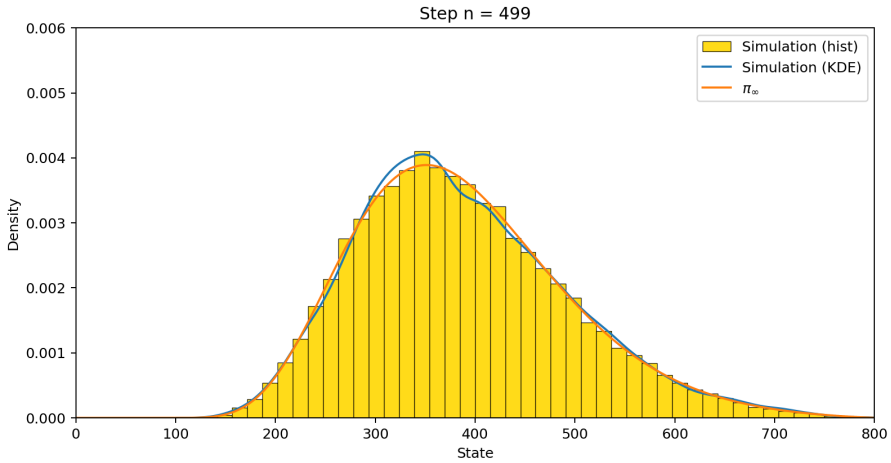
$$\|P^n(i, \cdot) - \pi\|_{\mathrm{TV}} \leq C\,|\lambda_2|^n.$$

A Markov chain with a steady-state distribution

- Example parameters: $L_0 \sim \mathcal{U}(200, 400)$, $a = 4$, $L_s = 90$, $\beta = 0.045$, $p = 0.026$.
- Simulation loop: 173.21 s vs. π_∞: 0.21 s.



Initial distribution

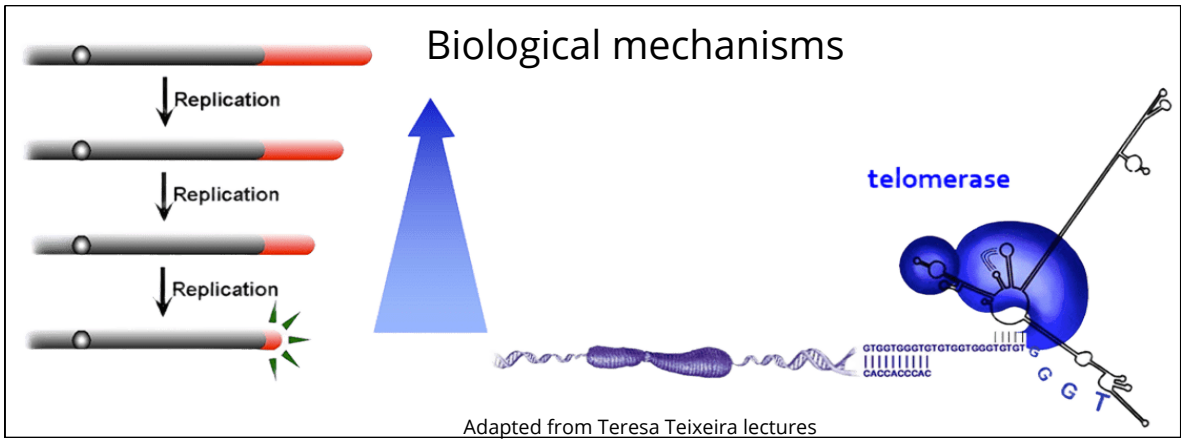


Steady-state distribution

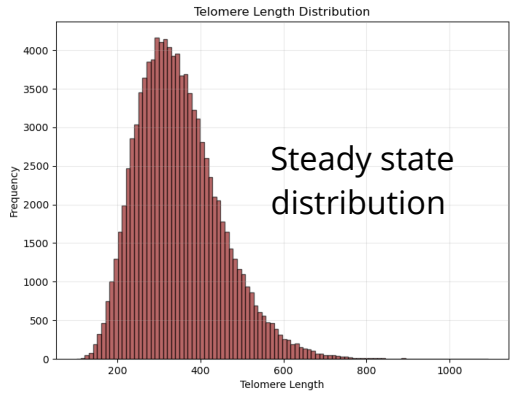
Why is this useful?

This is what we expect from a model of telomere homeostasis: after enough divisions, the system settles into a steady-state distribution.

Modeling telomere dynamics



Adapted from Teresa Teixeira lectures



Inverse problem: parameter inference

Inference principle

Parameter vector

$$\theta = (a, L_s, \beta, p)$$

Compare model and data at stationarity

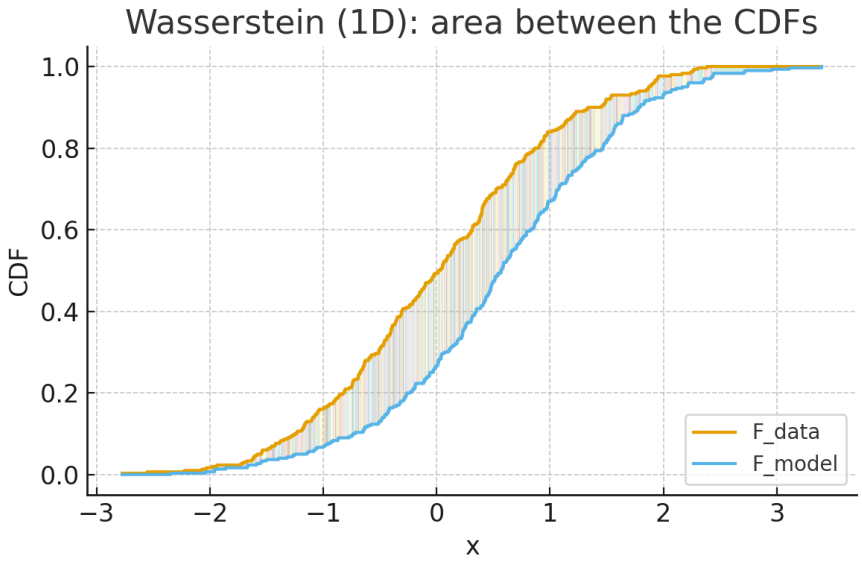
- **Data:** empirical distribution π_{exp} from Nanopore reads.
- **Model:** stationary distribution $\pi_{\text{mod}}(\theta)$.

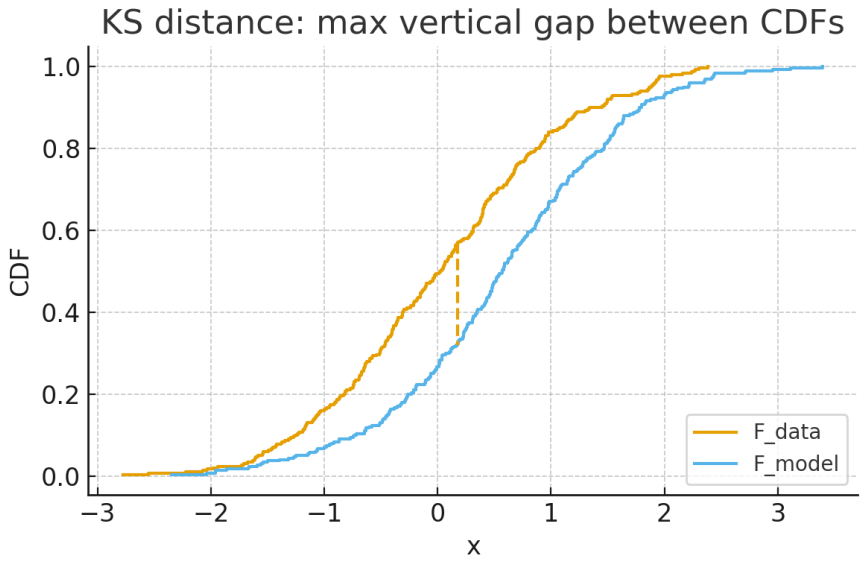
Estimation

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} W_1(\pi_{\text{exp}}, \pi_{\text{mod}}(\theta)),$$

where W_1 denotes the 1-Wasserstein distance.

- The Wasserstein distance is useful here because it compares the whole distribution.
- In practice, optimization is performed with CMA–ES (**Covariance Matrix Adaptation Evolution Strategy**).





Wasserstein distance:

$$W_1(\mu,\nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\, dx$$

in dimension 1. It measures the **area between the two CDFs**.

Kolmogorov–Smirnov distance:

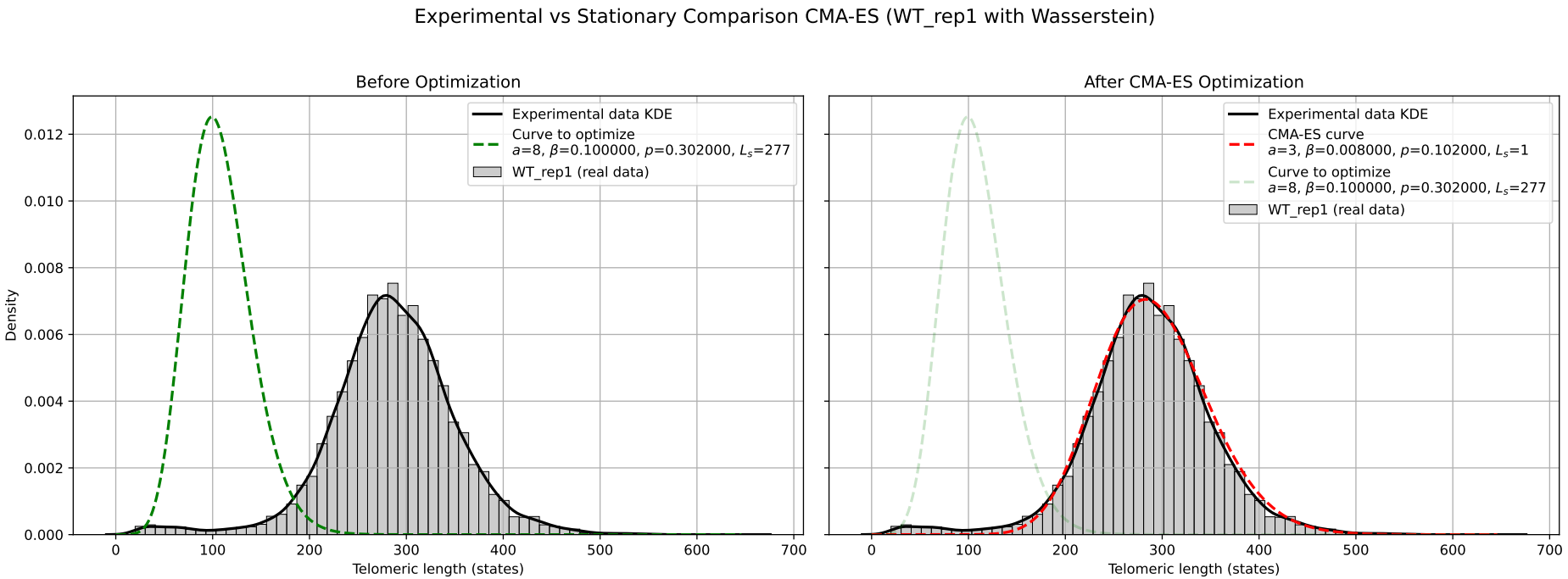
$$d_{KS}(\mu,\nu) = \sup_{x\in\mathbb{R}} |F_\mu(x) - F_\nu(x)|$$

It measures only the **largest vertical gap** between the two CDFs.

Biological interpretation

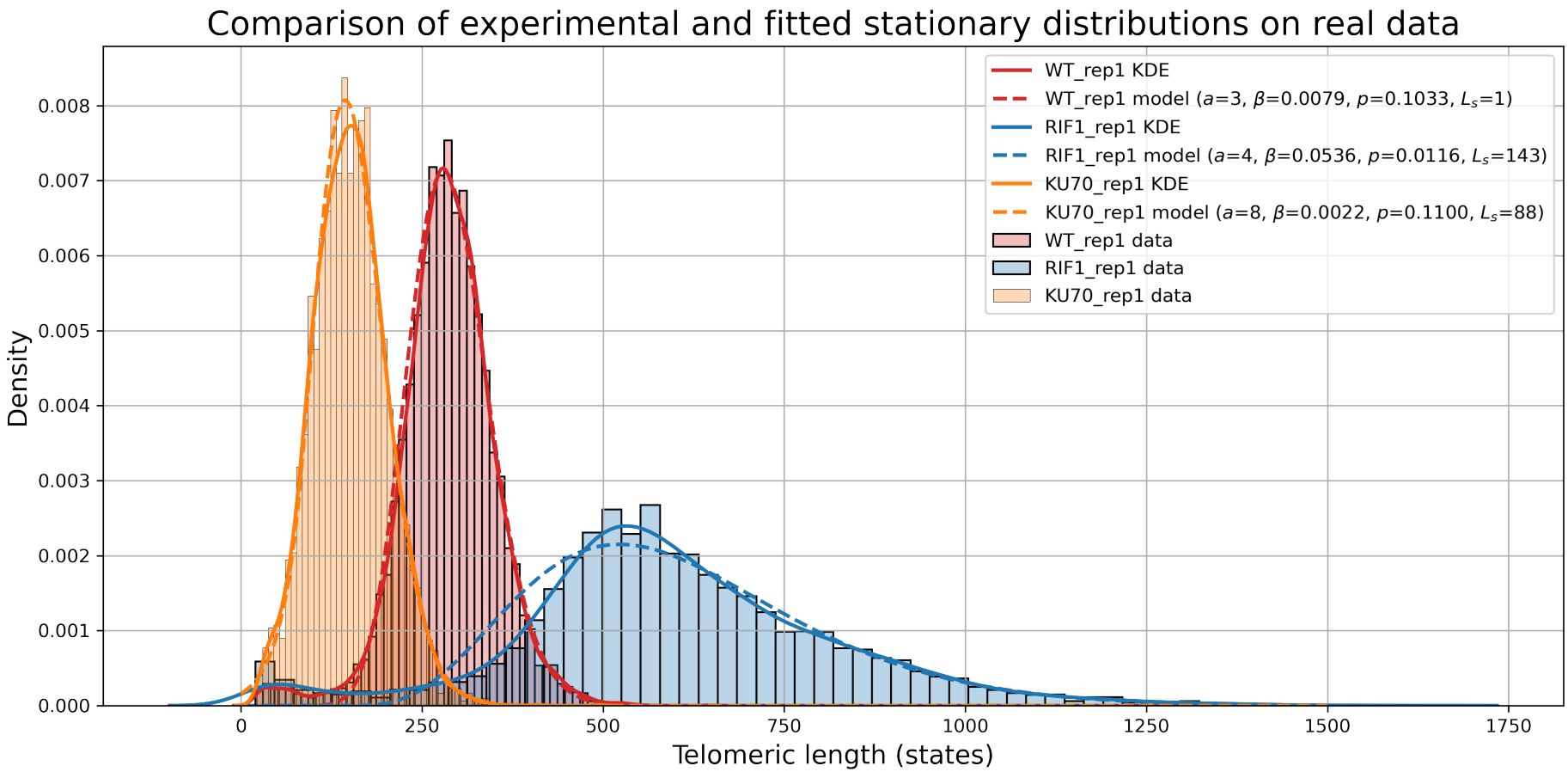
We want to compare the **whole telomere-length distribution**, not just the worst local error. Wasserstein measures the redistribution needed to match model and data, making it more relevant for comparing the **full distribution shape**.

The model reproduces the main empirical features

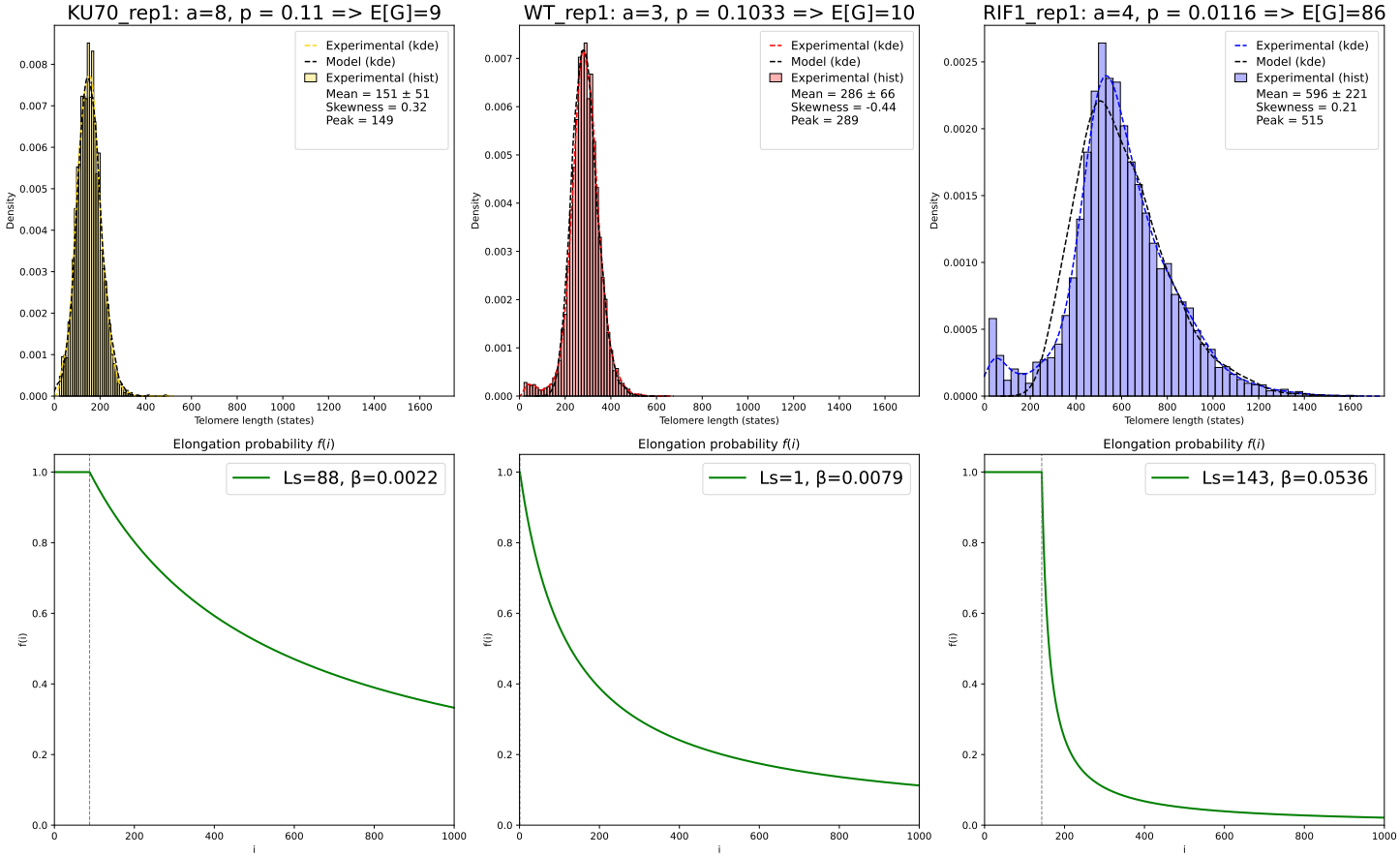


- The fitted stationary distribution reproduces the main empirical shape.
- The framework captures differences in WT (first replicate).

Fitted stationary distributions across strains



- The fitted stationary laws reproduce the main empirical shapes.
- They capture strong differences between WT, `yku70Δ`, and `rif1Δ`.



Strain	a	$1/p$	(L_s, β)	**Interpretation**	**Literature**
WT	3	10	(1, 0.0079)	balanced	–
`yku70Δ`	8	9	(88, 0.0022)	stronger shortening, slower decay	3' overhang problem *Soudet et al., (2014)*
`rif1Δ`	4	86	(143, 0.0536)	stronger elongation, extended plateau then faster decay	Increased elongation *Teixeira et al.* (2004)

Limite discret to continous

In the discrete model, the stationary law is obtained from

$$\pi = \pi P,$$

where

$$P(i,j) = (1 - f(i))\,\mathbf{1}_{\{j=(i-a)_+\}} + f(i)\,p(1-p)^{j-(i+a)_+-1}\,\mathbf{1}_{\{j\geq(i-a)_++1\}}.$$

With CMA-ES, this computation is repeated for many parameter sets, which is costly.

The continuous limit gives

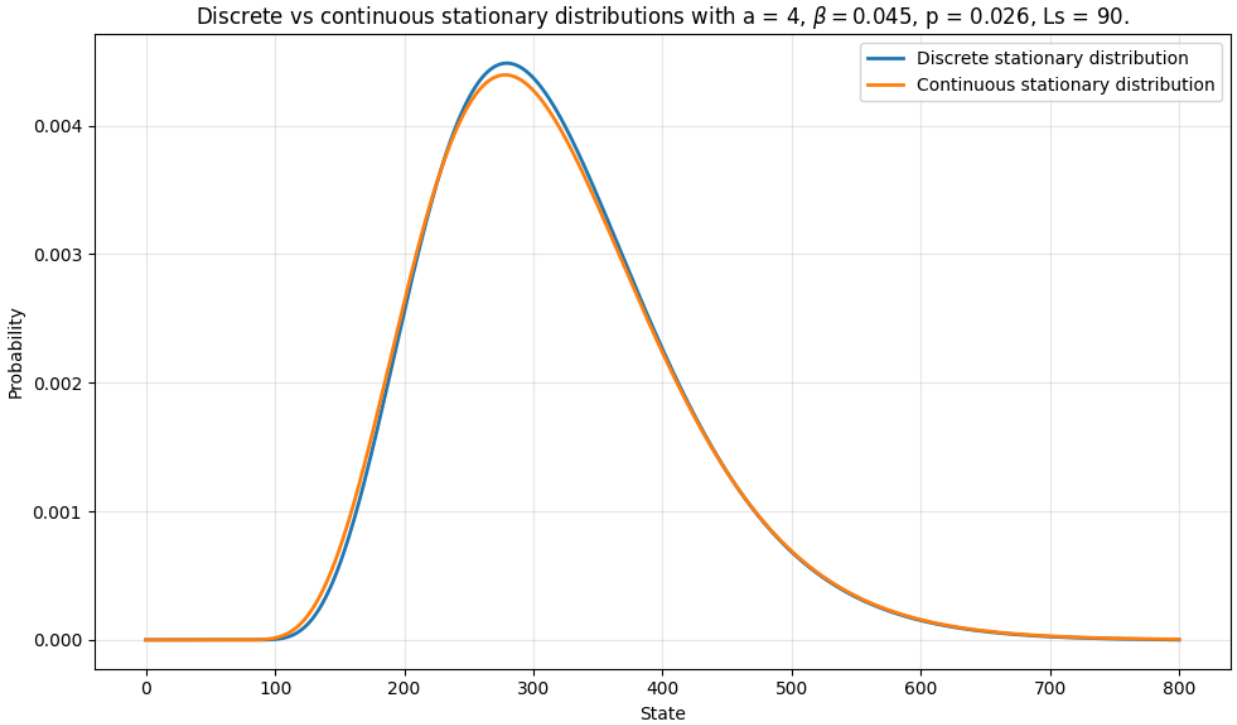
$$\partial_t C(x,t) = \alpha\,\partial_x\big((1-f(x))C(x,t)\big) - f(x)C(x,t) + \mu\int_0^x f(y)C(y,t)e^{-\mu(x-y)}\,dy,$$

and at equilibrium

$$\partial_t C(x,t) = 0.$$

Interest

The continuous model avoids repeated matrix computations and can strongly reduce fitting time while keeping a good approximation of the stationary distribution.



Model	Mean (s)	Std (s)	Speed-up
Discrete	1.1258	0.1634	1.00
Continuous	0.0002	0.0000	5184.64

Main point

Almost same fit, but more than 5000× faster. Crucial for repeated evaluations in optimization by CMA-ES.

The continuous formulation is not only a numerical approximation. It also raises several mathematical questions.

Main objective

Show that the solution of the continuous equation converges toward a stationary profile:

$$\|C(\cdot, t) - \overline{C}(\cdot)\| \longrightarrow 0, \qquad t \to \infty.$$

Ideally, one would like an exponential convergence estimate:

$$\|C(\cdot, t) - \overline{C}(\cdot)\| \leq Ce^{-\nu t}\|C(\cdot, 0) - \overline{C}(\cdot)\|.$$

- Prove well-posedness of the integro-differential equation.
- Prove existence and uniqueness of a stationary state.
- Study convergence toward equilibrium (semigroup or entropy methods).
- Justify that the continuous stationary profile approximates the discrete stationary law.

Key point

This would justify replacing repeated computations of π_θ by the equilibrium profile of the continuous model during parameter fitting.

Take-home messages

1. A simple stochastic model with shortening and preferential rescue of short telomeres captures the main experimental trends.
2. The model provides a bridge between a biological mechanism and an experimentally observed steady-state distribution.
3. Across strains, the fitted parameters support distinct biological regimes of telomere maintenance.
4. The **continuous model** gives a very good approximation of the stationary distribution while being much faster than the discrete model, which is crucial for repeated evaluations in CMA-ES.

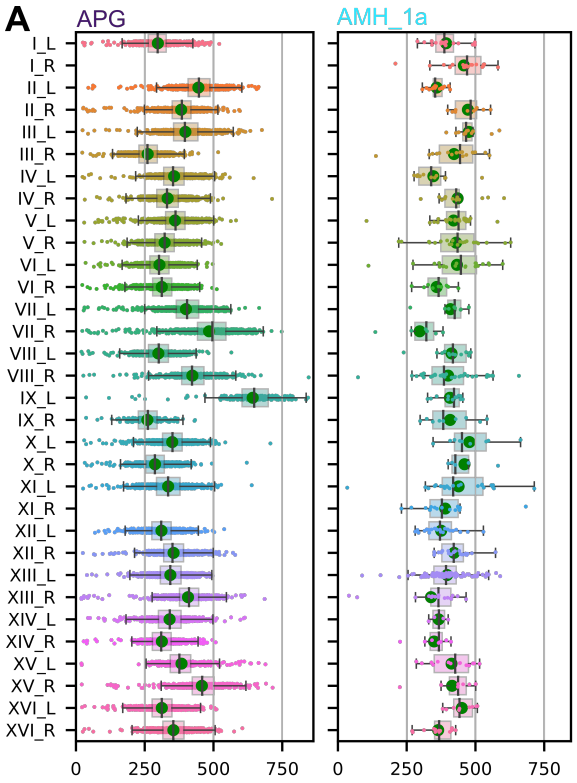
Limits and perspectives

Practical identifiability

- Fix L_s when independent biological or empirical information is available.
- Use **likelihood profiles** to detect flat directions and weakly identified parameters.
- Use a **Bayesian approach** such as **ABC** to include prior biological information and quantify uncertainty.

Model limitations and next steps

- The current model follows one representative telomere, not all chromosome ends jointly.
- Next steps: include **end-specific heterogeneity**.



Example of end-specific heterogeneity.

Acknowledgements

Supervision and scientific guidance:
Zhou Xu, **Marie Doumic**

Teams and collaborators:
Telomere and Genome Stability team (CQSB), Merge team (CMAP)

Funding:
PhD fellowship supported by the **MITI** program of the **CNRS**.

Special thanks:
Clotilde Garrido, Nina Vittorelli, Louis Ollivier and my lab friends.

Thank you for your attention!