Modeling relaxation experiments with a mechanistic model of gene expression

Maxime Estavoyer, Marion Dufeu, Grégoire Ranson, Sylvain Lefort, Thibault Voeltzel, Veronique Maguer-Satta, Olivier Gandrillon, Thomas Lepoutre

BMC Bioinformatics

RESEARCH

Open Access

Modeling relaxation experiments with a mechanistic model of gene expression

Maxime Estavoyer[1], Marion Dufeu[2], Grégoire Ranson[3,4], Sylvain Lefort[5], Thibault Voeltzel[5^], Véronique Maguer-Satta[5], Olivier Gandrillon[6,7] and Thomas Lepoutre[1*]

^T. Voeltzel: deceased.

*Correspondence: thomas.lepoutre@inria.fr

[1] Inria, CNRS, Ecole Centrale de Lyon, INSA Lyon, Universite Claude Bernard Lyon 1, Université Jean Monnet, ICJ UMR5208, 69603 Villeurbanne, France
[2] "Tumor Cell Plasticity in Melanoma", Institut Convergence Plascan, Centre de Recherche en Cancérologie de Lyon, INSERM U1052-CNRS UMR5286, Centre Léon Bérard, Université de Lyon, Université Claude Bernard Lyon1, 69008 Lyon, France
[3] Universite Claude Bernard Lyon 1, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Jean Monnet,, Universite Claude Bernard Lyon 1, Université Jean MonnetICJ UMR5208, Inria, 69622 Villeurbanne, France
[4] Laboratory for Industrial and Applied Mathematics (LIAM), Department of Mathematics and Statistics, York University, Toronto ON M3J 1P3, Canada
[5] Cancer Research Center of Lyon (CRCL), CNRS UMR5286, INSERM U1052, Léon Bérard Center, Lyon 1 university, Lyon, France
[6] ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie Site Jacques Monod, Univ Lyon, 69007 Lyon, France
[7] Inria, Paris, France

Abstract

Background: In the present work, we aimed at modeling a relaxation experiment which consists in selecting a subfraction of a cell population and observing the speed at which the entire initial distribution for a given marker is reconstituted.

Methods: For this we first proposed a modification of a previously published mechanistic two-state model of gene expression to which we added a state-dependent proliferation term. This results in a system of two partial differential equations. Under the assumption of a linear dependence of the proliferation rate with respect to the marker level, we could derive the asymptotic profile of the solutions of this model.

Results: In order to confront our model with experimental data, we generated a relaxation experiment of the CD34 antigen on the surface of TF1-BA cells, starting either from the highest or the lowest CD34 expression levels. We observed in both cases that after approximately 25 days the distribution of CD34 returns to its initial stationary state. Numerical simulations, based on parameter values estimated from the dataset, have shown that the model solutions closely align with the experimental data from the relaxation experiments.

Conclusion: Altogether our results strongly support the notion that cells should be seen and modeled as probabilistic dynamical systems.

Keywords: Relaxation experiments, Two-state model, Asymptotic profile

Background

Cells are neither machines [1] nor simple information processing devices [2, 3]. Their specific complexity sometimes led to the idea that they should be treated differently that classical physico-chemical systems [4]. Nevertheless like all living systems cells are rooted within a physico-chemical reality which they can not escape. We therefore argue that cells should be seen and modelled as probabilistic dynamical systems.

One obvious sign that cells should indeed be seen as such lie in the possibility to perform so-called "relaxation" experiments. This consists in selecting a subfraction of a cell population (potentially down to one cell) and observing the speed at which the entire

initial distribution for a given marker is reconstituted. Such relaxation experiments have already been published and analyzed on various cells and antigens.

Arguably the very first report to do so analyzed the distribution of the Sca1 antigen (Stem Cell Antigen 1) at the surface of EML cells, a multipotent mouse haematopoietic cell line. It was shown that it took more than 9 days before the three fractions (most Sca-1 negative, most Sca-1 positive and a central fraction) regenerated Sca-1 histograms similar to that of the parental (unsorted) population [5]. The authors proposed a phenomenological model which point toward discrete transitions in a dynamical system exhibiting multistability to quantitatively predict the relaxation dynamics of the sorted subpopulations [5]. For this they assumed the existence of two stable states, one of low and one of high Sca1 expression. Proliferation was assumed to be equal in both states.

Other studies have adopted a somewhat different approach with the knock-in of fluorescent reporter genes under the control of endogenous promoters [6, 7]. The first targeted promoter was that of Nanog in murine embryonic stem cells [6], an other gene classically considered as a stemness marker. Similarly to [5], the authors demonstrated that, although being in a Nanog low of in a Nanog high state is not biologically equivalent in term of fate, the transition between these two states can be adequately modelled using a fully probabilistic model, simulated using a Stochastic Simulation Algorithm [8].

The second targeted promoter was that of Tenascin-C in NIH 3T3 mouse fibroblasts [7]. In that case, the authors first proposed a phenomenological 2-states model, which proved to not correctly capture their data. They then turned toward a Langevin type stochastic differential equation to model the relaxation process. This led to an accurate prediction of the rates at which different phenotypes will arise from an isolated subpopulation of cells [7]. In contrast with [5], the authors assumed that each state had its own proliferation rate.

In the present work, we aimed at using a previously published mechanistic model of gene expression [9] to which we will add a stemness-dependant proliferation term, to fit relaxation data obtained by examining CD34 expression at the surface of TF1-BA cells.

Methods

Mathematical model

Case without proliferation

Throughout this work, we will use the classical two-state model (Fig. 1; see [9] and references therein), a refinement of the pioneer model introduced by [10].

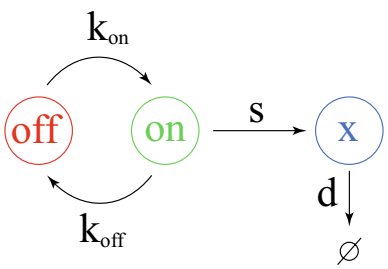

Fig. 1 The 2-state model of gene expression. The gene opens with a k_{on} rate and closes with a k_{off} rate. Similarly to [11] we only consider protein (x) production (with an s rate) and degradation (with a d rate)

This is the simplest model that accounts well for the specific nature of single-cell omics data (non-poisonnian [12], well fitted by Gamma distributions [13] and displaying a high proportion of zero counts [14]). More refined models with any number of possible gene configuration have been described [15] but their mathematical complexity makes them cumbersome to use for our purpose.

It is important to stress here that this is a mechanistic model, that differs from the phenomenological 2-states model described upper. Such models only considered a low and a high χ state, without describing the protein production process. Importantly here stochasticity is described at the core of the modelling and does not need to be introduced as a additional term in the model. We recently proposed a piecewise deterministic Markov process (PDMP) version of that model which rigorously approximates the original molecular model [9]. Furthermore, a moment-based method has been proposed for estimating parameter values from a given experimental distribution assumed to arise from the functioning of a 2-states model [11]. We recall here the mathematical description of the model through the PDMP (piecewise deterministic Markov process) formalism

$$\frac{d}{dt}\chi = s.E(t) - d\chi(t),$$

where $E(t) \in \{0,1\}$ switching from 0 to 1 (resp. 1 to 0) at a rate k_{on} (resp. k_{off}). In this process, the protein quantity $\chi(t)$ is structurally bounded by $X_{max} = s/d$. From this process, we can derive the Chapman Kolmogorov or master equation in the form

$$\begin{cases} \partial_t n_{\text{on}}(t,\chi) + \partial_\chi J_{\text{on}}(t,\chi) = -k_{\text{off}} n_{\text{on}} + k_{\text{on}} n_{\text{off}} \quad \chi \in]0, s/d[, \\ \partial_t n_{\text{off}}(t,\chi) + \partial_\chi J_{\text{off}}(t,\chi) = +k_{\text{off}} n_{\text{on}} - k_{\text{on}} n_{\text{off}} \quad \chi \in]0, s/d[, \\ J_{\text{on}}(t,\chi) = (s - d\chi) n_{\text{on}}(t,\chi),\ J_{\text{off}}(t,\chi) = -d\chi, \\ J_{\text{on}}(t,0) = J_{\text{off}}(t,0) = J_{\text{on}}(t,s/d) = J_{\text{off}}(t,s/d) = 0. \end{cases} \tag{1}$$

see [16–18] for similar derivations. master equation of the process in the absence of proliferation reads. We recall that the boundary conditions simply reflects the no-flux boundary conditions stating $(s - d\chi)n_{\text{on}}(t,\chi) = -d\chi n_{\text{off}}(t,\chi) = 0$ for $\chi = 0, s/d$. Moreover, because $s - ds/d = -d.0, = 0$, we only specify the boundary conditions when they give constraints on the densities. We define $X_{\text{max}} = s/d$ as the maximum value for the quantity of CD34 in a cell. Scaling the space by X_{max} allows us to consider the following system

$$\begin{cases} \partial_t n_{\text{on}} + d\partial_x((1-x)n_{\text{on}}) = -k_{\text{off}} n_{\text{on}} + k_{\text{on}} n_{\text{off}}, \quad x \in]0,1[, \\ \partial_t n_{\text{off}} + d\partial_x((-x)n_{\text{off}}) = k_{\text{off}} n_{\text{on}} - k_{\text{on}} n_{\text{off}}, \quad x \in]0,1[, \\ n_{\text{on}}(t,0) = n_{\text{off}}(t,1) = 0. \end{cases} \tag{2}$$

with $n_{\text{on}}(t,x)$ being the number of cells with a promoter in the on state at time t, with a (scaled) CD34 level x and $n_{\text{off}}(t,x)$ being the number of cells with a promoter in the off state. The total number of cells, denoted as $n(t,x)$, is given by $n_{\text{on}}(t,x) + n_{\text{off}}(t,x)$. This is the quantity we considered to be measured.

Steady state of the model. The system is mass preserving and it converges to a steady state $N_{\text{on,off}}$ which is characterized by

$$\begin{cases} d\partial_x((1-x)N_{\text{on}}) = -k_{\text{off}}N_{\text{on}} + k_{\text{on}}N_{\text{off}}, \quad x \in]0,1[, \\ d\partial_x((-x)N_{\text{off}}) = -k_{\text{off}}N_{\text{on}} - k_{\text{on}}N_{\text{off}}, \quad x \in]0,1[, \\ N_{\text{on}}(0) = N_{\text{off}}(1) = 0. \end{cases} \tag{3}$$

And the solution if nonnegative. An interesting feature of this system is the fact that we have an explicit solution. We recall here the computations that can be found in [16] because they might help for understanding the computations for the model with proliferation. Indeed, summing up the equations, we get

$$\partial_x((1-x)N_{\text{on}} - xN_{\text{off}}) = 0.$$

Therefore, this quantity is constant on $]0,1[$. Using the boundary condition, we can see that 0 is the only admissible constant. Therefore, in this precise case, we have necessarily

$$(1-x)N_{\text{on}} = xN_{\text{off}}.$$

Injecting in the equation we get

$$d\partial_x((1-x)N_{\text{on}}) = -k_{\text{off}}N_{\text{on}} + k_{\text{on}}\frac{(1-x)N_{\text{on}}}{x} = (1-x)N_{\text{on}}\left(-\frac{k_{\text{off}}}{1-x} + \frac{k_{\text{on}}}{x}\right).$$

From this, we get

$$N_{\text{on}} = C(1-x)^{\frac{k_{\text{off}}}{d}-1}x^{\frac{k_{\text{on}}}{d}}, \quad N_{\text{off}} = C(1-x)^{\frac{k_{\text{off}}}{d}}x^{\frac{k_{\text{on}}}{d}-1}.$$

and quite remarkably, we have

$$N(x) = N_{\text{on}}(x) + N_{\text{off}}(x) = C(1-x)^{\frac{k_{\text{off}}}{d}-1}x^{\frac{k_{\text{on}}}{d}-1}. \tag{4}$$

If we choose $C = \frac{\Gamma((k_{\text{on}}+k_{\text{off}})/d)}{\Gamma(k_{\text{on}}/d)\Gamma(k_{\text{off}}/d)}$ we normalize this to 1 and get a β law $B(k_{\text{on}}/d, k_{\text{off}}/d)$, so that we end up with

$$(n_{\text{on}}(t,x), n_{\text{off}}(t,x)) \to \left(\int_0^1 n^0_{\text{on}}(x) + n^0_{\text{off}}(x)\right)(N_{\text{on}}(x), N_{\text{off}}(x)).$$

Case with proliferation

HSCs mostly reside in a quiescent state, although they can occasionally divide during homeostasis [19]. We therefore will consider $CD34^+$ TF1-BA cells as immature slowly proliferating cells and $CD34^-$ TF1-BA cells as mature highly proliferating cells. Therefore the proliferation rate will depend on the x variable representing the CD34 content but not on the on/off status.

Moreover, we consider that cells keep their on/off status during a division. This is in line with the demonstration of a memory of transcriptional activity in mammalian cells [20, 21].

$$\begin{cases} \partial_t n_{\text{on}} + d\partial_x((1-x)n_{\text{on}}) = -k_{\text{off}}n_{\text{on}} + k_{\text{on}}n_{\text{off}} + r(x)n_{\text{on}}(t,x), \quad x \in]0,1[, \\ \partial_t n_{\text{off}} + d\partial_x((-x)n_{\text{off}}) = k_{\text{off}}n_{\text{on}} - k_{\text{on}}n_{\text{off}} + r(x)n_{\text{off}}(t,x), \quad x \in]0,1[, \\ n_{\text{on}}(t,0) = n_{\text{off}}(t,1) = 0. \end{cases} \tag{5}$$

Since the system is structurally non-conservative, it makes no sense to look for steady state here. However, one can investigate for stable exponential profile, that is to look for positive solutions with shape

$$e^{\lambda t}(N_{\text{on}}(x), N_{\text{off}}(x)).$$

Such solution satisfy the system

$$\begin{cases} \lambda N_{\text{on}} + d\partial_x((1-x)N_{\text{on}}) = -k_{\text{off}}N_{\text{on}} + k_{\text{on}}N_{\text{off}} + r(x)N_{\text{on}}(x), & x \in]0,1[, \\ \lambda N_{\text{off}} + d\partial_x((-x)N_{\text{off}}) = k_{\text{off}}N_{\text{on}} - k_{\text{on}}N_{\text{off}} + r(x)N_{\text{off}}(t,x), & x \in]0,1[, \\ N_{\text{on}}(0) = N_{\text{off}}(1) = 0, \quad N_{\text{on,off}} > 0. \end{cases} \tag{6}$$

In the sequel, we will focus on the normalized representant so that we will assume

$$\int_0^1 N_{\text{on}}(x) + N_{\text{off}}(x)dx = 1.$$

We also introduce the adjoint eigenprofile. It can be obtained as exponentially growing solutions for the adjoint differential operators, this is the continuous equivalent of left and right eigenvector for the same eigenvalues in matrix analysis ($MU = \lambda U, v^T M = \lambda V$ or equivalently $M^T V = \lambda V$).

$$\begin{cases} \lambda\phi_{\text{on}} - d(1-x)\partial_x\phi_{\text{on}} = -k_{\text{off}}\phi_{\text{on}} + k_{\text{off}}\phi_{\text{off}} + r(x)\phi_{\text{on}}, & x \in]0,1[, \\ \lambda\phi_{\text{off}} - d(-x)\partial_x(\phi_{\text{off}}) = -k_{\text{on}}\phi_{\text{off}} + k_{\text{on}}\phi_{\text{on}} + r(x)\phi_{\text{off}}, & x \in]0,1[, \\ \phi_{\text{on,off}} > 0, \quad \int_0^1 N_{\text{on}}\phi_{\text{on}} + N_{\text{off}}\phi_{\text{off}}dx = 1. \end{cases} \tag{7}$$

We emphasize in particular the following property, for any initial condition of the system, one has

$$\int_0^1 n_{\text{on}}(t,x)\phi_{\text{on}}(x) + n_{\text{off}}(t,x)\phi_{\text{off}}(x)dx = e^{\lambda t}\int_0^1 n_{\text{on}}(0,x)\phi_{\text{on}}(x) + n_{\text{off}}(0,x)\phi_{\text{off}}(x)dx = C^0 e^{\lambda t},$$

and moreover,

$$\int_0^1 |e^{-\lambda t}n_{\text{on}}(t,.) - C^0 N_{\text{on}}|\phi_{\text{on}}(x)dx + \int_0^1 |e^{-\lambda t}n_{\text{off}}(t,.) - C^0 N_{\text{off}}|\phi_{\text{off}}(x)dx \to 0.$$

$$e^{-\lambda t}(n_{\text{on}}(t,.), n_{\text{off}}(t,.)) \to C^0(N_{\text{on}}, N_{\text{off}}).$$

In the weighted norm $\|(f_{\text{on}}, f_{\text{off}})\|_\phi = \int_0^1 |f_{\text{on}}|\phi_{\text{on}} + |f_{\text{off}}|\phi_{\text{off}}$. In case $\phi_{\text{on,off}}$ is lower bounded, this implies classical L^1 convergence. This lower bound is established below. For more details on this, we refer to [22] for an introduction to eigenvectors in this context. In particular, thanks to our normalization, the triplet (λ, N, ϕ) is uniquely defined. Note that in the conservative case ($r = 0$), $\lambda = 0$, N is given by the renormalized steady state and the adjoint eigenvector is simply the constant vector (1, 1). Note also that this guarantees that for any initial data, we have in case $\lambda \geq 0$ and $\phi_{\text{on,off}}$ are lower bounded (as it will be established below)

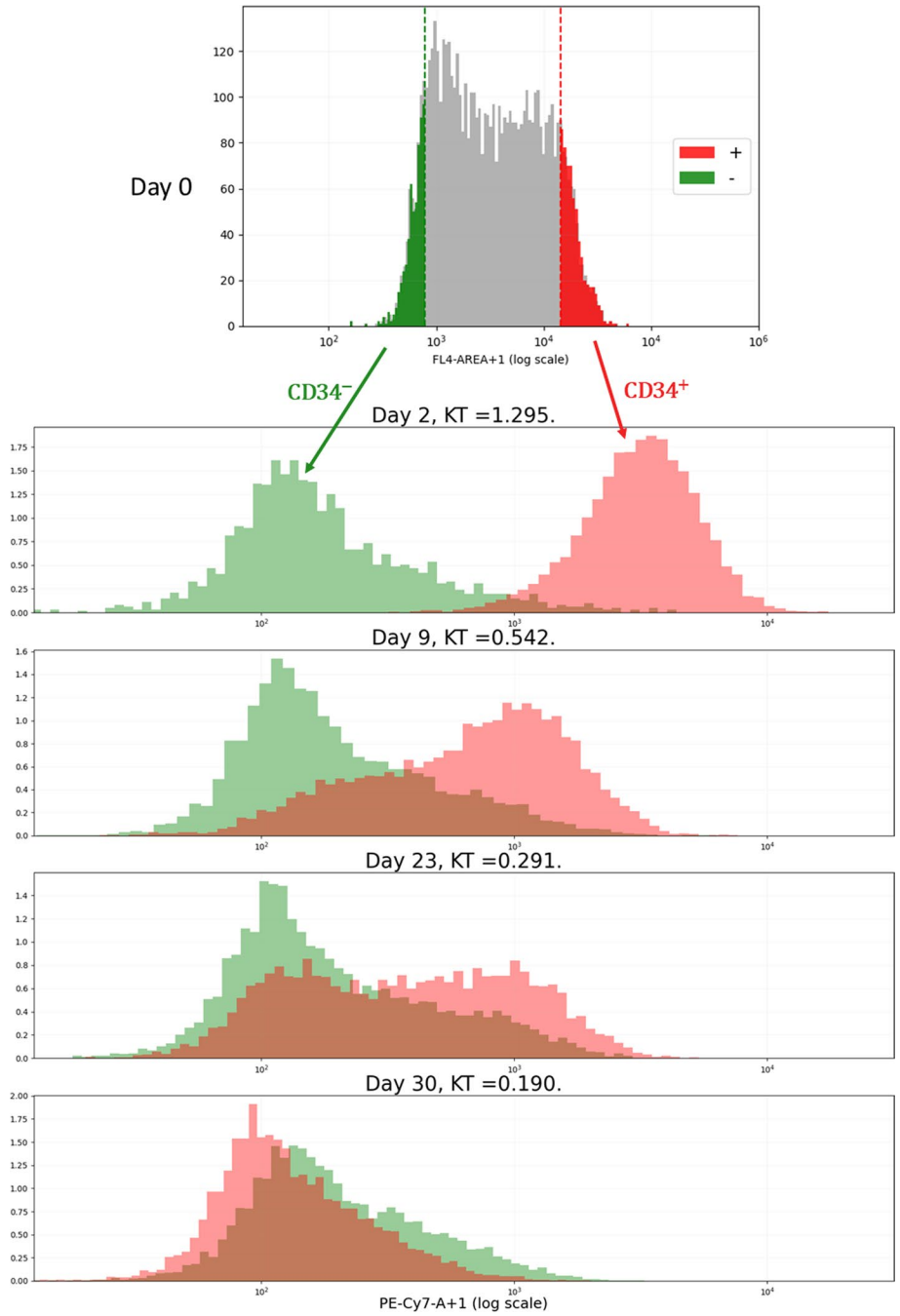

Fig. 2 The relaxation experiment. TF1-BA cells well labelled with an anti-CD34 antibody and FACS-sorted. The 10 percent most CD34 positive and the 10 percent most CD34 negative cells were sorted, grown in culture for the indicated period of time, where the distribution of cell-surface CD34 expression was assessed. KT: the modified Kantorovich–Rubinstein distance, defined by the Eq. (14), between the two distributions [26]

$$\begin{aligned}
&\frac{1}{\int_0^1 n_{\text{on}}(t,x) + n_{\text{off}}(t,x)dx}(n_{\text{on}}(t,.), n_{\text{off}}(t,.)) \\
&= \frac{1}{e^{-\lambda t}\int_0^1 n_{\text{on}}(t,x) + n_{\text{off}}(t,x)dx} e^{-\lambda t}(n_{\text{on}}(t,.), n_{\text{off}}(t,.)) \\
&\quad \to \frac{1}{C^0} C^0(N_{\text{on}}, N_{\text{off}}) = (N_{\text{on}}, N_{\text{off}}).
\end{aligned}$$

And regarding the observations of the steady profile in Fig. 2, our normalized asymptotic profile should be $N : x \longmapsto N_{\text{on}}(x) + N_{\text{off}}(x)$.

We assume that, for the initial dimensional system, the proliferation rate is linear, i.e. $r : \chi \longmapsto \tilde{r}_0 + \tilde{r}_1\chi$. Scaling again the space by $X_{\max}$, the proliferation term becomes, $r : x \to r_0 + r_1 x$ with $r_0 = \tilde{r}_0$ and $r_1 = \tilde{r}_1 X_{\max}$. We assume that the constant proliferation rate is positive, $r_0 > 0$. Conversely, to model the fact that $CD34^-$ cells divide more frequently than $CD34^+$ cells, we assume that the linear proliferation term, r_1, is negative. However, to preserve the positivity of the proliferation rate, the constant r_1 must satisfy the following condition, $r_1 \in [-r_0, 0]$.

We show in the Results section that, under this proliferation assumption, it is theoretically possible to derive the normalized asymptotic profiles $(N_{\text{on}}, N_{\text{off}})$.

The biological setting

Relaxation experiments

Chronic Myeloid Leukemia (CML) is a myeloproliferative disorder arising at the hematopoietic stem cell (HSC) level. It is associated with the recurrent chromosomal (Philadelphia) translocation t(9;22)(q34;q11) which leads to the oncogenic chimeric gene that fuses Bcr and Abl genes and results in the expression of a constitutively active unique tyrosine kinase named BCR-ABL [23].

Véronique Maguer-Satta's group has developed the TF1-BA cell line, a unique model of immature human hematopoietic cells (TF1) transformed with BCR-ABL by lentiviral infection. This model was shown to recapitulate early alterations following the BCR-ABL oncogene appearance as identified using primary samples of CML patients at diagnosis and in chronic phase [24].

We decided to analyze the relaxation dynamics for the CD34 antigen at the surface of those TF1-BA cells (Fig. 2). CD34 is a transmembrane phosphoglycoprotein which is predominantly regarded as a marker of Haematopoietic Stem Cells (HSCs) [25]. We reasoned that CD34 surface expression could therefore be seen as a proxy for stemness of our TF1-BA cells. Interestingly, one observes a relaxation in both directions: $CD34^-$ cells are regenerated from $CD34^+$ cells, as biologists would expect, but one also see that $CD34^+$ cells are regenerated from $CD34^-$ cells, establishing that stemness is not a fixed quality but the result of an underlying dynamical system as previously shown in other cellular systems ([5, 6]; see upper)

Data processing

Two types of data were collected on days 2, 5, 9, 13, 19, 23, 26 and 30: cell counts and fluorescence distribution. The cell counts allow us to quantify proliferation whereas the fluorescence measure the distribution of CD34 expression.

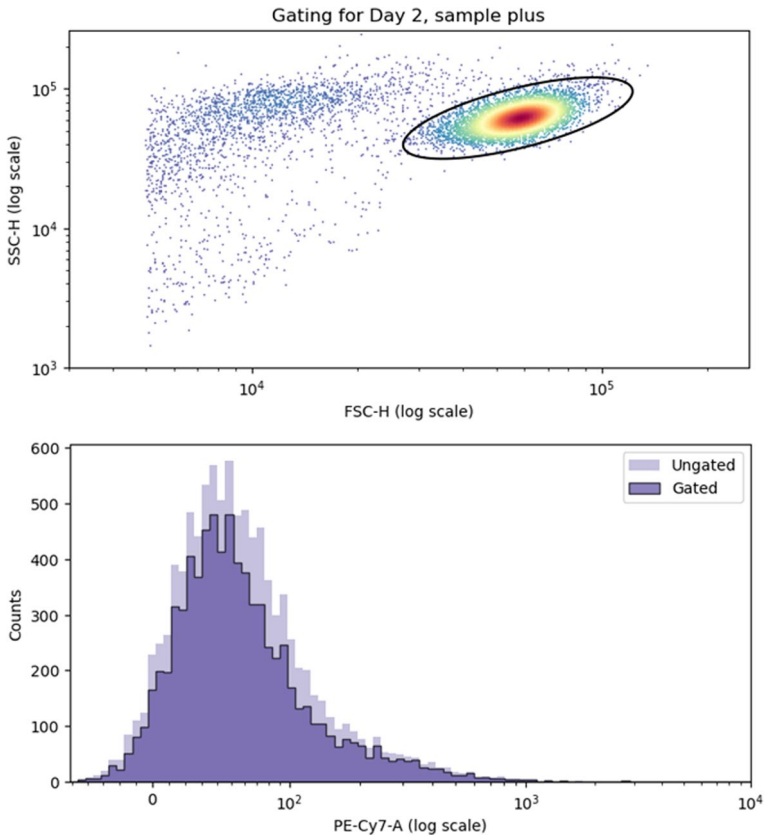

Fig. 3 Example of flow cytometry gating. *Top.* Example of SSC-H along FSC-H plot for raw data from the plus subpopulation on day 2. As the data contain a high proportion of debris cells, we select only those viable cells lying within the black ellipse. *Bottom.* Fluorescence data before gating (*Ungated*) and after gating (*Gated*). For the figures and the ellipse, we used the python package "FlowCal" [27]

Gating.

As usual for cytometric data, we initiate the analysis with a gating step. We use SSC-H and FSC-H data to distinguish viable and debris cells. FSC (Forward Scatter) data are generally assimilated to the size of the cells analyzed, and correspond to the light scattered along the laser path. SSC (Side Scattered) data, on the other hand, are usually linked to the granularity and correspond to the light scattered at a 90-degree angle. The "H" stands for Height and is one component of this type of data. Cell debris are characterized by relatively low size and high granularity relative to their size.

To select viable cells, we plot the values of SSC-H along those of FSC-H. An example of such a graph is provided in Fig. 3 using data from day 2 of the $CD34^+$ cell experiment. Visually, viable cells can be identified as the cluster of points with high FSC-H and SSC-H values. Using the "FlowCal" python package [27], we draw an ellipse as a filter to select only these viable cells. At the bottom of Fig. 3, we represent fluorescence data with and without the gating phase (*Ungated* and *Gated*, respectively). Note that fluorescence distribution is only slightly affected by the removal of debris cells.

Shifting.

Even after gating, some cells exhibit a negative fluorescence level, which is inconsistent as these values are intended to represent the amount of proteins in each cell.

To avoid this problem, we added a shifting step. This step occurs immediately after the gating process and consists in subtracting the minimum value of each distribution (for each sub-population and for each day) from all the values, bringing the minimum to 0. This transformation, once again, does not distort the fluorescence distribution. This corresponds to the interpretation of negative values as compensation of a baseline value.

Numerical simulations

Linking data to mathematical model: The cell counts are interpreted as snapshots of the total population $\int_0^{X_{\max}}(n_{\text{on}}(t,x)+n_{\text{off}}(t,x))dx$. The fluorescence distribution is considered as a sample from the distribution $\frac{n_{\text{on}}(t,x)+n_{\text{off}}(t,x)}{\int_0^{X_{\max}}(n_{\text{on}}(t,x)+n_{\text{off}}(t,x))dx}$. As we have no information on the repartition on/off for the initial data, we apply the following rule: for t_0, starting of our simulation (DAY 2), we choose the repartition to be the same as in the steady distribution N. More precisely, we fix the proportion with the equation

$$n_{\text{on/off}}(t_0,x)=\underbrace{\frac{N_{\text{on/off}}(x)}{N_{\text{on}}(x)+N_{\text{off}}(x)}}_{\text{parameter dependent}}\underbrace{(n_{\text{on}}(t_0,x)+n_{\text{off}}(t_0,x)}_{\text{data}}). \tag{8}$$

Numerical scheme. For the Eq. (5), we use an explicit upwind scheme. Setting $a_{\text{on}}: x\longmapsto d(1-x)$ and $a_{\text{off}}: x\longmapsto -dx$, the scheme is given by

$$\begin{cases} \frac{{n_{\text{on}}}_j^{n+1}-{n_{\text{on}}}_j^n}{\Delta t}+\frac{{a_{\text{on}}}_{j+1/2}^n{n_{\text{on}}}_j^n-{a_{\text{on}}}_{j-1/2}^n{n_{\text{on}}}_{j-1}^n}{\Delta x}=-k_{\text{off}}{n_{\text{on}}}_j^n+k_{\text{on}}{n_{\text{off}}}_j^n+r_j{n_{\text{on}}}_j^n,\\ \frac{{n_{\text{off}}}_j^{n+1}-{n_{\text{off}}}_j^n}{\Delta t}+\frac{{a_{\text{off}}}_{j+1/2}^n{n_{\text{off}}}_{j+1}^n-{a_{\text{off}}}_{j-1/2}^n{n_{\text{on}}}_j^n}{\Delta x}=-k_{\text{on}}{n_{\text{off}}}_j^n+k_{\text{off}}{n_{\text{on}}}_j^n+r_j{n_{\text{off}}}_j^n, \end{cases} \tag{9}$$

with ${n_{\text{on/off}}}_j^n=n_{\text{on/off}}(j\Delta x,n\Delta t)$, $r_j=r(j\Delta x)$, ${a_{\text{on/off}}}_{j+1/2}^n=(a_{\text{on/off}}((j+1)\Delta x,n\Delta t)+a_{\text{on/off}}(j\Delta x,n\Delta t))/\Delta x$ and ${a_{\text{on/off}}}_{j-1/2}^n=(a_{\text{on/off}}(j\Delta x,n\Delta t)+a_{\text{on/off}}((j-1)\Delta x,n\Delta t))/\Delta x$. In the Results section, Fig. 7 illustrates a comparison between the result of the numerical scheme and the theoretical asymptotic profile of Eq. (5).

Estimation of the distance to the data.

To calibrate the parameter values of our system, we use our experimental data. Initially, in order to estimate the exponential growth rate of cells, we perform a linear regression analysis on the temporal evolution data of the cell count. To determine the values of other system parameters, we seek values that make our numerical results as close as possible to the experimental data. To characterize this notion of closeness between our numerical results and the data, we introduce the Kantorovich–Rubinstein distance. Given two probability distribution p_1,p_2 on $\mathbb{R}_+$, we define their cumulative distribution function $P_i(x)=Pr(X<x,\text{under distribution } p_i)=\int_0^x p_i(dx)$. Using these functions we can define the Kantorovich Rubinstein (also known as Wasserstein) distance by

$$dist_{\text{KT}}(p_1,p_2)=\int_0^{\infty}|P_1(x)-P_2(x)|\text{dx}. \tag{10}$$

In our specific case, we want to compare at each step the (normalized) distribution generated by the model at time t_i (hereby denoted as $model(t_i, dx)$ with cumulative distribution $M(t_i, .)$) and the corresponding distribution of the data at time t_i denoted as $data(t_i, dx)$with cumulative distribution $D(t_i, .)$. We would therefore compute

$$dist_{KT}(t_i) = dist_{\mathrm{KT}}(model(t_i), data(t_i)) = \int_0^\infty |M(t_i, x) - D(t_i, x)|dx.$$

Note that in our case the integral is in fact taken on the finite interval [0, 1] for scaled variables.

Considering the distribution profile of the data, we prefer to study this distance on a logarithmic scale. We therefore make the following change of variables $y = \log(xX_{\max} + 1)$, and we define the modified Kantorovich–Rubinstein distance as follows

$$dist_{\mathrm{KT}}^{\log}(t) = \int_0^{\log(X_{\max}+1)} \left|\tilde{D}(y, t) - \tilde{M}(y, t)\right| \mathrm{dy}, \tag{11}$$

with $\tilde{D}$ and $\tilde{M}$ the two cumulative distribution functions, defined below, in the new logarithmic scale. Note that, following this change of variable, this "distance" can be greater than 1.

We are looking for a function $\tilde{m}$ that satisfies the following relation, for all $b \in [0, \log(1 + X_{max})]$

$$\tilde{M}(b, t) = M\left(\frac{e^b - 1}{X_{max}}, t\right). \tag{12}$$

In particular, the link between the corresponding densities is immediately given by

$$\tilde{m}(b, t) = m\left(\frac{e^b - 1}{X_{max}}, t\right)\frac{e^b}{X_{max}}. \tag{13}$$

The space [0, 1] is discretized uniformly with $J + 1$ points, and this sequence is denoted $(x_j)_j$.

$$(x_j)_{j\in\{0,1,\ldots,J\}} : \ x_j = j\Delta x, \qquad \text{with } \Delta x = 1/J.$$

We also define the sequences $(y_j)_{j\in\{0,1,\ldots,J\}}$ and $(\ell_j)_{j\in\{0,1,\ldots,J-1\}}$ as follows

$$(y_j)_{j\in\{0,1,\ldots,J\}} : \ y_j = \log(X_{\max}x_j + 1) = \log\left(1 + \frac{X_{\max}j}{J}\right),$$

and

$$(\ell_j)_{j\in\{0,1,\ldots,J-1\}} : \ \ell_j = y_{j+1} - y_j.$$

Consequently, the estimator of the cumulative distribution function M is given by

$$\hat{M}(y_j,t)=\frac{1}{\sum_i \tilde{m}(y_i,t)\ell_i}\sum_{i<j}\tilde{m}(y_i,t)\ell_i=\frac{\sum_{i<j} m(x_i,t)\left(\frac{1}{X_{\max}}+x_i\right)\ell_i}{\sum_i m(x_i,t)\left(\frac{1}{X_{\max}}+x_i\right)\ell_i},\quad \forall j\in\{0,1,\dots,J\}.$$

where we have used (13) to estimate $\tilde{m}$. We need to renormalize to ensure we are comparing probability distributions.

The estimator of the cumulative distribution function D is

$$\hat{D}(y_j,t)=\frac{1}{\#\{d_k(t),\ \forall k\}}\sum_{i<j}h_i,\quad \forall j\in\{0,1,\dots,J\},$$

with $h_j(t)=\#\{d_i:\ \log(d_i(t)+1)\in[y_{j-1},y_j]\}$, where the operator # corresponds to the cardinal of a set and the data d_i correspond to the fluorescence data obtained after data processing. These data, d_i, are real numbers between 0 and $X_{\max}$. The term $\#\{d_k(t),\ \forall k\}$ corresponds to the number of cells present in the data on day t after the gating operation.

Therefore, the distance between the experimental data and the mathematical model is as follows

$$\widehat{dist}^{\log}_{\mathrm{KT}}(t;\text{parameters})=\sum_{j=0}^{J-1}\left|\hat{D}(y_j,t)-\hat{M}(y_j,t;\text{parameters})\right|\ell_j. \tag{14}$$

To calibrate the parameters of our model, we will minimize the sum of the modified Kantorovich–Rubinstein distance for the different days at our disposal and for the two experiments. The distance associated with $CD34^+$ data is denoted $\widehat{dist}^{\log,+}_{\mathrm{KT}}$, while that associated with $CD34^-$ data is denoted $\widehat{dist}^{\log,-}_{\mathrm{KT}}$. We also introduce the distance, denoted $\widehat{dist}^{\log,\pm}_{\mathrm{KT}}$, corresponding to the sum of these two distances. Thereby we look for one set of parameters for fitting the two datasets jointly. Mathematically, the optimization problem is given by the following formula

$$\text{parameters}^{\text{opt}}=\underset{\text{parameters}}{\arg\min}\left(\sum_{t\in\text{Days}}\widehat{dist}^{\log,\pm}_{\mathrm{KT}}(t;\text{parameters})\right). \tag{15}$$

The results of this optimization work are detailed in the Results section.

Results

Mathematical analysis—derivation of explicit normalized asymptotic profile (N_{on}, N_{off})

Under the assumption of a linear proliferation rate $r(x)=r_0+r_1x$, we obtain the following result

Theorem 1 *Assume $r(x)=r_0+r_1x$. Define the matrix M by*

$$\begin{pmatrix} r_1-k_{\text{off}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{on}}\end{pmatrix}.$$

Denote $s(M)$ the largest eigenvalue of M, then the left and right eigenvector

$$V^t M = s(M)V^t M, \quad MU = s(M)U,$$

can be chosen positive. Moreover, the solution to (6) *is given by*

$$\lambda = r_0 + s(M) = r_0 + \frac{r_1 - k_{\text{on}} - k_{\text{off}} + \sqrt{(r_1 - k_{\text{on}} - k_{\text{off}})^2 + 4k_{\text{on}}r_1}}{2}. \tag{16}$$

And

$$\begin{pmatrix} N_{\text{on}}(x) \\ N_{\text{off}}(x) \end{pmatrix} = \begin{pmatrix} C(1-x)^{\frac{k_{\text{off}}}{d\eta}-1} x^{\frac{k_{\text{on}}\eta}{d}} e^{\frac{r_1}{d}x} \\ C\eta(1-x)^{\frac{k_{\text{off}}}{d\eta}} x^{\frac{k_{\text{on}}\eta}{d}-1} e^{\frac{r_1}{d}x} \end{pmatrix}, \tag{17}$$

with C an arbitrary positive constant and η *given by* $\eta = 1 + s(M)/k_{\text{on}}$.

In particular, we have

$$N(x) = C(1-x)^{\frac{k_{\text{off}}}{d\eta}-1} x^{\frac{k_{\text{on}}\eta}{d}-1} e^{\frac{r_1}{d}x}(\eta + (1-\eta)x), \tag{18}$$

Proof

We notice that the existence and uniqueness (up to a multiplicative factor) of the triplet s(M), V, U is a straightforward consequence of the classical Perron Frobenius theorem which applies here because the off-diagonal entries of M are positive [28]. We introduce the ration $\eta = \frac{V_{\text{on}}}{V_{\text{off}}}$ *and notice that by construction, we have*

$$\eta = \frac{V_{\text{on}}}{V_{\text{off}}} = \frac{k_{\text{off}}}{s(M) - r_1 + k_{\text{off}}} = \frac{s(M) + k_{\text{on}}}{k_{\text{on}}}. \tag{19}$$

Then, consider the system satisfied by $\psi_{\text{on,off}} = e^{r_1 x}\phi_{\text{on,off}}$, where $\phi_{on/off}$ is the solution of (7). We get

$$\begin{cases} \lambda\psi_{\text{on}} - d(1-x)\partial_x\psi_{\text{on}} = -k_{\text{off}}\psi_{\text{on}} + k_{\text{off}}\psi_{\text{off}} + (r_0 + r_1)\psi_{\text{on}}(t,x), \quad x \in]0,1[, \\ \lambda\psi_{\text{off}} - d(-x)\partial_x(\psi_{\text{off}}) = -k_{\text{on}}\psi_{\text{off}} + k_{\text{on}}\psi_{\text{on}} + r_0\psi_{\text{off}}(x), \quad x \in]0,1[. \end{cases}$$

This system can be summarized as

$$d\begin{pmatrix} -(1-x)\partial_x\psi_{\text{on}} \\ -(-x)\partial_x\psi_{\text{off}} \end{pmatrix} = \left(\begin{pmatrix} r_0 + r_1 - k_{\text{off}} & k_{\text{off}} \\ k_{\text{on}} & r_0 - k_{\text{on}} \end{pmatrix} - \lambda I_2\right)\begin{pmatrix} \psi_{\text{on}} \\ \psi_{\text{off}} \end{pmatrix}.$$

We recognize the matrix $r_0 + M^t$. Therefore, we have a solution independent on x $\psi_{\text{on,off}} = V_{\text{on,off}}$ and $\lambda = r_0 + s(M)$ and $\phi_{\text{on,off}} = e^{-\frac{r_1}{d}x}V_{\text{on,off}}$. Similarly, if we denote 0

$$\begin{cases} \lambda P_{\text{on}} + d\partial_x((1-x)P_{\text{on}}) = -k_{\text{off}}P_{\text{on}} + k_{\text{on}}P_{\text{off}} + (r_0 + r_1)P_{\text{on}}, \quad x \in]0,1[, \\ \lambda P_{\text{off}} + d\partial_x((-x)P_{\text{off}}) = k_{\text{off}}P_{\text{on}} - k_{\text{on}}P_{\text{off}} + r_0P_{\text{off}}, \quad x \in]0,1[, \\ P_{\text{on}}(0) = P_{\text{off}}(1) = 0, \quad P_{\text{on,off}} > 0. \end{cases}$$

This can be condensed into

$$d\partial_x \begin{pmatrix} (1-x)P_{\text{on}} \\ -xP_{\text{off}} \end{pmatrix} = (r_0 - \lambda + M) \begin{pmatrix} P_{\text{on}} \\ P_{\text{off}} \end{pmatrix}.$$

We can now proceed as for the conservative case and notice that since $V^t(r_0 - \lambda - M) = 0$ multiplying the equation by V.

$$\partial_x(V_{\text{on}}(1-x)P_{\text{on}}(x) - xV_{\text{off}}P_{\text{off}}(x)) = 0,$$
$$P_{\text{off}}(x) = \frac{V_{\text{on}}(1-x)}{xV_{\text{off}}}P_{\text{on}}.$$

After the appropriate substitution, we obtain

$$\begin{aligned} d\partial_x((1-x)P_{\text{on}}) &= \left(\frac{-k_{\text{off}} + r_0 + r_1 - \lambda}{(1-x)} + \frac{k_{\text{on}}V_{\text{on}}}{xV_{\text{off}}}\right)(1-x)P_{\text{on}} \\ &= \left((r_1 - k_{\text{off}} - s(M))\frac{1}{(1-x)} + (s(M) + k_{\text{on}})\frac{1}{x}\right)(1-x)P_{\text{on}}. \end{aligned}$$

This leads to, for a suitable renormalization constant $C > 0$,

$$\begin{cases} P_{\text{on}} = C(1-x)^{\frac{k_{\text{off}}+s(M)-r_1}{d}-1}x^{\frac{k_{\text{on}}+s(M)}{d}}, \\ P_{\text{off}} = C\frac{V_{\text{on}}}{V_{\text{off}}}(1-x)^{\frac{k_{\text{off}}+s(M)-r_1}{d}}x^{\frac{k_{\text{on}}+s(M)}{d}-1}. \end{cases}$$

We introduce then the notation $\eta = \eta(r_1) = \frac{V_{\text{on}}}{V_{\text{off}}}$ and go back the N variables to write

$$\begin{cases} N_{\text{on}} = C(1-x)^{\frac{k_{\text{off}}}{d\eta}-1}x^{\frac{k_{\text{on}}\eta}{d}}e^{\frac{r_1}{d}x}, \\ N_{\text{off}} = C\eta(1-x)^{\frac{k_{\text{off}}}{d\eta}}x^{\frac{k_{\text{on}}\eta}{d}-1}e^{\frac{r_1}{d}x}. \end{cases} \tag{20}$$

In particular, we have

$$N(x) = C(1-x)^{\frac{k_{\text{off}}}{d\eta}-1}x^{\frac{k_{\text{on}}\eta}{d}-1}e^{\frac{r_1}{d}x}(\eta + (1-\eta)x).$$

Going back to the definition of η we notice

$$k_{\text{on}}\eta = k_{\text{on}} + s(M), \quad k_{\text{off}}/\eta = k_{\text{off}} + s(M) - r_1. \tag{21}$$

□

Simulation analysis

Estimation of exponential growth rate.

Using experimental data on cell numbers for different days, we estimate the exponential growth rate λ. For both relaxation experiments, we perform a linear regression of the natural logarithm of the number of cells. In the case of $CD34^+$ cell relaxation, the linear regression line is given by the slope $\lambda^+ \approx 0.418$, and for $CD34^-$ cells, the slope is $\lambda^- \approx 0.422$. We estimate the parameter λ by the average of these two slopes, $\lambda \approx 0.42$. Figure 4 shows that the estimate of the exponential growth rate is in good agreement with the experimental data.

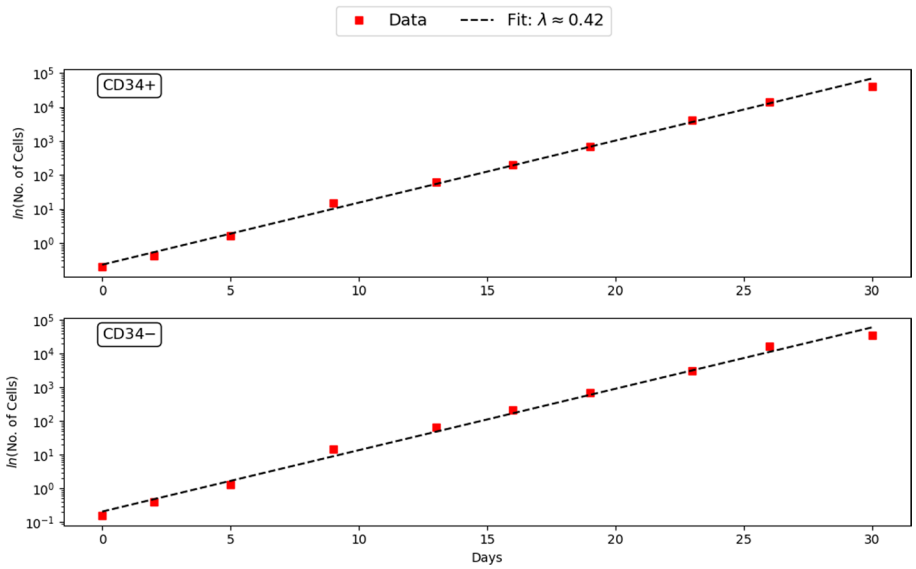

Fig. 4 Estimation of the exponential growth rate λ. The red squares correspond to the number of cells (log scale) at different times for the relaxation experiments: *Top.* $CD34^+$, *Bottom.* $CD34^-$. The dotted line in black illustrates the optimal fit of the experimental data. The average of the slopes of the linear regressions minimizing the two experiments is given by the slope $\lambda \approx 0.42$

Table 1 Estimated parameter values

Parameter	value	Units	Description	Estimation method
$X_{\max}$	2×10^4	proteins	Maximum value for the quantity of CD34 in a cell	Data-driven selection
r_0	0.426	h^{-1}	Constant proliferation rate	Estimating the proliferation rate λ and using the relation (16)
r_1	−0.426	h^{-1}	Linear proliferation rate	KT distance minimization (22)
k_{on}	0.261	h^{-1}	Rate at which the gene/promoter is turned "on"	KT distance minimization (22)
k_{off}	19.178	h^{-1}	Rate at which the gene/promoter is turned "off"	KT distance minimization (22)
d	0.21	h^{-1}	Degradation rate	KT distance minimization (22)
s	4214	proteins.h^{-1}	Synthesis rate of proteins when gene is on	Relation: $s = dX_{\max}$

Calibration of parameters.

A study of the maximum for each day and each experiment of the "PE-Cy7-A" fluorescence data reveals an $X_{\max}$ close to 20,000. To reduce the numerical complexity of the 5-parameter optimization $(r_0, r_1, k_{on}, k_{off}, d)$, we will use the estimate of λ to reduce our optimization problem to just 4 parameters. Indeed, using the theoretical relationship (16) and the previous estimate of λ, we can define the parameter r_0 as a function of the other model parameters,

$$r_0 = s(M) - \hat{\lambda} = s(M) - 0.42.$$

For the estimation of the other parameters r_1, k_{on}, k_{off} and d, we use the modified Kantorovich–Rubinstein distance minimization strategy, presented in the Method section and given by the following formula

$$(r_1^\star, k_{on}^\star, k_{off}^\star, d^\star) = \underset{r_1,k_{on},k_{off},d}{\arg\min} \left(\sum_{t\in \text{Days}} \widehat{dist}_{KT}^{\log,\pm}(t; r_1, k_{on}, k_{off}, d) \right). \tag{22}$$

To determine numerically this minimum, we calculate the modified Kantorovich–Rubinstein distance for the two relaxation experiments, for each point on a large grid of parameter values. We then adjust our grid to obtain the location of the minimum of the sum of two distances. This method gives us the following four parameter values $r_1^\star = 0.426$, $k_{on}^\star = 0.261$, $k_{off}^\star = 19.178$ and $d^\star = 0.21$. Following this optimization strategy, the optimal choice for the function r is to choose r_1 such that $r(x) = r_0 \times (1 - x)$. Nevertheless, we will see in the next section that the choice of r_1 is not decisive for a good fit between the model result and the experimental data.

Finally, using the relation, $s = dX_{max}$, we can calculate the value of the synthesis rate, $s = 4214$. All parameter value estimates are given in Table 1.

We compared those values to the litterature. The estimated half-life of CD34 in this model was estimated to be equal to $log(2)/0.261$, that is about 4 h. This is in the low range of the estimated distribution for proteins half-life [29]. Regarding the X_{max} estimation, its value is in the range of observed value, slightly over the median that is 1.6×10^4 [29]. The estimated k_{on} value gives an estimated frequency of 4 (1/0.261) bursts per hour on average, which is close from the expected range from a burst every 30 min to up to 10 h [30]. The k_{off} value display the expected ratio ($k_{on} << k_{off}$ an $d1 << k_{off}$) in the case of a bursty regime [9]. Altogether all of our estimated parameters are thus in the expected range.

Profile likelihood.

To investigate the robustness of our parameter estimates and the significance of each parameter in minimizing the modified Kantorovich–Rubinstein distance, we employ an approach analogous to the *profile likelihood* concept [31, 32] in the context of our optimization problem.

First, we examine the influence of the parameter d. Let d be fixed, we calculate, in the same way, the triplet of parameters (r_1, k_{on}, k_{off}) that minimizes the modified Kantorovich–Rubinstein distance under the fixed d constraint. These optimal parameters are, therefore, functions dependent on d, denoted as $\widehat{r_1}$, $\widehat{k_{on}}$, and $\widehat{k_{off}}$. Mathematically, they are defined by the following relation

$$\left(\widehat{r_1}(d), \widehat{k_{on}}(d), \widehat{k_{off}}(d)\right) = \underset{r_1,k_{on},k_{off}}{\arg\min} \left(\sum_{t\in \text{Days}} \widehat{dist}_{KT}^{\log,\pm}(t; r_1, k_{on}, k_{off}, d) \right).$$

Once these functions have been calculated, we can determine the modified Kantorovich–Rubinstein distance associated with them, denoted by S_d and defined by the following equality,

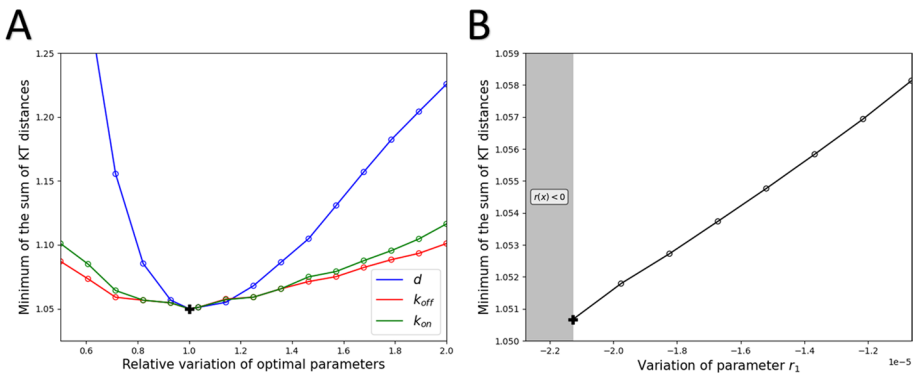

Fig. 5 Likelihood profiles for for k_{on}, k_{off} and d in A and for r_1 in B. A. The blue curve represents the function S_d, the red curve $S_{k_{off}}$ and the green curve $S_{k_{on}}$. The function S_d is introduced into Eq. (23). B. The grey area corresponds to the range of parameter values for r_1 such that the function r is non-positive. Compared with the other parameters, variation in the r_1 parameter has a small impact on the minimum Kantorovich–Rubinstein distance

$$S_d(d/d^\star) = \sum_{t\in \text{Days}} \widehat{dist}_{\text{KT}}^{\log,\pm}\left(t; \widehat{r_1}(d), \widehat{k_{\text{on}}}(d), \widehat{k_{\text{off}}}(d), d\right). \tag{23}$$

For the argument of the function, we choose $d/d^\star$, to study the distance associated with the relative variation of the optimal parameter. By definition of the function S_d, it reaches its minimum at $d = 1$, corresponding to $d = d^\star$. Similarly, we can define the functions $S_{k_{\text{on}}}$, $S_{k_{\text{off}}}$, and S_{r_1}.

In Fig. 5.A, we plotted the S_d, $S_{k_{\text{on}}}$ and $S_{k_{\text{off}}}$ functions. The impact of the relative variation of the two transition rates around the optimal value, on the modified Kantorovich–Rubinstein distance is quite similar. For the degradation rate, d, we note that a fine estimate of this is crucial to obtain good accuracy between the data and the mathematical model.

Conversely, the parameter r_1 has a minor impact on the minimum of the modified Kantorovich–Rubinstein distance. Specifically, when r_1 deviates from its optimal value, new optimal parameter values emerge, resulting in distances very close to the optimal distance. This result is illustrated in Fig. 5.B.

Comparison between model and experimental data. In Fig. 6 we compared data from relaxation experiments with the results of our model for the parameter values presented in Table 1.

To initialize our model on day 2, we use the Eq. (8), it follows this following initial conditions

$$n_{\text{on/off}}(t_0, x) = \frac{N_{\text{on/off}}(x)}{N_{\text{on}}(x) + N_{\text{off}}(x)} \times \sum_{j=0}^{J-1} h_j \mathbf{1}_{x\in[x_j, x_{j+1}]}(x), \qquad x \in [0,1], \tag{24}$$

where h_j corresponds to the number of cells on day 2 with fluorescence between $X_{\max} \times x_j$ and $X_{\max} \times x_{j+1}$, where $(x_j)_j$ corresponds to the uniform discretization of

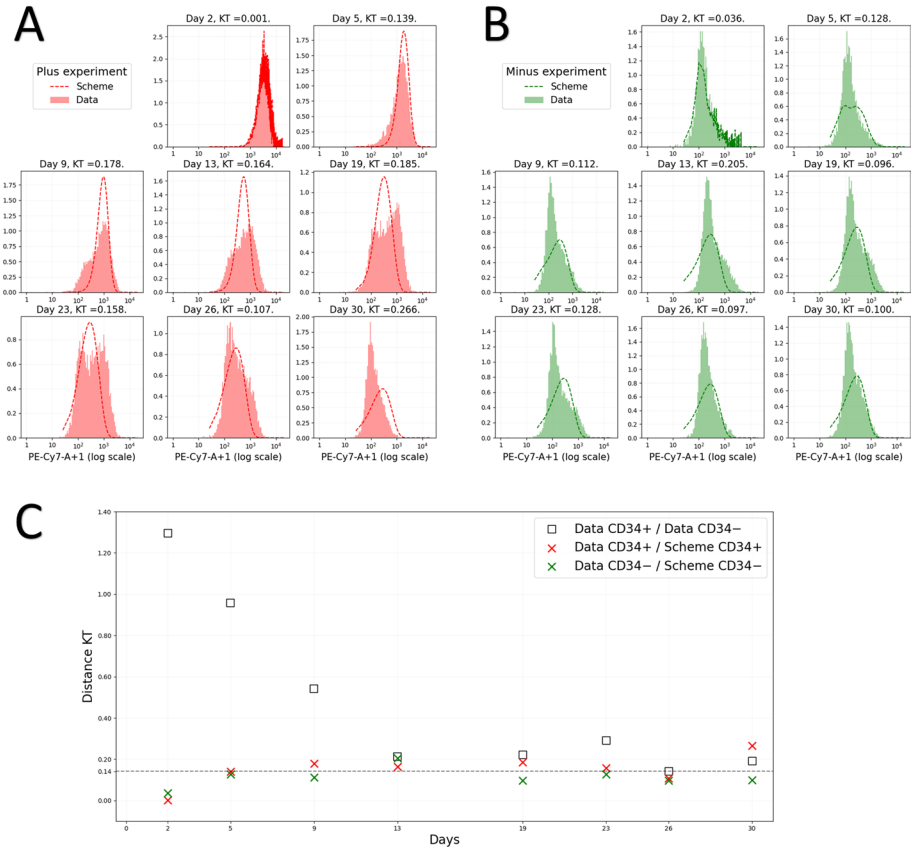

Fig. 6 Comparison of model and data. On the left the fitting of the $CD34^+$ relaxation experiment (in A) and on the right in green of the $CD34^-$ (in B) experiments. Experimental data in logarithmic scale are represented by plain histograms and the numerical results of model (5) are represented by the dotted curves. We initialize the model on day 2, using the biological data. The initial condition is given by (24). Parameter values are given in Table 1. KT: the modified Kantorovich–Rubinstein distance, defined by the Eq. (14). C. Time-dependent evolution of the Kantorovich–Rubinstein distance between model and experimental data. For different days of the experiment, the modified Kantorovich–Rubinstein distance between the two relaxation experiments is depicted using black squares. The minimum distance, reached on day 26, is illustrated by a horizontal dotted line. The red crosses correspond to the modified Kantorovich–Rubinstein distance between the model for the parameter values from Table 1, and the $CD34^+$ cell relaxation experiment. Similarly, the green crosses represent the distance for the $CD34^-$ cell relaxation experiment

space [0, 1]. That is, the $(h_j)_j$ correspond to the heights of the histogram of the data renormalized by the maximum $X_{\max}$.

Visually, Fig. 6.A,B reveals a high degree of proximity between the experimental data and the mathematical model. To quantify this closeness, we once again employ the modified Kantorovich–Rubinstein distance. In Fig. 6.C, we represent, by black squares, the temporal evolution of distance between the experimental data of the two relaxation experiments. Due to the antinomic nature of the two experiments, distances are considerable in the early days of the experiment. It then gradually decreases as the two distributions converge towards the stationary distribution. From day 26, both profiles reached the stationary stage. At this point, in the absence of noise, these two distributions are expected to be similar. Therefore, the minimum

distance, dist = 0.142 (illustrated by a dotted line), attained on day 26, corresponds to a reference distance to determine the proximity of two distributions.

In this figure, we also illustrate the distance between the model and the two relaxation experiments. The $CD34^+$ cell relaxation experiment is represented by the red crosses, and the $CD34^-$ cell relaxation experiment by the green crosses. To quantitatively assess the proximity of the model to experimental data, we employ the reference distance represented by the horizontal line. For the $CD34^-$ cell relaxation experiment, we observe that the distances are always less than the reference distance, except on day 13 for which the distance is slightly greater. Concerning the $CD34^+$ cell relaxation experiment, this time the distances are more regularly greater than the reference distance, but are still within an acceptable order of magnitude. These results show that our proposed model is very close to the experimental data.

Discussion

Although we had to infer a number of model parameters which could not be deduced from the literature (like for example the half-life of the CD34 protein), the overall fitting ability of our model proved to be quite satisfactory. Using Kantorovich distances, we indeed observed that the model-to-experiment distance was within the range of the experiment-to-experiment distance, so in the range of experimental variability.

We assumed that the proliferation rate would depend upon the level of expression of the very gene that is being modelled. In our case, that proved to be useful since we wanted to fit relaxation data obtained from CD34 expression. CD34 is a known marker for stemness and we hypothesized that, in line with the existing litterature [25], $CD34^+$ cells would proliferate less that $CD34^-$ more mature cells. We should nevertheless stress that such a behaviour can be true for normal hematopoietic stem cells, but can be questioned regarding cancer stem cells.

The methodology presented in this paper is applicable to any cell systems for which one can perform simultaneously relaxation experiments and proliferation measurements. Biological systems for which the half-life of the protein of interest is known should be preferred, since this will remove the need for estimating an important model parameter.

One of the difficulties we faced when comparing the model's output with experimental data, lies in the need for common units By default, our model output is a value between 0 (no CD34 expressed) and 1 (maximum level of CD34 expression). FACS data are corrected fluorescent values, that can be negative in the raw acquisition dataset. We therefore processed the data with a gating phase, a shifting phase, and finally normalized them in order to obtain comparable values with the model.

It is crucial to emphasize that within a cell population displaying a stationary distribution of phenotypic states, no cell remains in a permanent state over time. Given a sufficiently long time, one can assume that all cells will have visited all possible states (i.e. all possible values for their surface CD34 expression). In other terms, in the state versus identity long standing debate [33], we clearly side with the view that stemness is an emerging dynamical property

Several points shall be investigated further. The first point that can be enriched is the form of the division rate $r(x) = r_0 - r_1 x$. The linear form ensures the explicit formulation of the stable distribution and facilitates the scaling by X_{max} but is not necessary for the existence of a profile. Moreover, we made strong assumptions here that the daughter cells have the same concentration of markers than the mother cell and that the division has no impact on the on/off status of the cell. This later is a reasonable assumption in the light of the existence of transcriptional memory [20, 21], but it might be gene-dependent.

One missing aspect of our model is the absence of any explicit death term. On the other hand, an expression independent death rate could immediately be considered by relaxation of the constraint of positive division rate (which would then correspond to a net growth rate). In terms of parameters, this would affect r_0.

Another missing aspect of our model is the fact that the CD34 gene expression is modelled in isolation. It is quite obvious that in cells its expression level will be constrained by its positioning in a complex web of genes-to-genes interactions known as a Gene Regulatory Network (GRN). Inference of such GRNs is a notoriously difficult task (see e.g. [34]), and performing relaxation experiments from such complex objects is yet to be done.

One of the future goal of our work would be to assess its predictive ability. A promising lead would be to go further in the analysis in order to estimate the influence of parameters on the relaxation time. Mathematically, this could be analysed through the spectral gap which is beyond the scope of this work. It would be especially interesting to identify the effects of various parameters on it, in particular the parameter d which represents the degradation rate. Note that in this case, the distribution and the value of λ are expected to change. Interestingly, the model's prediction in this case could be tested experimentally by modifying the endogenous CD34 protein stability.

Conclusion

In the present work, we proposed a revised two-state probabilistic model for gene expression which explicitly incorporates a proliferation term. This model was analysed and we obtained an analytical solution for our model's steady state. The same model was then used for simulating the transient behavior of FACS-sorted cells leading to the progressive relaxation towards the steady state distribution. Altogether, our work shows that a two-state description for CD34 gene expression is well suited to explain the relaxation experiments. This support the notion that cells should be seen and modeled as probabilistic dynamical systems.

Supplementary Information

Appendix A: Supplementary Figures

See Fig. 7.

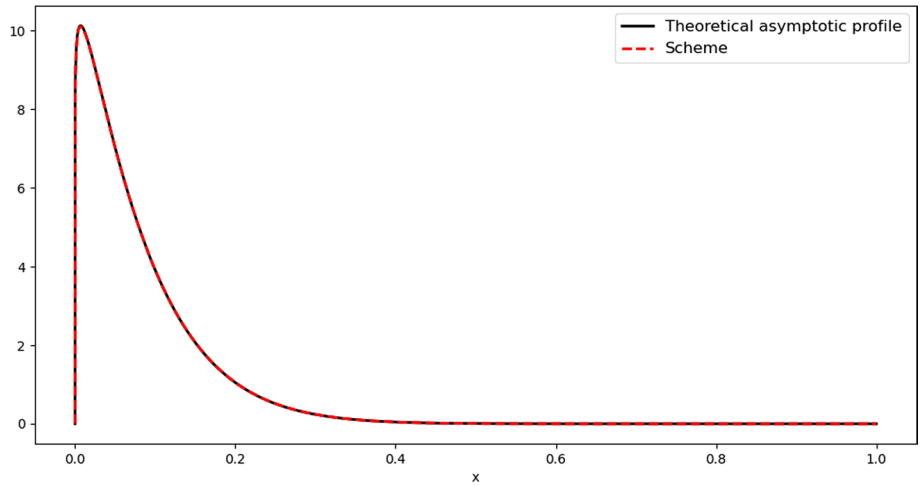

Fig. 7 Comparison between the theoretical asymptotic profile N (black line) presented in Theorem 1 and the numerical solution obtained using the numerical scheme (red dotted line) defined by Eq. (9). Parameter values are given by $r_0 = 3.4, r_1 = 3, k_{on} = 0.85, k_{off} = 12.71, d = 1, X_{max} = 2 \times 10^4$

Abbreviations

GRN Gene regulatory network
PDMP Piecewise deterministic Markov process

Acknowledgements
We thank Ulysse Herbach for his help with inferring 2-states parameter values. We also thank the BioSyL Federation and the LabEx Ecofect (ANR-11-LABX-0048) of the University of Lyon for inspiring scientific events.

Author Contributions
ME, TL and GR performed the mathematical derivation of the profile. VMS, SL and TV produced the data. MD, OG and ME analyzed and shaped the data. MD and ME performed the numerical simulations. ME performed the parameter inference. ME, OG and TL interpreted the results. ME, OG and TL wrote the manuscript. All authors but TV read and approved the final manuscript.

Funding
Maxime Estavoyer is funded by ANR PLUME (ANR-21-CE13-0040). For this project, Marion Dufeu was funded by IXXI (http://www.ixxi.fr)

Availability of data and materials
The data described in this article can be freely and openly accessed at gitlab: https://gitlab.inria.fr/lepoutre_public/relaxation_modeling.

Declarations

Conflict of interest
None

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Received: 21 March 2024 Accepted: 20 May 2024
Published online: 20 August 2024

References

1. Nicholson DJ. Is the cell really a machine? J Theor Biol. 2019;477:108–26. https://doi.org/10.1016/j.jtbi.2019.06.002.
2. Kupiec JJ. A probabilistic theory for cell differentiation, embryonic mortality and DNA c-value paradox. Specul Sci Technol. 1983;6(5):471–8.

3. Noble D. Genes and causation. Philos Transact A Math Phys Eng Sci. 2008;366(1878):3001–15. https://doi.org/10.1098/rsta.2008.0086.
4. Schrödinger E. What is life? The physical aspect of the living cell. Cambridge: Cambridge University Press; 1944.
5. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453(7194):544–7.
6. Kalmar T, Lim C, Hayward P, Munoz-Descalzo S, Nichols J, Garcia-Ojalvo J, Martinez Arias A. Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol. 2009;7(7):1000149. https://doi.org/10.1371/journal.pbio.1000149.
7. Sisan DR, Halter M, Hubbard JB, Plant AL. Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. Proc Natl Acad Sci USA. 2012;109(47):19262–7. https://doi.org/10.1073/pnas.1207544109.
8. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys. 1976;22(4):403–34.
9. Herbach U, Bonnaffoux A, Espinasse T, Gandrillon O. Inferring gene regulatory networks from single-cell data: a mechanistic approach. BMC Syst Biol. 2017;11(1):105. https://doi.org/10.1186/s12918-017-0487-0.
10. Ko MS. A stochastic model for gene induction. J Theor Biol. 1991;153:181–94.
11. Peccoud J, Ycart B. Markovian modelling of gene product synthesis. Theor Popul Biol. 1995;48:222–34.
12. Raj A, Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135(2):216–26.
13. Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M, Tay S. Digital quantification of proteins and MRNA in single mammalian cells. Mol Cell. 2016;61(6):914–24. https://doi.org/10.1016/j.molcel.2016.02.030.
14. Sarkar A, Stephens M, Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. bioRxiv, 2020; 2020–0407030007 https://doi.org/10.1101/2020.04.07.030007
15. Herbach U. Stochastic gene expression with a multistate promoter: breaking down exact distributions. SIAM J Appl Math. 2019;79:1007–29.
16. Bressloff PC. Stochastic processes in cell biology, vol. 41. Cham: Springer; 2014.
17. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic MRNA synthesis in mammalian cells. PLoS Biol. 2006;4(10):309.
18. Karmakar R, Bose I. Graded and binary responses in stochastic gene expression. Phys Biol. 2004;1(4):197.
19. Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. Nature. 2018;553(7689):418–26. https://doi.org/10.1038/nature25022.
20. Fourneaux C, Racine L, Koering C, Dussurgey S, Vallin E, Moussy A, Parmentier R, Brunard F, Stockholm D, Modolo L, Picard F, Gandrillon O, Paldi A, Gonin-Giraud S. Differentiation is accompanied by a progressive loss in transcriptional memory. BMC Biol. 2024;22(1):58.
21. Phillips NE, Mandic A, Omidi S, Naef F, Suter DM. Memory and relatedness of transcriptional activity in mammalian cell lineages. Nat Commun. 2019;10(1):1208.
22. Perthame B. Transport equations in biology. Cham: Springer; 2006.
23. Chereda B, Melo JV. Natural course and biology of cml. Ann Hematol. 2015;94(Suppl 2):107–21. https://doi.org/10.1007/s00277-015-2325-z.
24. Laperrousaz B, Jeanpierre S, Sagorny K, Voeltzel T, Ramas S, Kaniewski B, Ffrench M, Salesse S, Nicolini FE, Maguer-Satta V. Primitive cml cell expansion relies on abnormal levels of bmps provided by the niche and on bmprib overexpression. Blood. 2013;122(23):3767–77. https://doi.org/10.1182/blood-2013-05-501460.
25. Sidney LE, Branch MJ, Dunphy SE Dua, Harminder S, Hopkinson, A. Concise review: evidence for cd34 as a common marker for diverse progenitors. Stem Cells. 2014;32(6):1380–9. https://doi.org/10.1002/stem.1661.
26. Vershik AM. Long history of the Monge–kantorovich transportation problem. Math Intell. 2013;35:1–9.
27. Castillo-Hair SM, Sexton JT, Landry BP, Olson EJ, Igoshin OA, Tabor JJ. Flowcal: a user-friendly, open source software tool for automatically converting flow cytometry data from arbitrary to calibrated units. ACS Synth Biol. 2016;5(7):774–80.
28. Horn RA, Johnson CR. Matrix analysis. Cambridge university press, 2012.
29. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337–42. https://doi.org/10.1038/nature10098.
30. Nicolas D, Phillips NE, Naef F. What shapes eukaryotic transcriptional bursting? Mol BioSyst. 2017;13(7):1280–90.
31. Venzon D, Moolgavkar S. A method for computing profile-likelihood-based confidence intervals. J Roy Stat Soc: Ser C. 1988;37(1):87–94.
32. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25(15):1923–9.
33. Zipori D. The nature of stem cells: state rather than entity. Nat Rev Genet. 2004;5(11):873–8.
34. Ventre E, Herbach U, Espinasse T, Benoit G, Gandrillon O. One model fits all: Combining inference and simulation of gene regulatory networks. PLoS Comput Biol. 2023;19(3):1010962. https://doi.org/10.1371/journal.pcbi.1010962.

Publisher's Note