Estimer les effets des mutations sur la valeur sélective d'une bactérie

Jérémy Andréoletti et Nathanaël Boutillon
Projet encadré par Marie Doumic et Lydia Robert

22 février 2021

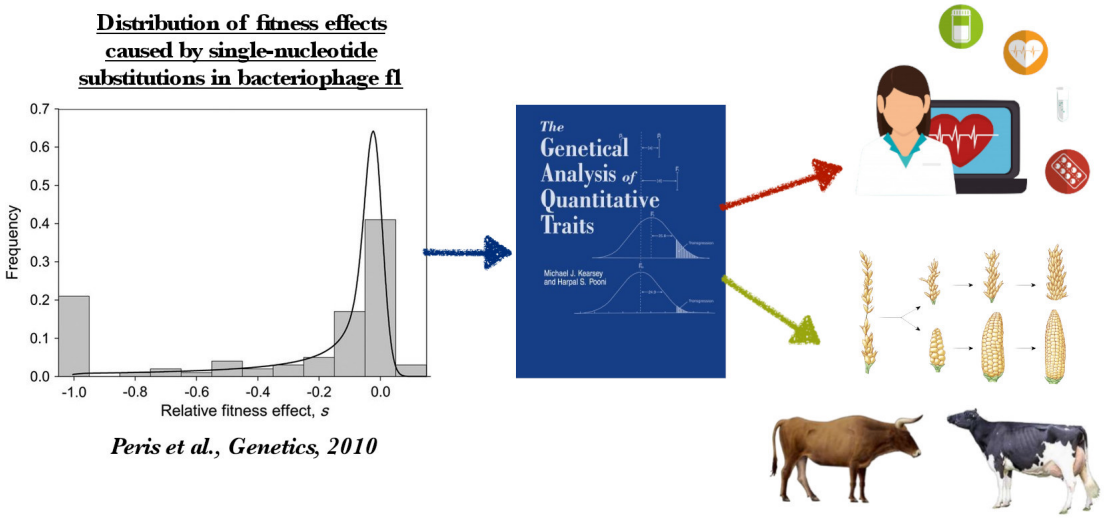
Table des matières

1 Introduction

NOTE : pas oublier de préciser ce qu'on a fait nous et ce qu'on n'a pas fait nous.

1.1 Contexte

La génétique des traits quantitatifs [1] est l'étude des phénotypes qui sont modulés génétiquement par un très grand nombre de loci à faible effets, et plus spécifiquement des liens entre variations génotypiques à ces loci et variations phénotypiques des traits associés. Bien comprendre ces mécanismes est essentiel non seulement en médecine humain, afin de prédire les risques de certaines maladies génétiques et mettre au point des traitements individuels, mais aussi en agriculture, afin de guider les programme de sélection artificielle de plantes ou animaux domestiques. Et un élément important de ce processus est d'être capable de prédire les effets sur la fitness de chaque nouvelle mutation – par exemple on s'attend à avoir beaucoup de mutations faiblement délétères et peu de mutations avantageuses ou fortement délétères – et dans l'idéal avoir accès à une distribution de probabilité complète de ces effets, que l'on appellera à partir de maintenant la **DFE = Distribution of Fitness Effects**.



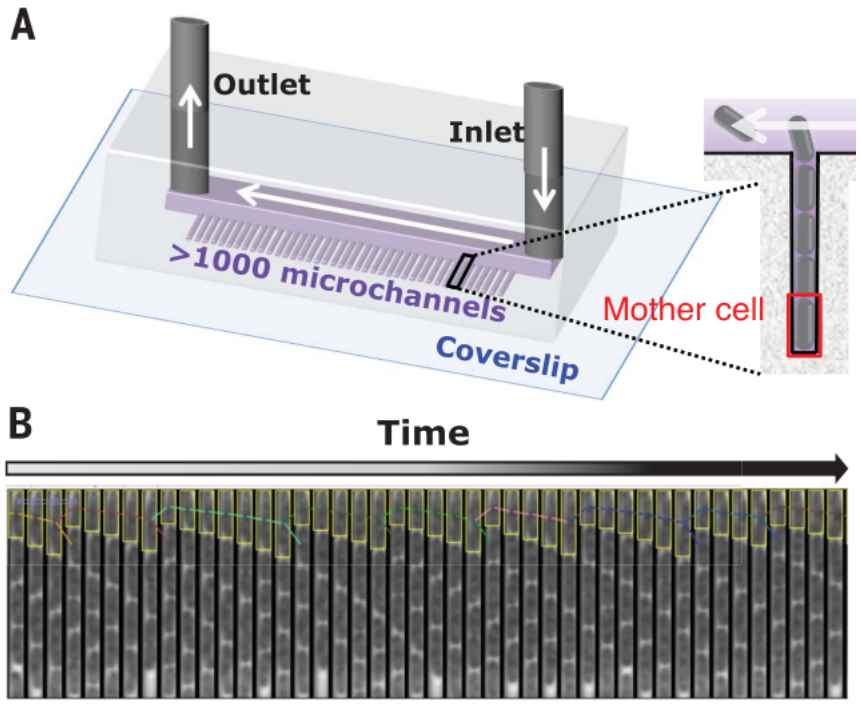
Le point de départ de ce projet est un article de Lydia Robert *et al.* [2] dont l'objectif est justement d'aider à comprendre l'impact des mutations sur la fitness. En particulier, les auteur.ice.s ont mis au point un protocole permettant pour la première fois de suivre en temps réel les dynamiques d'apparition des mutations et de croissance-fragmentation chez *Escherichia coli*. Dans les expériences classiques d'accumulation de mutations dans des colonies bactériennes, les observations sont moyennées sur plusieurs générations, induisant des biais par l'effet de la sélection naturelle éliminant les mutations trop délétères, voire létales, et par le nombre limité de lignées pouvant être suivies.

1.2 Expériences

Dans l'article de Lydia Robert *et al.* [2] sont présentées 2 types d'expériences dépassant une partie de ces limitations, en se basant sur l'observation de bactéries piégées dans 1000 micro-canaux d'une puce micro-fluidique :

1. **microfluidic Mutation Accumulation (μMA) experiment** : Suivi du taux d'élongation (proxy de la fitness) des bactéries sur de nombreuses générations en supprimant les effets de la sélection naturelle.
2. **Mutation Visualization (MV) experiment** : Détection en temps réel des mutation apparaissant dans l'ADN bactérien ;

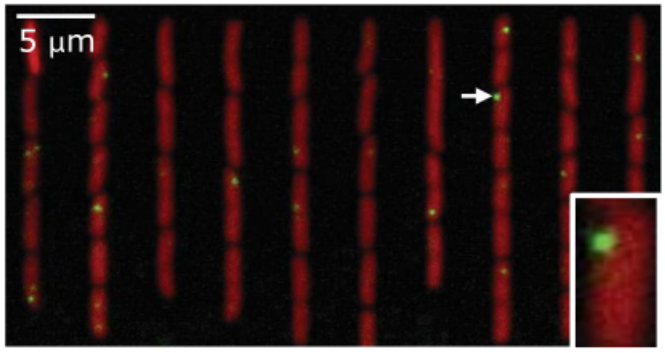
Détails de l'expérience μMA Dans cette expérience les bactéries sont placées dans une *Mother Machine* dans laquelle ~ 1500 cellules-mères vont se multiplier dans des micro-canaux individuels, leurs descendantes étant évacuées au fur et à mesure. En observant ces cellules au cours du temps, comme représenté dans la figure ci-dessous tirée de l'article, on peut ainsi reconstruire l'évolution des taux de croissance sur plusieurs générations et essayer d'extraire la part de ses variations dues à l'apparition de mutations.



Le dispositif élimine bien totalement la sélection naturelle puisque même les cellules-mères mortes suite à des mutations létales sont conservées.

Détails de l'expérience MV Dans cette expérience, on se place toujours dans la *Mother Machine* et on observe par fluorescence l'apparition de mismatchs entre paires de nucléotides dus

à des erreurs de réplication de l'ADN (via l'insertion d'un rapporteur YFP-MutL). En prenant des mutants (*MutH*, *MutT*) dont le système de réparation est dysfonctionnel, on obtient alors que tout mismatch donne une mutation donc les observations correspondent bien à l'apparition de mutations.



À partir de ces expériences plusieurs questions sont posées dans l'article : Est-ce que les mutations sont ponctuelles (dynamique d'apparition poissonnienne) ou arrivent groupées ? Est-ce que les baisses soudaines du taux de croissance sont dues à l'arrivée de mutations délétères isolées ou bien à une accumulation de mutations en interaction ? Quelle est la dynamique d'apparition de mutations létales ? Et surtout pour le problème qui nous intéresse, peut-on inférer la DFE ?

1.3 Modélisation mathématique

Notations On se place dans le cadre de l'expérience μMA. Notons W_t le taux de croissance au temps $t \geqslant 0$ d'une lignée de cellule, et N_t le nombre de mutations qu'il y a eu sur cette lignée depuis le début de l'expérience. Notons

$$s_i = \frac{W_{t_{i-1}} - W_{t_i}}{W_{t_{i-1}}} \in]-\infty, 1]$$

l'effet relatif de la i^e mutation. On a en particulier :

$$\frac{W_t}{W_0} = \prod_{i=1}^{N_t} (1 - s_i) \tag{1}$$

Enfin, on note λ le taux de mutation, que l'on suppose constant.

Objectif Le but du projet est de répondre au problème suivant :

> **Énoncé du problème :** Estimer la loi des s_i sachant (1), sachant que l'on peut mesurer expérimentalement W_t et sachant que N_t suit une loi de Poisson de paramètre λt. Trouver une mesure pertinente pour exprimer l'erreur entre la loi estimée et la loi « réelle ».

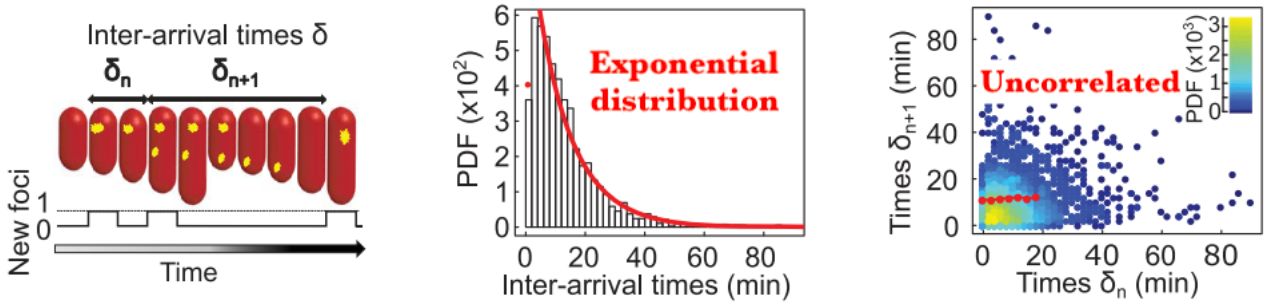
Hypothèses On fait les hypothèses suivantes sur les cellules :

1. Les effets des mutations s_i sont indépendants et identiquement distribués ;
2. Les mutations arrivent selon une dynamique poissonnienne ;
3. Le taux de mutation λ est constant. En particulier, il ne dépend pas de la taille de la cellule ;
4. Le taux de croissance change instantanément après la mutation (on ne prend pas en compte la division des cellules) ;
5. Les taux de croissance des cellules ne changent que à cause des mutations.

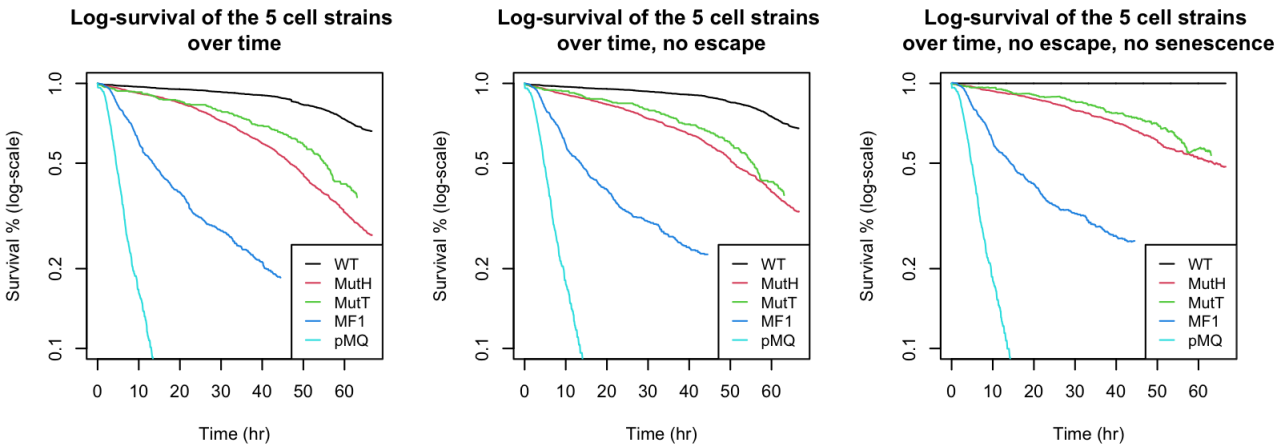
Remarque. L'hypothèse 1 permet de dire que la DFE, que l'on recherche, est bien définie. Dans la section suivante, on montre que l'hypothèse 2 est justifiée expérimentalement.

1.4 Dynamique poissonnienne des mutations

Afin de tester expérimentalement l'hypothèse d'apparition poissonnienne des mutations, les auteur.ice.s ont évalué les intervalles de temps δ entre 2 apparitions de mutations dans l'expérience de MV, ces intervalles définissant un processus de Poisson s'ils suivent une distribution exponentielle et si leurs longueurs sont indépendantes 2 à 2 (et en particulier d'un interval δ_n au suivant δ_{n+1}. Et effectivement ces deux critères sont vérifiés, ainsi que présenté dans la figure ci-dessous (en corrigeant la distribution attendue pour des observations discontinues toutes les 4 minutes). Nous avons répliqué ces résultats dans le notebook *I4_Replication_Dynamique-apparition-mutations*.



De même, on peut vérifier avec la seconde expérience (μMA) que les mutations létales suivent une dynamique semblable. Cela n'est pas évident a priori si on imagine que certaines mutations ont plus de chance d'être létales si la cellule a déjà accumulé précédemment des mutation délétères dans des voies partiellement redondants qui ne peuvent donc plus compenser une nouvelle perte de fonction. Mais ce n'est pas ce que l'on observe empiriquement, puisque les courbes de survie suivent une courbe exponentielle (droites dans les affichages log ci-dessous tirés du même notebook). On remarque aussi phénomène de sénescence des cellules commence à émerger, bien visible à partir de ∼40 heures chez la souche sauvage *WT* qui a peu de mutation, qui peut être corrigé en divisant par le taux de croissance des cellules WT (de taux d'apparition de mutations létales très faible). Seule la souche MF1 semble s'écarter légèrement d'une trajectoire exponentielle.



1.5 Tentative naïve pour déterminer la DFE

2 Simulations et tâtonnements

2.1 Simulations

2.2 Commentaire

3 Problème des moments

3.1 Détermination des moments

Dans cette section, nous allons donner une méthode permettant d'estimer les moments de la DFE, à partir des moments de la loi des W_t, $t \geqslant 0$.

Définissons

$$E_n(t) = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \ln\left(\mathbb{E}\left[W_t^k\right]\right)$$

Proposition 3.1.
Pour $t \geqslant 0$ et $n \in \mathbb{N}$:

$$E_n(t) = (\lambda \mathbb{E}\left[S^n\right]) t$$

Démonstration. Tout d'abord, remarquons que :

$$\begin{aligned}
\mathbb{E}\left[W_t^k\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{N_t}(1-s_i)^k \mid N_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(1-s)^k\right]^{N_t}\right] \\
&= e^{-\lambda t}\sum_{i=0}^{+\infty}\frac{\mathbb{E}\left[(1-s)^k\right]^i\lambda^i}{i!} \\
&= e^{-\lambda t}e^{\lambda t\mathbb{E}\left[(1-s)^k\right]}
\end{aligned}$$

Maintenant, nous pouvons calculer :

$$\begin{aligned}
E_n(t) &= \sum_{k=1}^{n}(-1)^k\binom{n}{k}\left(-\lambda t+\lambda\mathbb{E}\left[(1-s)^k\right]\right) \\
&= \lambda t\sum_{k=1}^{n}\left((-1)^k\binom{n}{k}\sum_{l=1}^{k}(-1)^l\binom{k}{l}\mathbb{E}\left[s^l\right]\right) \\
&= \lambda t\sum_{l=1}^{n}\mathbb{E}\left[s^l\right]\left((-1)^k\sum_{k=l}^{n}(-1)^{k+l}\binom{n}{k}\binom{k}{l}\right) \\
&= \lambda t\sum_{l=1}^{n}\mathbb{E}\left[s^l\right]\left((-1)^k\binom{n}{l}\sum_{k=l}^{n}(-1)^{k+l}\binom{n-l}{k-l}\right) \\
&= \lambda t\mathbb{E}\left[s^n\right]
\end{aligned}$$

□

En traçant les fonctions $t \mapsto E_n(t)$ (que l'on est capable de calculer à partir des observations de l'expérience μMA), on devrait obtenir des droites dont les pentes seront directement reliées aux moments de la DFE par un facteur λ. Dans l'annexe A.2, nous montrons que [2] a obtenu des droites et nous expliquons comment les obtenir ; nous estimons également, avec cette méthode, les moments de la DFE.

3.2 Estimation de la DFE

Nous avons vu qu'il était possible de déterminer les moments de la loi de S à partir de la donnée de la distribution des $W_t, t \geqslant 0$. Maintenant, nous allons tenter de trouver la loi de S à partir des moments de S : cela s'appelle le problème des moments.

Estimation de la fonction caractéristique À partir de tous les moments de S, on peut calculer la fonction caractéristique de la loi de S, grâce à l'expression suivante :

$$\varphi_S(\xi) := \mathbb{E}\left[e^{i\xi S}\right] = \mathbb{E}\left[\sum_{k=0}^{+\infty}\frac{(iS\xi)^k}{k!}\right] = \sum_{k=0}^{+\infty}\frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right]$$

Il est légal d'inverser la somme et l'espérance car on fait l'hypothèse que S est bornée. On a alors, pour tout $N \in \mathbb{N}$:

$$\begin{aligned}\varphi_S(\xi) := \mathbb{E}\left[e^{i\xi S}\right] &= \sum_{k=0}^{N} \frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right] + \sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right] \\ &= \sum_{k=0}^{N} \frac{(i\xi)^k}{k!}m_k + \sum_{k=0}^{N} \frac{(i\xi)^k}{k!}\left(\mathbb{E}\left[S^k\right] - m_k\right) + \sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!}\mathbb{E}\left[X^k\right]\end{aligned}$$

où $m_0, \ldots, m_n$ sont les moments que l'on a estimés par la méthode de la section précédente.

Notons

$$\hat{\varphi}_X(\xi) = \sum_{k=0}^{N} \frac{(i\xi)^k}{k!}m_k$$

qui est la meilleure estimation que l'on peut avoir de la fonction caractéristique à partir des N premiers moments.

Estimation de la DFE On remarque que, si la loi de S a une densité f :

$$\varphi_X(\xi) = 2\pi\mathcal{F}^{-1}f(\xi)$$

À partir de la fonction caractéristique φ_S, on peut donc trouver la densité f de S :

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \varphi_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi$$

Seulement, on ne connaît que les $N+1 > 0$ premiers moments, chacun avec une certaine erreur $(\varepsilon_k)_{0\leqslant k\leqslant N}$. Ainsi, on ne va calculer $\varphi_S(\xi)$ avec une erreur raisonnable que pour $\xi \leqslant A$, ce qui induira une erreur sur l'estimation de f.

Notons $\hat{f}$ la fonction que l'on calcule par cette méthode, qui est une estimation de f :

$$\hat{f}(x) = \frac{1}{2\pi}\int_{|\xi|\leqslant A} \hat{\varphi}_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi \tag{2}$$

3.3 Bornes sur l'erreur commise

On remarque que l'estimation (2) de la DFE f contient trois approximations :

1. l'erreur de régularisation qui consiste à ne pas considérer $\xi > A$;
2. l'erreur sur le calcul de $\hat{\varphi}_S$ qui consiste à ne considérer que les N premiers moments ;
3. l'erreur sur l'estimation des moments considérés.

Nous retrouverons ces trois erreurs dans la borne suivante sur l'erreur entre $\hat{f}$ et f :

Proposition 3.2.
Soient $A > 0$, $N \geqslant 1$, $k \geqslant 2$. On a alors :

$$(\forall x \in \mathbb{R}) \qquad \left|\hat{f}(x) - f(x)\right| \leqslant \alpha_1 + \alpha_2 + \alpha_3$$

avec :

$$\begin{aligned}
\alpha_1 &= \frac{\|f^{(k)}\|_1}{2\pi^2(k-1)A^{k-1}} \\
\alpha_2 &= \frac{A^{N+1}}{\pi(N+1)!}\mathbb{E}\left[S^N(e^{AS-1})\right] \\
\alpha_3 &= \frac{\|\varepsilon(N)\|_\infty(e^A-1)}{\pi}
\end{aligned}$$

où $\|\varepsilon(N)\|_\infty$ est l'erreur maximale commise sur le calcul des N premiers moments.

Démonstration. On peut décomposer $f(x)$ selon la régularisation des coefficients ξ et l'estimation de $\varphi_S(\xi)$:

$$\begin{aligned}
f(x) &= \frac{1}{2\pi}\int_{\mathbb{R}} \varphi_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi \\
&= \frac{1}{2\pi}\int_{|\xi|\leqslant A} \varphi_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi + \underbrace{\frac{1}{2\pi}\int_{|\xi|>A} \varphi_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi}_{a_1} \\
&= \frac{1}{2\pi}\int_{|\xi|\leqslant A} \left(\hat{\varphi}_S(\xi) + \sum_{k=0}^{N}\frac{(i\xi)^k}{k!}\left(\mathbb{E}\left[S^k\right]-m_k\right) + \sum_{k=N+1}^{+\infty}\frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right]\right)e^{-ix\xi}\,\mathrm{d}\xi + a_1 \\
&= \underbrace{\frac{1}{2\pi}\int_{|\xi|\leqslant A} \hat{\varphi}_S(\xi)e^{-ix\xi}\,\mathrm{d}\xi}_{\hat{f}(x)} + a_1 + \underbrace{\frac{1}{2\pi}\int_{|\xi|\leqslant A}\sum_{k=N+1}^{+\infty}\frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right]e^{-ix\xi}\,\mathrm{d}\xi}_{a_2} \\
&\quad + \underbrace{\frac{1}{2\pi}\int_{|\xi|\leqslant A}\sum_{k=0}^{N}\frac{(i\xi)^k}{k!}\left(\mathbb{E}\left[S^k\right]-m_k\right)e^{-ix\xi}\,\mathrm{d}\xi}_{a_3} \\
&= \hat{f}(x) + a_1 + a_2 + a_3
\end{aligned}$$

On obtient donc comme majoration de l'erreur d'approximation de f, pour $x \in \mathbb{R}$ et $\alpha_i = |a_i|$:

$$\left|f(x) - \hat{f}(x)\right| = |a_1 + a_2 + a_3| \leqslant |a_1| + |a_2| + |a_3| = \alpha_1 + \alpha_2 + \alpha_3$$

où

— α_1 est l'erreur que l'on commet en omettant de calculer $\varphi_S(\xi)$ pour $\xi > A$ (erreur de régularisation).

$$\alpha_1 = \frac{1}{2\pi}\left|\int_{|\xi|>A} \varphi_S(\xi) e^{-ix\xi}\,\mathrm{d}\xi\right|$$

Pour tout $k \geqslant 1$: $2\pi\left(\mathcal{F}^{-1}(f^{(k)})\right)(\xi) = (-i\xi)^k \varphi(\xi)$ donc :

$$\begin{aligned}
4\pi^2\alpha_1 &= \left|\int_{|\xi|>A} \frac{1}{(-i\xi)^k} 2\pi\mathcal{F}^{-1}(f^{(k)})(\xi) e^{-ix\xi}\,\mathrm{d}\xi\right| \\
&\leqslant \int_{|\xi|>A} \frac{1}{|\xi|^k} \underbrace{\left|\mathcal{F}^{-1}(f^{(k)})(\xi)\right|}_{\leqslant \|f^{(k)}\|_1} \mathrm{d}\xi \\
&\leqslant 2\|f^{(k)}\|_1 \int_{\xi>A} 1/(\xi^k)\,\mathrm{d}\xi
\end{aligned}$$

On a donc, pour tout $k \geqslant 1$:

$$\alpha_1 \leqslant \frac{\|f^{(k)}\|_1}{2\pi^2(k-1)A^{k-1}}$$

Pour que cette borne soit bonne, il faut d'une part faire certaines hypothèse sur f, d'autre part prendre A assez grand ;

— α_2 est l'erreur que l'on commet en omettant dans notre calcul les moments d'ordre plus grand que $N+1$.

$$\alpha_2 = \frac{1}{2\pi}\left|\int_{|\xi|\leqslant A} \sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right] e^{-ix\xi}\,\mathrm{d}\xi\right|$$

On a :

$$\begin{aligned}
2\pi\alpha_2 &\leqslant \int_{-A}^{A}\left|\sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!}\mathbb{E}\left[S^k\right]\right| e^{-i\xi x}\,\mathrm{d}\xi \\
&\leqslant \int_{-A}^{A} \sum_{k=N+1}^{+\infty} \frac{|\xi|^k}{k!}\mathbb{E}\left[S^k\right]\mathrm{d}\xi = 2\int_0^A \sum_{k=N+1}^{+\infty} \frac{\xi^k}{k!}\mathbb{E}\left[S^k\right]\mathrm{d}\xi \\
&\leqslant 2\sum_{k=N+1}^{+\infty} \frac{A^{k+1}}{(k+1)!}\mathbb{E}\left[S^k\right] = 2\mathbb{E}\left[\frac{1}{S}\sum_{k=N+1}^{+\infty}\frac{(AS)^{k+1}}{(k+1)!}\right]
\end{aligned}$$

D'après la formule de Taylor avec reste intégral :

$$\begin{aligned}
\sum_{k=N+2}^{+\infty}\frac{(AS)^k}{k!} &= e^{AS} - \sum_{k=0}^{N+1}\frac{(AS)^k}{k!} \\
&= \sum_{k=0}^{N+1}\frac{(AS)^k}{k!} + \int_0^{AS}\frac{(AS-t)^{N+1}e^t}{(N+1)!}\,\mathrm{d}t - \sum_{k=0}^{N+1}\frac{(AS)^k}{k!} \\
&= \int_0^{AS}\frac{(AS-t)^{N+1}e^t}{(N+1)!}\,\mathrm{d}t = \frac{(AS-t)^{N+1}(e^{AS}-1)}{(N+1)!}
\end{aligned}$$

d'où :

$$\alpha_2 \leqslant 2\mathbb{E}\left[\frac{(AS)^{N+1}(e^{AS}-1)}{2\pi S(N+1)!}\right] = \frac{A^{N+1}}{\pi(N+1)!}\mathbb{E}\left[S^N(e^{AS}-1)\right]$$

Pour que cette borne soit bonne, il faut prendre A assez petit et N assez grand ;

— α_3 est l'erreur que l'on commet qui provient des erreurs sur le calcul des moments.

$$\alpha_3 = \frac{1}{2\pi}\left|\int_{|\xi|\leqslant A}\sum_{k=0}^{N}\frac{(i\xi)^k}{k!}\left(\mathbb{E}\left[S^k\right]-m_k\right)e^{-ix\xi}\,\mathrm{d}\xi\right|$$

On a :

$$\begin{aligned}
2\pi\alpha_3 &\leqslant \int_{-A}^{A}\sum_{k=0}^{N}\left|\frac{(i\xi)^k}{k!}\left(\mathbb{E}\left[S^k\right]-m_k\right)e^{-i\xi x}\right|\mathrm{d}\xi \\
&\leqslant 2\|\varepsilon\|_\infty\int_0^A\sum_{k=0}^{N}\xi^k/(k!)\,\mathrm{d}\xi \quad \text{brutal} \\
&\leqslant 2\|\varepsilon\|_\infty\int_0^A e^{\xi}\,\mathrm{d}\xi
\end{aligned}$$

Avec :

$$\|\varepsilon\|_\infty = \max_{k=1,\dots,N}\left\{\mathbb{E}\left[S^k\right]-m_k\right\}$$

On a donc :

$$\alpha_3 \leqslant \frac{\|\varepsilon\|_\infty(e^A-1)}{\pi}$$

□

Optimisation du paramètre A On remarque que α_1 diminue quand A augmente, mais α_3 augmente quand A augmente. La proposition suivante donne la borne que l'on obtient lorsque l'on prend le meilleur compromis pour A.

Proposition 3.3.
Supposons que :

- *l'on soit capable de calculer un nombre arbitrairement grand de moments de f avec une erreur bornée par $\varepsilon > 0$;*
- *il existe $k \in \mathbb{N}$ tel que $d_k = \|f^{(k)}\|_1 < +\infty$;*

Alors on a, pour tout $x \in \mathbb{R}$:

$$\left|f(x)-\hat{f}(x)\right| \sim \left|\frac{1}{\pi\ln(\varepsilon)}\right| \quad \textit{quand } \varepsilon \to 0$$

Démonstration. On traite d'abord le cas particulier $k=2$, pour introduire les idées.

1. Cas particulier $k=2$: on a donc une erreur que l'on peut majorer par

$$\alpha_1+\alpha_2+\alpha_3 \leqslant \alpha_N(A) = \frac{d_2}{2\pi^2 A} + \frac{A^{N+1}}{(N+1)!\pi}\mathbb{E}\left[S^N(e^{AS}-1)\right] + \frac{\|\varepsilon\|_\infty(N)(e^A-1)}{\pi}$$

Supposons que l'on soit capable de prendre $N\to\infty$, avec une erreur sur les moments $\|\varepsilon\|_\infty(N)$ bornée par $\varepsilon>0$. De cette manière, on a $\alpha_2=0$. Posons $x=d_2/(2\pi^2)$ et $y=\varepsilon/\pi$. On a alors :

$$\alpha_N(A) \xrightarrow{N\to\infty} \alpha(A) = x/A + y(e^A-1)$$

donc $\alpha'(A) = -x/A^2 + ye^A$. On veut A tel que α soit minimum, c'est-à-dire $\alpha'(A)=0$ d'où :

$$A^2e^A = x/y$$

On obtient alors le paramètre A qui minimise la borne :

$$A = 2W\left(\frac{\sqrt{x/y}}{2}\right) = 2W\left(\sqrt{\frac{d_2}{8\pi\varepsilon}}\right)$$

avec W la fonction W de Lambert telle que $z=W(z)e^{W(z)}$.

2. Pour k général, on a les mêmes calculs, mais avec $x=\frac{d_k}{2\pi^2(k-1)}$, ce qui donne :

$$A = 2W\left(\sqrt{\frac{d_k}{8\pi\varepsilon(k-1)}}\right)$$

3. Regardons maintenant le comportement de A quand $\varepsilon\to 0$. Comme, quand $z\to\infty$:

$$W(z) = \ln z - \ln\ln z + o(1)$$

on a

$$A_\varepsilon = \ln\left(\frac{d_k}{8(k-1)\pi\varepsilon}\right) - \ln\ln\left(\frac{d_k}{8(k-1)\pi\varepsilon}\right) + o(1) = \ln(1/\varepsilon) - \ln\ln(1/\varepsilon) + o(1)$$

d'où :

$$\alpha(A_\varepsilon) = y(e^{A_\varepsilon}-1) + \frac{x}{{A_\varepsilon}^{k-1}} = \frac{\varepsilon}{\pi}e^{A_\varepsilon} + o(1) = \frac{1}{\pi\ln\left(\frac{1}{\varepsilon}\right)} + o(1)$$

d'où, quand $\varepsilon\to 0$:

$$\alpha(A_\varepsilon) \sim |1/(\pi\ln(\varepsilon))|$$

ce qui permet de conclure la proposition.

□

Commentaire Nous pouvons faire les remarques et les commentaires suivants :

1. On est capable de borner la norme infinie entre la DFE réelle f et la DFE estimée $\hat{f}$. Les bornes que l'on obtient ne sont pas optimales.
2. Sous des hypothèses très favorables, la norme infinie se comporte comme $\left|\frac{1}{\ln \varepsilon}\right|$, ce qui est une décroissance très lente de la borne (d'autant plus qu'il est largement exagéré de supposer que l'on sera capable de calculer les moments avec une grande précision) ;
3. Toutefois, la norme infinie n'est pas forcément la plus pertinente dans notre situation. On pourrait penser à d'autres méthode pour mesurer la distance entre ces deux distributions : par exemple, la norme L^2, la distance de Kolmogorov, ou la divergence de Kullback-Leibler.

L'objectif de la partie suivante est de présenter une nouvelle méthode pour estimer la DFE.

4 Étude d'une EDP

Dans cette partie, nous présentons une modélisation du problème par une EDP sur la densité de la loi de $\ln W_t$. D'après l'EDP, nous aurons une expression explicite de la loi de $\ln(1-S)$.

Nous commençons par introduire l'EDP (3), puis nous déduisons une expression explicite pour la loi de $\ln(1-S)$; ensuite, nous faisons le lien avec un problème de fragmentation ; ensuite, nous étudierons le comportement des solution de l'EDP en temps court et en temps long.

4.1 Nouveau point de vue sur le problème

Transformations initiales D'après (1), on a, tant que $W_t > 0$:

$$\ln W_t = \sum_{i=1}^{N_t} \ln(1-s_i)$$

On fait les deux hypothèses suivantes :

1. Pour tout $t > 0$, la loi de $\ln W_t$ peut s'écrire :

   $$m(t)\delta_{-\infty} + u(t,\cdot)$$

   où $m(t)$ représente la probabilité qu'une cellule soit morte au temps t et $u(t,\cdot) \in C^\infty(\mathbb{R})$ est la densité de la loi de $\ln W_t$ en omettant les cellules mortes : ainsi, $\int_{\mathbb{R}} u(t,\cdot) = 1 - m(t)$;
2. La loi de $\ln(1-S)$ peut s'écrire :

   $$\mu\delta_{-\infty} + f(\cdot)$$

   où μ est le taux de mutations létales et $f(\cdot) \in C^\infty(\mathbb{R})$ est la « densité » de la loi de $\ln(1-s)$ sans prendre en compte les mutations létales : ainsi, $\int f = 1 - \mu$.

Remarquons tout de suite que l'on peut déduire la DFE à partir de la donnée de f et du taux de mutations létales.

Introduction du modèle Soit λ le taux de mutation. Considérons :

$$\partial_t u(t,x) = \lambda\left(\int_{\mathbb{R}} f(x-y)u(t,y)\,\mathrm{d}y - \int_{\mathbb{R}} f(y)u(t,x)\,\mathrm{d}y\right) - \lambda\mu u(t,x)$$

Cette expression est plutôt naturelle ; elle peut se comprendre ainsi :

changement de densité de fitness entre t et $t+\mathrm{d}t$
= taux de mutations $\times$ (gens qui arrivent sur ma fitness $-$ gens qui partent de ma fitness)
$-$ gens qui meurent

où μ est le taux de mortalité (qui comprend la mortalité due à la sénescence et la mortalité due aux mutations létales).

Faisons quelques transformations pour simplifier l'expression. Comme $\int_{\mathbb{R}} f(y)\,\mathrm{d}y = 1-\mu$:

$$\partial_t u(t,x) = \lambda\left(\int_{\mathbb{R}} f(x-y)u(t,y)\,\mathrm{d}y - (1-\mu)u(t,x)\right) - \lambda\mu u(t,x)$$

soit :

$$\partial_t u(t,x) = \lambda(f * u(t))(x) - \lambda u(t,x)$$

que l'on notera, en notant $u_t(\cdot) = (u(t))(\cdot) = u(t,\cdot)$:

$$\partial_t u_t(x) = \lambda(f * u_t)(x) - \lambda u_t(x) \tag{3}$$

Vérification On veut vérifier que cette EDP est crédible. Pour cela, on peut par exemple vérifier que le nombre total de cellules $N(t)$ décroît comme $\exp(-\lambda\mu t)$:

$$\begin{aligned} N'(t)\partial_t\left(\int_{\mathbb{R}} u(t,x)\,\mathrm{d}x\right) = \int_{\mathbb{R}} \partial_t u &= \lambda\int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(y)(u(t,x-y)-u(t,x))\,\mathrm{d}y - \mu u(t,x)\right)\mathrm{d}x \\ &= \lambda\int_{\mathbb{R}} f(y)\left(\int_{\mathbb{R}} u(t,x-y)\,\mathrm{d}x - \int_{\mathbb{R}} u(t,x)\,\mathrm{d}x\right)\mathrm{d}y - \lambda\mu\int_{\mathbb{R}} u(t,x)\,\mathrm{d}x \\ &= -\lambda\mu\int_{\mathbb{R}} u(t,x)\,\mathrm{d}x = -\lambda\mu t \end{aligned}$$

ce qui donne, comme prévu, une décroissance malthusienne au taux $\lambda\mu$ (taux de mutation $\times$ proportion de mutations létales) du nombre total de cellules :

$$N(t) = e^{-\lambda\mu t}N(0)$$

Estimation de la DFE On peut mesurer u et on aimerait estimer f, en sachant que l'EDP ci-dessus est vérifiée : on connaît donc la solution mais on ne connaît pas l'EDP. On a la proposition suivante :

Proposition 4.1.
Supposons que $f \in .$ Soit $u \in \cdot$ une solution classique de (3)*. Pour tout $x \in \mathbb{R}$:*

$$f(x) = \mathcal{F}^{-1}\left(\xi \mapsto \frac{\partial_t\left(\mathcal{F}u_t(\xi)\right)}{\lambda \mathcal{F}u_t(\xi)} + 1\right) \tag{4}$$

Démonstration. En prenant la transformée de Fourier des deux côtés dans (3), on a :

$$\mathcal{F}\left(\partial_t u_t\right)(\xi) = \lambda \mathcal{F}f(\xi)\mathcal{F}u_t(\xi) - \lambda \mathcal{F}u_t(x)$$

soit :

$$\left(\partial_t \mathcal{F}u_t\right)(\xi) = \lambda \mathcal{F}f(\xi)\mathcal{F}u_t(\xi) - \lambda \mathcal{F}u_t(x)$$

et donc :

$$\begin{aligned}\mathcal{F}f(\xi) &= \frac{\partial_t \mathcal{F}u_t(\xi)}{\lambda \mathcal{F}u_t(\xi)} + 1 \\ &= \frac{1}{\lambda}\partial_t \ln\left(\mathcal{F}u_t(\xi)\right) + 1\end{aligned}$$

Ainsi :

$$f(x) = \mathcal{F}^{-1}\left(\frac{1}{\lambda}\partial_t \ln\left(\mathcal{F}u_t(\xi)\right) + 1\right)(x)$$

□

4.2 Détermination de la DFE à partir de (4)

L'expression (4) donne une forme explicite pour $f(x)$. Il faut, pour l'exploiter, être capable de calculer la valeur, pour chaque ξ, de

$$a_\xi = \frac{\partial_t \mathcal{F}u_t(\xi)}{\mathcal{F}u_t(\xi)}$$

Il est intéressant de remarquer que $a_\xi \in \mathbb{C}$ ne dépend pas du temps. Cela permet d'affirmer que :

$$\mathcal{F}u_t(\xi) = e^{a_\xi t}\mathcal{F}u_0(\xi)$$

et donc :

$$\left|\mathcal{F}u_t(\xi)\right| = e^{\Re(a_\xi)t}\left|\mathcal{F}u_0(\xi)\right| \qquad \text{et} \qquad \arg\left(\mathcal{F}u_t(\xi)\right) = \arg \mathcal{F}u_0(\xi) + t\Im(a_\xi)$$

On a donc deux expressions valables pour tout $t \in \mathbb{R}_+$:

$$\begin{aligned}\ln\left|\mathcal{F}u_t(\xi)\right| &= \ln\left|\mathcal{F}u_0(\xi)\right| + t \times \Re(a_\xi) \\ \arg\left(\mathcal{F}u_t(\xi)\right) &= \arg \mathcal{F}u_0(\xi) + t \times \Im(a_\xi)\end{aligned}$$

Traçons ces deux fonctions de t et vérifions qu'elles sont affines ; si c'est le cas, leurs pentes nous donneront la partie réelle et la partie imaginaire de a_ξ.

[ATTENDRE D'AVOIR DES RESULTATS COMPLETS POUR PRESENTER ICI DES DFE]

4.3 Lien avec un problème de fragmentation

Présentation du problème de fragmentation Nous présentons ici une transformation de (3) qui est étudiée dans [3], [4] dans le cadre d'un problème de fragmentation. Un problème de fragmentation vise à étudier la distribution des tailles de cellules lors de divisions cellulaires successives. L'intérêt du problème réside dans le fait que les cellules ne se divisent pas toujours en leur milieu, mais plutôt en un point aléatoire.

On note $k(x,y)$ la densité de probabilité pour une cellule de longeur x de se diviser en une cellule de longueur y et une cellule de longueur $x-y$. Le noyau de fragmentation k vérifie alors

$$k(x,y) = k(x,x-y) \qquad \int_0^x k(x,y)\,\mathrm{d}y = 1$$

Un problème intéressant, par exemple, est de savoir si k est bimodal ou non (*ie* : si les cellules ont tendance à se diviser en deux cellules de tailles à peu près égales, ou bien si une cellule fille a tendance à être beaucoup plus grosse que l'autre).

Ce modèle est très proche du nôtre : jusqu'à présent, les cellules changeaint de taux de croissance ; maintenant, les cellules changent de taille.

La seule véritable différence avec notre modèle est qu'une cellule se divise nécessairement en une cellule plus petite et une cellule plus grande ; cependant, cette différence n'est pas fondamentale car il est tout à fait légitime de supposer que l'immense majorité des mutations sont délétères, ce qui revient à négliger les mutations bénéfiques.

Mise en équation L'équation étudiée dans [3], [4] est :

$$\partial_t v(t,x) = \int_x^{+\infty} k\left(\frac{x}{y}\right) v(t,y)\,\mathrm{d}y - v(t,x) \tag{5}$$

Comparaison avec notre modèle Maintenant, nous allons montrer que (5) est une version multiplicative de (3). Considérons $v(t,\cdot)$ la densité de la loi de W_t et g la densité de $1-s$. Posons $x' = e^x$: on a alors $x'v(t,x') = u(t,x)$. Avec le changement de variable $y' = e^y$, on a :

$$\begin{aligned}
\partial_t(x'v(t,x')) &= \lambda \int_{\mathbb{R}} f(x-y)u(t,y)\,\mathrm{d}y - \lambda u(t,x) \\
&= \lambda \int_{\mathbb{R}_+} \frac{1}{y'} f(\ln(x') - \ln(y'))u(t,\ln(y'))\,\mathrm{d}y' - \lambda u(t,x) \\
&= \lambda \int_{\mathbb{R}_+} \frac{y'x'}{y'} g(x'/y')v\left(t,y'\right)\,\mathrm{d}y' - \lambda\left(x'v(t,x')\right)
\end{aligned}$$

donc

$$\partial_t v(t,x') = \lambda \int_{\mathbb{R}_+} g(y) v\left(t,\frac{x'}{y}\right)\,\mathrm{d}y - \lambda v(t,x')$$

ce qui est équivalent à (5).

Remarque. En réalité, (3) est une somme car on a transformé le produit (1) en somme en prenant le ln.

Résultat en temps long L'intérêt de comparer notre problème au problème de fragmentation étudié dans [3], [4] est que ces articles ont obtenu des résultats sur le comportement en temps long des solutions. Dans notre cas précis, il est montré que la distribution des taux de croissance (ou des tailles de cellules) converge vers un Dirac en 0.

Malheureusement, nous ne pouvons pas exploiter ce résultat en temps long car le temps de l'expérience n'est pas assez « long » pour que l'on puisse affirmer que l'on observe un comportement asymptotique. Nous avons pu, tout de même, faire tourner nos simulations pendant très longtemps et constater que la distribution convergeait effectivement vers un Dirac en 0.

Dans le cas où le taux de division *dépend de la taille de la cellule* (λ est de la forme x^γ, où x est la taille de la cellule et $0 < \gamma < 1$), [4] montre que la distribution converge vers une distribution particulière, qui n'est pas forcément un Dirac. Il y a donc une stationnarité qui apparaît. Ce résultat peut être intéressant dans le cas où nous aimerions faire évoluer notre modèle vers un modèle qui prendrait en compte le fait que les cellules qui croissent lentement ont un métabolisme plus lent et, par conséquent, subissent moins de mutations.

Références

[1] Trudy F. C. Mackay, Eric A. Stone, Julien F. Ayroles, *The genetics of quantitative traits : challenges and prospects*, Nature Reviews Genetics, 565–577, 2009

[2] Robert et al., *Mutation dynamics and fitness effects followed in single cells*, Science 359, 1283–1286, 16 March 2018

[3] Doumic, Escobedo, *Time asymptotics for a critical case in fragmentation and growth-fragmentation equations*, submitted 2015

[4] Beal et al., *The Division of Amyloid Fibrils : Systematic Comparison of Fibril Fragmentation Stability by Linking Theory with Experiments*, iScience, 25 September 2020

A Annexe : Présentation des résultats

Nous avons répliqué certains résultats dans des notebooks disponibles sur `https://github.com/Jeremy-Andreoletti/MSV_Project_DFE`.

A.1 Dynamique poissonienne des mutations

A.2 Calcul des premiers moments

Commentaires sur les figures La figure 1, tirée de [2], présente des graphes de $E_n(t)$. Dans 3.1, on a démontré que l'on s'attendait à obtenir des droites de pente $\lambda\mathbb{E}[S^n]$: les droites que l'on obtient sont très satisfaisantes. On peut effectuer à partir des ces droites des régressions linéaires pour estimer les pentes $\lambda\mathbb{E}[S^n]$ de ces droites. Les résultats obtenus dans [2] sont présentés dans le tableau 2 pour un nombre plus grand de souches et de moments. Enfin, le tableau 3 présente des estimations de la moyenne, de l'écart-type, du coefficient d'asymétrie et de la kurtosis pour les DFE de trois souches (*mutH*, *mutT*, **MF1**). Ces moments particuliers permettent d'obtenir des indices qualitatifs sur la forme des DFE : par exemple, comme le coefficient d'asymétrie et la kurtosis sont très élevés, on s'attend à ce que les DFE soient très piquées et très asymétriques, *ie* : très peu de mutations sont très délétères, la plupart des mutations sont quasiment neutres.

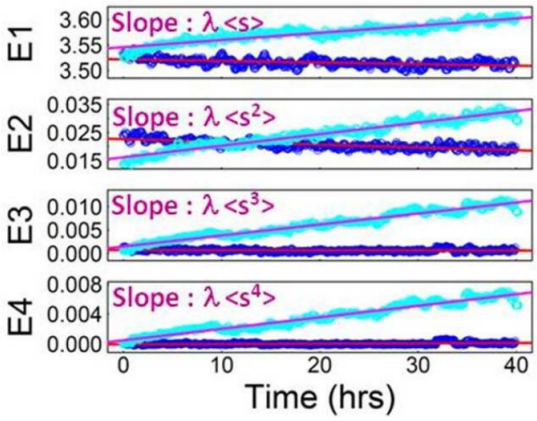


FIGURE 1 – Graphes de $t \mapsto E_n(t)$ pour $n = 1, 2, 3, 4$, obtenus dans [2], pour les souches WT (en bleu foncé) et *MutH* (en bleu clair). Les droites correspondent aux régressions linéaires effectuées.

Traitement initial des données Les données obtenues dans l'expérience μMA sont très bruitées. Nous indiquons ici les transformations que les auteurs de [**?**] ont effectuées sur ces données afin d'obtenir des résultats corrects :

— Sélection des cellules qui sont encore vivantes à la 44^e heure. En effet, les cellules qui croissent très lentement ou qui sont mortes induisent du bruit dans l'estimation de la DFE. Nous avons pu vérifier que le conditionnement effectué (« on suppose que la cellule

	mutH **Exp1**	*mutH* **Exp2**	*mutH* **Exp3**	*mutT*	**MF1**	**WT Exp1**	**WT Exp2**	**WT Exp3**
Slope of E_1**:** $\lambda < s > (\cdot 10^5)$	1.9	1.6	1.5	1.2	39	-0.3	0.1	0.3
Slope of E_2**:** $\lambda < s^2 > (\cdot 10^6)$	6.8	5.2	3.0	2.2	166	-1.5	-1.6	-0.1
Slope of E_3**:** $\lambda < s^3 > (\cdot 10^6)$	3.9	2.7	1.2	1.6	89	-0.1	-0.09	0.03
Slope of E_4**:** $\lambda < s^4 > (\cdot 10^7)$	25	16	7.3	9.4	560	0.1	-0.2	0.6
Slope of E_5**:** $\lambda < s^5 > (\cdot 10^7)$	18	11	4.8	5.5	390	-0.01	0.06	0.4
Slope of E_6**:** $\lambda < s^6 > (\cdot 10^7)$	13	7.5	3.3	3.5	290	-0.05	0.01	0.2
Slope of E_7**:** $\lambda < s^7 > (\cdot 10^7)$	9.7	5.5	2.3	2.3	220	-0.04	-0.007	0.1
Slope of E_8**:** $\lambda < s^8 > (\cdot 10^7)$	7.4	4.2	1.6	1.6	180	-0.03	0.002	0.07
Slope of E_9**:** $\lambda < s^9 > (\cdot 10^7)$	5.6	3.3	1.2	1.1	150	-0.02	0.001	0.05
Slope of E_{10}**:** $\lambda < s^{10} > (\cdot 10^7)$	4.3	2.7	0.8	0.8	130	-0.02	-0.002	0.03

FIGURE 2 – Pentes obtenues dans [2] pour les 10 premiers moments sur les souches *mutH* (trois expériences), *mutT*, **MF1**, **WT** (trois expériences).

	mutH **Exp1**	*mutH* **Exp2**	*mutH* **Exp3**	*mutT*	**MF1**
Mean (%)	0.35	0.30	0.28	0.22	0.35
CV	10.0	10.3	8.2	9.2	11
Skewness	16.0	16.6	17.3	35	14
Kurtosis	302	329	446	1040	220

FIGURE 3 – Estimations obtenues dans [2] de la moyenne, de l'écart-type, du coefficient d'asymétrie et de la kurtosis pour les DFE de trois souches (*mutH*, *mutT*, **MF1**)

ne meurt pas au cours de l'expérience ») n'induit pas de biais majeur dans l'estimation des moments ;

— Pour éviter les erreurs d'analyse d'image, les taux de croissances des lignées ayant un taux de croissance très faible (inférieur à 0.015) ont été vérifiés visuellement ;

— L'analyse d'image peut créer des valeurs aberrantes. Ces valeurs aberrantes ont été supprimées ainsi : si une valeur au temps t diffère d'au moins 30% de la médiane des valeurs prises avant le temps t, et diffère d'au moins 30% de la médiane des valeurs prises après le temps t, alors ont supprime cette valeur.

Prise en compte du bruit multiplicatif Les auteurs de [2] on montré que, en supposant que le bruit sur la mesure des taux de croissance est multiplicatif, alors la pente de $E_n(t)$ n'est pas modifiée par le bruit. Un bruit purement multiplicatif n'influence donc pas l'estimation des moments. Plus précisément, notons W'_{t_i} le taux de croissance mesuré au temps t_i, et W_{t_i} le taux de croissance réel au tems t_i. L'hypothèse du bruit multiplicatif suppose que $W'_{t_i} = W_{t_i}(1+\varepsilon_i)$ où les ε_i, $i \in \mathbb{N}$ sont des variables aléatoires indépendantes, identiquement distribuées, et indépendantes des W_t. On a :

$$\ln \mathbb{E}\left[W'_{t_i}\right] = \mathbb{E}[W_{t_i}] + \mathbb{E}[1+\varepsilon_i] = \mathbb{E}[W_{t_i}] + \mathbb{E}[1+\varepsilon]$$

ce qui fait que l'estimation $E'_n(t)$ que l'on fait de $E_n(t)$ diffère de $E_n(t)$ d'un terme indépendant de t. Ainsi, l'ordonnée à l'origine de la droite est modifiée mais pas sa pente.

B Annexe : Simulation