

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

Années 2020 – 2021

Résumé

Le but de ce projet est d'estimer les effets des mutations sur la valeur sélective d'une bactérie.

Dans un premier temps, nous présentons le contexte général dans lequel se place cette étude, et, en particulier, nous décrivons les expériences réalisées dans [2] dont sont issues les données sur lesquelles nous nous sommes basés. Nous introduisons le modèle mathématique qui permet leur analyse et l'interprétation de certains résultats de l'article initial que nous avons répliqués. Une procédure permettant de réaliser des simulations réalistes des expériences est également décrite.

Ensuite, nous proposons une première méthode pour trouver la distribution des effets des mutations sur la valeur sélective (appelée DFE pour *distribution of fitness effects*), et calculons des bornes sur l'erreur commise. Nous remarquons que la recherche de la DFE à partir des données des expériences réalisées dans [2] est un problème inverse sévèrement mal posé.

Enfin, nous abordons le problème par un autre point de vue, en considérant le processus comme la solution d'une EDP avec un terme intégral. Cette approche nous permet alors d'ouvrir quelques pistes de réflexion sur l'étude de cette EDP et sur l'aide qu'elle pourrait nous fournir dans l'estimation de la DFE.

Les notebooks qui complètent ce rapport (simulations, etc.) sont trouvables à l'adresse :
https://github.com/Jeremy-Andreoletti/MSV_Project_DFE.

Table des matières

1	Introduction	3
1.1	Contexte	3
1.2	Expériences	4
1.3	Modélisation mathématique	5
1.4	Dynamique poissonnienne des mutations	6
1.5	Tentative naïve pour déterminer la DFE	7
2	Simulations	9
3	Problème des moments	12
3.1	Détermination des moments	12
3.2	Estimation de la DFE	13
3.3	Bornes sur l'erreur commise	14
4	Étude d'une EDP	19
4.1	Nouveau point de vue sur le problème	19
4.2	Détermination de la DFE	22
4.3	Lien avec un problème de fragmentation	26
5	Conclusion et perspectives	27
A	Annexe : Présentation des résultats de [2]	30
B	Annexe : Vérification des résultats de 4.2	32
C	Annexe : Animations issues des simulations	33

1 Introduction

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 effet, et plus spécifiquement l'étude des liens entre variations génotypiques à ces loci et variations phénotypiques des traits associés. Bien comprendre ces mécanismes est essentiel non seulement en médecine humaine, afin de prédirer les risques de certaines maladies génétiques et mettre au point des traitements individuels, mais aussi en agriculture, afin de guider les programmes de sélection artificielle de plantes ou d'animaux domestiques. 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, d'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)**.

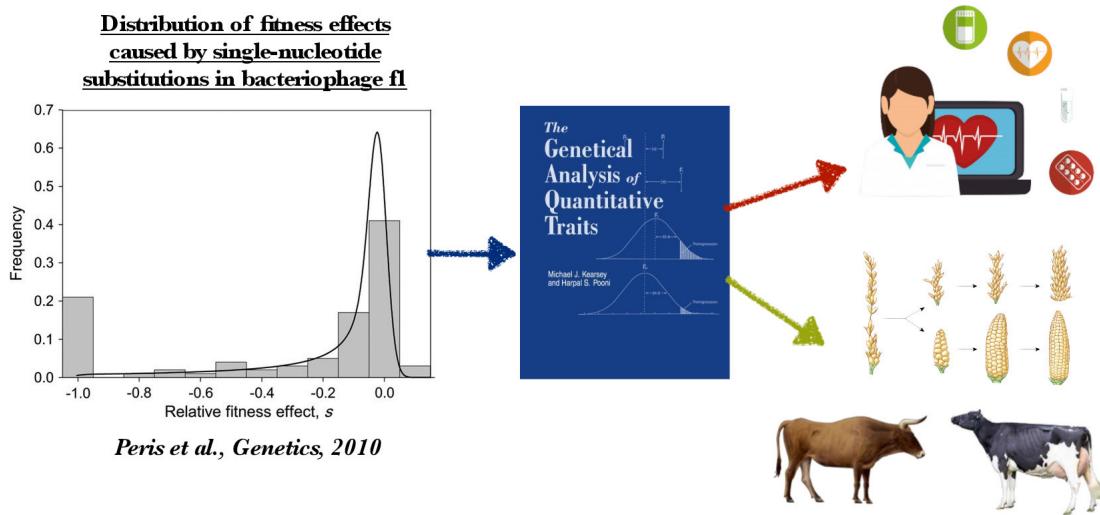


FIGURE 1 – L'étude de la DFE permet de l'analyse génétique des traits quantitatifs. Cela a des applications en médecine et en sélection artificielle.

Le point de départ de ce projet est un article de Lydia Robert *et al.* [2] dont l'objectif est précisément d'aider à comprendre l'impact des mutations sur la fitness. En particulier, les auteurs 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 présentés deux types d'expérience qui dépassent une partie de ces limitations, en se basant sur l'observation de bactéries piégées dans 1476 micro-canaux d'une puce micro-fluidique :

1. **microfluidic Mutation Accumulation (μ MA) experiment** : suivi du taux de croissance des bactéries sur de nombreuses générations en supprimant les effets de la sélection naturelle. Le taux de croissance des bactéries représentera la fitness des cellules ;
2. **Mutation Visualization (MV) experiment** : détection en temps réel des mutations apparaissant dans l'ADN bactérien. Le suivi est effectué sur une durée d'environ 60 heures.

Détails de l'expérience μ MA Dans cette expérience, les bactéries sont placées dans une *Mother Machine* dans laquelle 1476 cellules-mères se multiplient 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 2 tirée de l'article, on peut reconstruire l'évolution des taux de croissance sur plusieurs générations et essayer d'extraire la part de ces variations due à l'apparition de mutations.

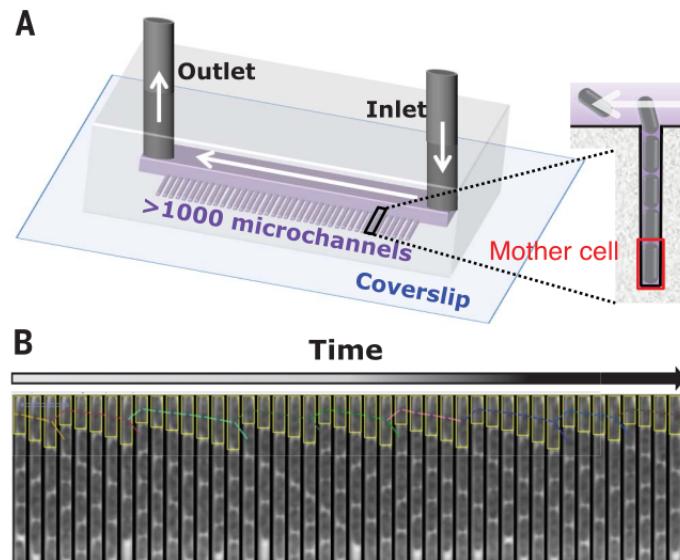


FIGURE 2 – (A) Vue d'ensemble de la *Mother Machine* utilisée pour les expériences MV et μ MA. (B) Évolution temporelle de l'état d'un unique canal. La cellule mère, qui est celle dont on mesure le taux de croissance dans l'expérience μ MA, est colorée en jaune.

Le dispositif élimine 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 (figure 3) l’apparition de mésappariements entre paires de nucléotides dues 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 s’assure que toute erreur d’appariement donne à terme une mutation. Les apparitions de taches fluorescentes correspondent donc bien à l’apparition d’une mutation.

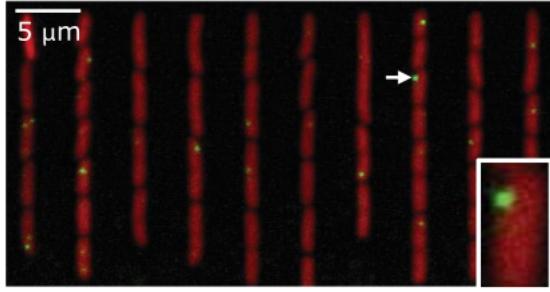


FIGURE 3 – Prise de vue reconstituée de l’expérience MV. Les mutations détectées correspondent aux points fluorescents jaunes.

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 des 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 μ MA d’accumulation de mutations. On note W_t le taux de croissance au temps $t \geq 0$ d’une lignée de cellules, et N_t le nombre de mutations qui ont eu lieu sur cette lignée depuis le début de l’expérience. On note

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

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 que l'on peut mesurer expérimentalement W_t , que N_t suit une loi de Poisson de paramètre λt et que l'on a l'expression (1). 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 du taux de croissance ;
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 changent uniquement à cause des mutations.

Remarque. L'hypothèse 1 permet de dire que la DFE, que l'on recherche, est bien définie. Dans la section 1.4, 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 auteurs ont évalué les intervalles de temps δ entre deux apparitions de mutations dans l'expérience de MV. Ces intervalles définissent un processus de Poisson si leurs longueurs suivent une distribution exponentielle et sont indépendantes deux à deux (et en particulier d'un intervalle δ_n au suivant δ_{n+1}). La figure 4 montre que ces deux critères sont vérifiés. Nous avons répliqué ces résultats dans le notebook *I4_Replication_Dynamique-apparition-mutations*¹.

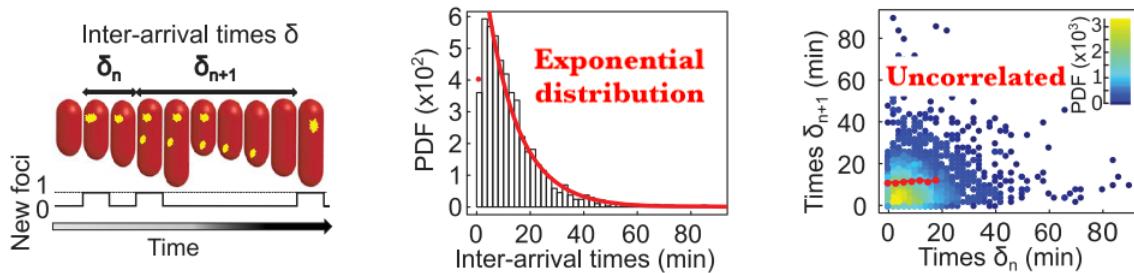


FIGURE 4 – Schémas tirés de [2] montrant que les mutations apparaissent selon un processus de Poisson. Les intervalles suivent une distribution exponentielle corrigée (car il faut corriger le fait que les prises de vue sont effectuées seulement toutes les 4 minutes) et les longueurs δ_n et δ_{n+1} sont indépendantes.

1. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/I4_Replication_Dynamique-apparition-mutations

De même, on peut vérifier avec la seconde expérience (μ MA) que les mutations létales en particulier suivent une dynamique semblable. Cela n'est pas évident a priori : on pourrait par exemple imaginer que certaines mutations ont plus de chances d'être létales quand la cellule a déjà accumulé des mutation délétères dans des voies partiellement redondantes qui ne peuvent donc plus compenser une nouvelle perte de fonction. Empiriquement, on observe que les courbes de survie (figure 5) suivent une courbe exponentielle. On remarque aussi qu'un phénomène de sénescence des cellules commence à émerger, bien visible à partir d'environ 40 heures chez la souche sauvage WT qui a peu de mutations. Le phénomène de sénescence peut être corrigé en divisant par le taux de croissance des cellules WT. Après ces corrections, seule la souche MF1 semble s'écarte légèrement d'une trajectoire exponentielle.

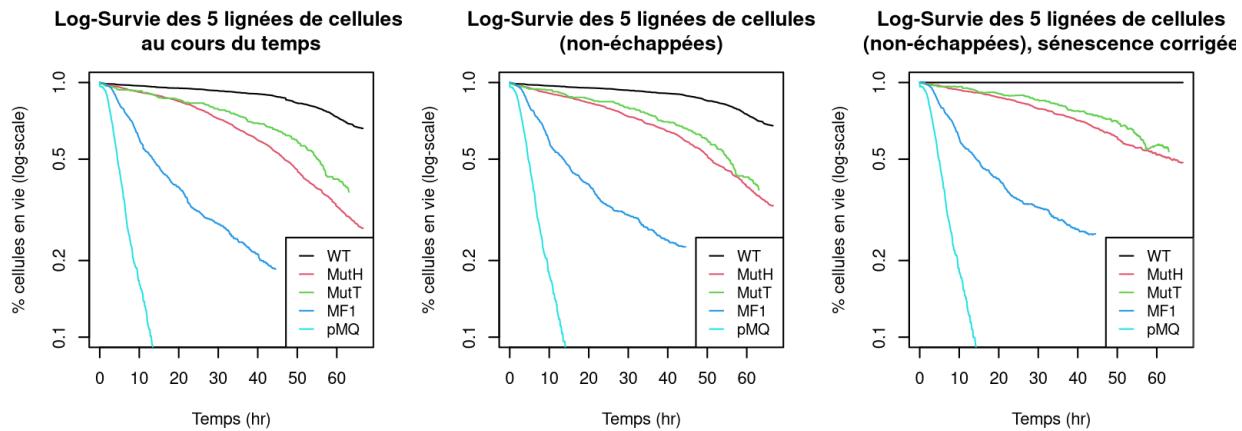


FIGURE 5 – Schéma tiré de notre réPLICATION DES RÉSULTATS, montrant les corrections effectuées pour obtenir une décroissance exponentielle du nombre de cellules au cours du temps. **Gauche :** sans correction ; **milieu :** suppression des canaux dont les cellules se sont échappées de leur canal, c'est-à-dire les cellules dont on a perdu la trace au cours de l'expérience ; **droite :** compensation des cellules qui meurent de vieillesse en divisant la proportion de cellules vivantes pour la souche par la proportion de cellules vivantes pour WT (on suppose la mort des cellules WT n'est causée que par la sénescence).

1.5 Tentative naïve pour déterminer la DFE

Nous avons essayé d'observer directement les variations des taux de croissance dues aux effets combinés des mutations et de la stochasticité des mesures, et si possible corriger cette dernière afin d'obtenir une estimation directe de la DFE. Ces tentatives sont présentées dans le notebook *I5_DFE-directe_Sauts-taux-de-croissance*².

2. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/I5_DFE-directe_Sauts-taux-de-croissance.ipynb

Nous nous basons sur le modèle de bruit multiplicatif dont la pertinence est justifiée dans l'article. Notons W_ε le taux de croissance mesuré et W le taux de croissance réel au temps t . On suppose que $W_\varepsilon = W(1 + \varepsilon)$ où les ε_i , $i \in \mathbb{N}$ sont des variables aléatoires indépendantes, identiquement distribuées, et indépendantes des W . On peut alors exprimer les effets relatifs bruités sur la fitness S_ε , avec W'_ε la mesure de fitness au temps t' suivant :

$$S_\varepsilon = \frac{W_\varepsilon - W'_\varepsilon}{W_\varepsilon} = \frac{W(1 + \varepsilon) - W'(1 + \varepsilon')}{W(1 + \varepsilon)} = 1 - \frac{W'}{W} \frac{1 + \varepsilon'}{1 + \varepsilon}$$

et donc, comme $|\varepsilon'|$ est très petit :

$$1 - S_\varepsilon \simeq \frac{W'}{W} (1 + \varepsilon')(1 - \varepsilon) \simeq \frac{W'}{W} (1 + \varepsilon' - \varepsilon)$$

Ce calcul montre que le bruit sur l'effet relatif sur la fitness calculé directement a une amplitude deux fois plus grande que le bruit sur la mesure. Cela est une première contre-indication à l'efficacité du calcul direct des effets des taux de croissance.

Nous avons tout de même essayé de regarder les distributions de variations relatives des taux de croissance pour les différentes souches de bactéries (figure 6). Ces distributions de variations relatives de taux de croissance ne sont pas normales, ni même symétriques. Les données utilisées pour ces tentatives sont celles de l'expérience MV.

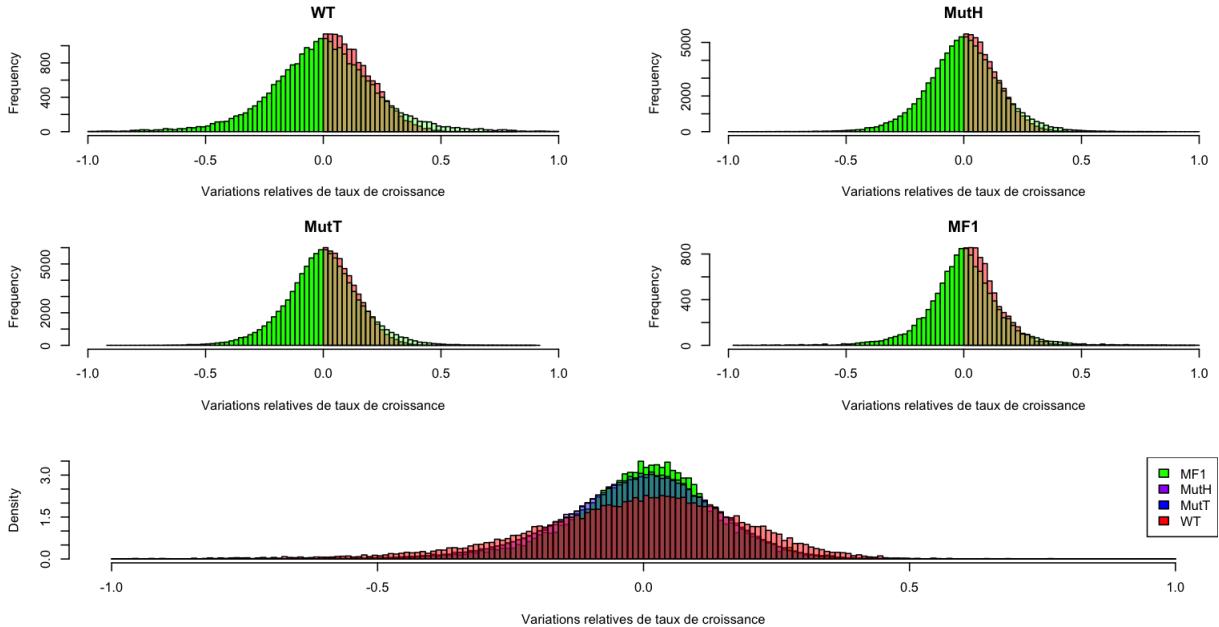


FIGURE 6 – Distribution des variations relatives des taux de croissance observées au cours de l'expérience, pour les différentes souches. Dans chaque cas la partie gauche de la distribution, en vert, a été reflétée de manière à illustrer l'asymétrie de ces distributions.

Plus le taux de mutation est faible ($\lambda_{WT} < \lambda_{MutH} \simeq \lambda_{MutT} < \lambda_{WT}$), plus la distribution des sauts relatifs de croissance est étalée. La part du bruit devient plus importante que celle du signal.

Malheureusement, les diverses tentatives subséquentes de retirer la distribution WT (quasi-mont uniquement du bruit) aux autres distributions, ou de faire des régressions pour des taux de mutations croissants se sont avérées être des échecs. Cela confirme la nécessité de faire appel à des outils mathématiques plus poussés afin de reconstruire la DFE.

2 Simulations

Afin de pouvoir tester les approches subséquentes, nous avons voulu réaliser des simulations aussi proches que possible de l'expérience réelle, en prenant en compte à la fois les mutations impactant les taux de croissance selon une DFE, la croissance des bactéries et le bruit des mesures. Les fonctions et animations associées sont présentées dans le notebook *II_Simulations*³.

Afin d'être les plus réalistes possibles, nos simulations suivent les étapes suivantes (figure 7) :

- Les taux de croissance sont initialisés avec ceux mesurés dans l'expérience (médiane des 10 premières valeurs) ;
- Une DFE est tirée dans un mélange de distributions au choix. Ici, nous avons pris la distribution Beta inférée dans l'article (de paramètres $\alpha = 0.0074$ et $\beta = 2.4$, voir figure 8) à laquelle on a rajouté une masse de Dirac en 1 (permettant d'incorporer 1% de mutations létales) ;
- Les effets s des mutations sur la fitness sont tirés dans cette DFE pour chacun des 1476 canaux de l'expérience et un grand nombre de mutations ;
- L'évolution du taux de croissance, mutation par mutation, est déduite des effets relatifs cumulés des mutations modifiant les taux de croissance initiaux ;
- Les durées entre 2 mutations sont tirées dans une loi exponentielle dont le paramètre est le taux de mutation ($\sim 0.32/\text{heure}$ en se basant sur le mutant *mutH*). Le taux de mutation peut être soit constant dans le temps et entre les cellules, soit être décroissant avec le taux de croissance (le cycle cellulaire et la réPLICATION étant ralentiS) ;
- À partir des temps d'apparition de chaque mutation on peut ne conserver que celles qui arrivent avant la fin de l'expérience (4000 minutes) et calculer l'évolution du taux de croissance, cette fois au cours du temps ;
- Enfin, le résultat est renvoyé en rajoutant (ou non) un bruit Gaussien multiplicatif à chaque mesure. Ce bruit est ensuite moyenné sur chaque génération de bactérie (comme indiqué dans l'article), en supposant que les bactéries se divisent dès qu'elles ont doublé de taille (même si un modèle de division incrémental serait plus réaliste).

Les images de la figure 9 sont extraites des animations présentées dans la figure 18 en annexe C et disponibles sur le notebook *II_Simulations*⁴. Elles représentent l'évolution au cours du temps

3. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/II_Simulations.ipynb

4. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/II_Simulations.ipynb

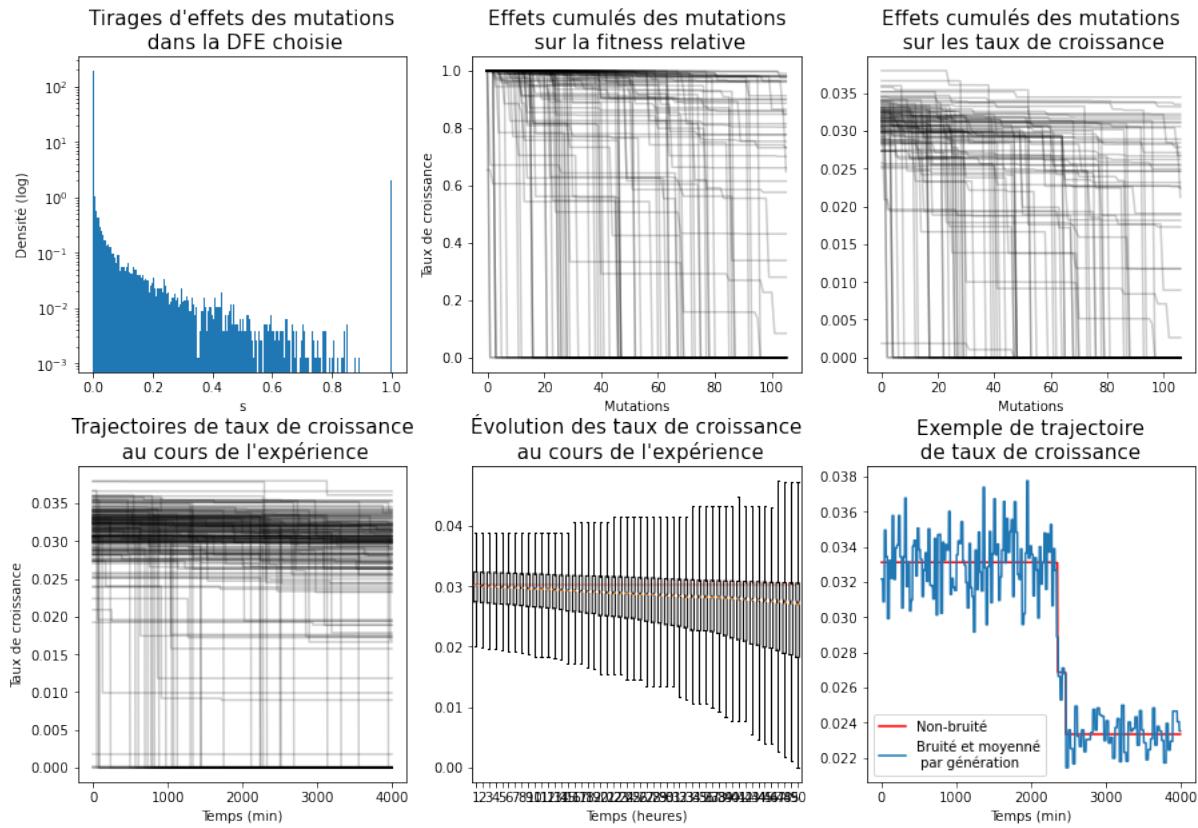


FIGURE 7 – Quelques étapes d’une simulation réaliste d’une expérience de croissance bactérienne, dont le taux de croissance est affecté par des mutations selon une DFE donnée et mesuré de manière bruitée

de la distribution des taux de croissance, en comparant les simulations aux données empiriques. On observe que l’évolution de la distribution simulée est au début en bonne adéquation avec celle observée. Toutefois, après 40 heures les deux commencent à diverger, probablement à cause du début de la sénescence des cellules, et de la DFE non adaptée (pas de mutations avantageuses, trop peu de mutations fortement délétères).

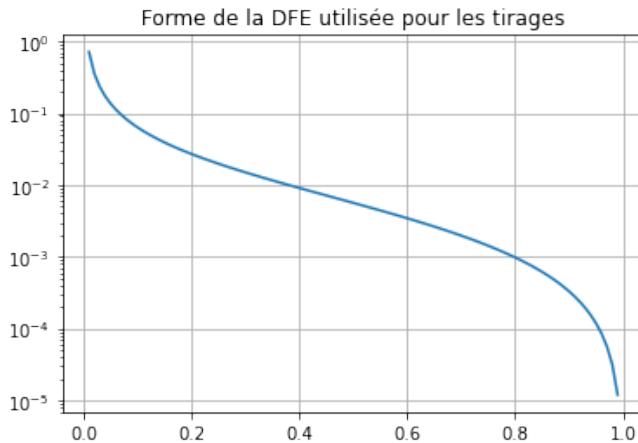


FIGURE 8 – Loi beta de paramètres $\alpha = 0.0074$ et $\beta = 2.4$.

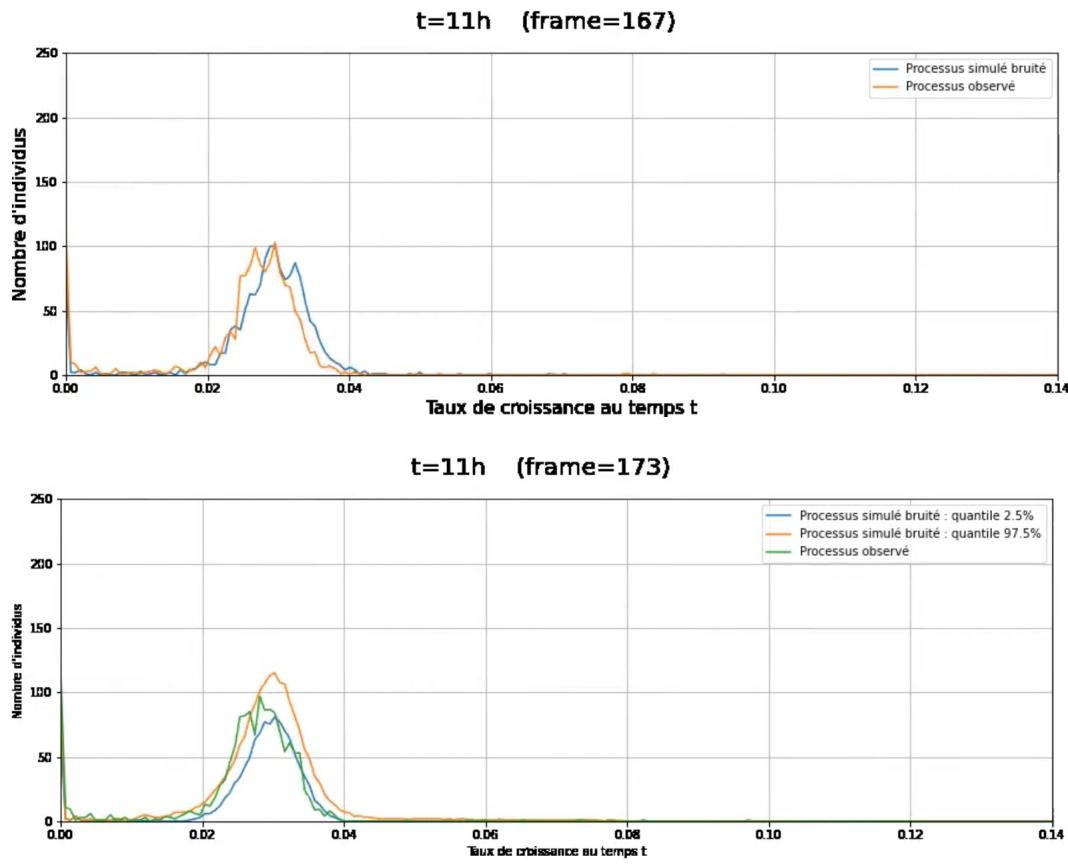


FIGURE 9 – Images tirées d’animations montrant l’évolution de la distribution des taux de croissance au cours du temps dans les simulations et l’expérience réelle. La seconde image affiche un intervalle des valeurs simulées sur 100 réplications.

3 Problème des moments

3.1 Détermination des moments

Dans cette section, nous allons présenter la méthode donnée dans [2] afin d'estimer les moments de la DFE à partir des moments de la loi des W_t , $t \geq 0$. Tous les résultats de la section 3.1 sont issus de [2].

Définissons

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

Notons S une variable aléatoire dont la loi est la DFE. Ainsi, S représente un effet de mutation typique.

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

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

Remarque. En traçant les fonctions $t \mapsto E_n(t)$ (qui peuvent être calculées à partir des observations de l'expérience μMA), on devrait obtenir des droites dont les pentes seront directement reliées aux moments de S (et donc de la DFE) par un facteur λ . Dans l'annexe A et le notebook *III1_Replication_Estimation-moments*⁵, nous montrons que [2] a bien obtenu des droites et nous expliquons les traitements permettant de reproduire ces résultats, et enfin d'estimer les moments de la DFE.

Démonstration. Tout d'abord, remarquons que, N_t suivant une loi de Poisson de taux λt :

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

5. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/III1_Replication_Estimation-moments.ipynb

Maintenant, nous pouvons calculer :

$$\begin{aligned}
E_n(t) &= \sum_{k=1}^n (-1)^k \binom{n}{k} (-\lambda t + \lambda t \mathbb{E}[(1-S)^k]) \\
&= \lambda t \sum_{k=1}^n (-1)^k \binom{n}{k} \left(\mathbb{E} \left[\sum_{l=0}^k \binom{k}{l} (-S)^l \right] - 1 \right) \\
&= \lambda t \sum_{k=1}^n \left((-1)^k \binom{n}{k} \sum_{l=1}^k (-1)^l \binom{k}{l} \mathbb{E}[S^l] \right) \\
&= \lambda t \sum_{l=1}^n \mathbb{E}[S^l] \left(\sum_{k=l}^n (-1)^{k+l} \binom{n}{k} \binom{k}{l} \right) \\
&= \lambda t \sum_{l=1}^n \mathbb{E}[S^l] \left(\binom{n}{l} \sum_{k=l}^n (-1)^{k+l} \binom{n-l}{k-l} \right) \\
&= \lambda t \sum_{l=1}^n \mathbb{E}[S^l] \left(\binom{n}{l} (1-1)^{n-l} \right) \\
&= \lambda t \mathbb{E}[S^n]
\end{aligned}$$

□

3.2 Estimation de la DFE

Nous avons vu dans 3.1 qu'il était possible de déterminer les moments de la loi de S (qui est la DFE) à partir de la donnée de la distribution des $W_t, t \geq 0$. 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 φ de la loi de S , grâce à l'expression suivante :

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

Il est légal d'inverser la somme et l'espérance si l'on fait l'hypothèse que $|S|$ est bornée. En particulier, cette hypothèse est vérifiée si toutes les mutations sont délétères : dans ce cas, en effet, $0 \leq S \leq 1$.

On a alors, pour tout $N \in \mathbb{N}$:

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

où m_0, \dots, 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.

Conventions utilisées pour la transformée de Fourier Nous définissons la transformée de Fourier $\mathcal{F} : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ et la transformée de Fourier inverse $\mathcal{F}^{-1} : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ par densité, à partir des expressions :

$$\mathcal{F}f(\xi) = \int_{\mathbb{R}} f(x)e^{-ix\xi} dx \quad \text{et} \quad \mathcal{F}^{-1}f(\xi) = \frac{1}{2\pi} \int_{\mathbb{R}} f(x)e^{ix\xi} dx$$

définies pour $f \in L^1(\mathbb{R})$. On a notamment :

$$\mathcal{F} \circ \mathcal{F}^{-1} = Id_{L^2(\mathbb{R})}$$

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

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

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

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

Seulement, on ne connaît que les $N+1 > 0$ premiers moments, chacun avec une certaine erreur $(\varepsilon_k)_{0 \leq k \leq N}$. Ainsi, on va calculer $\varphi_S(\xi)$ avec une erreur raisonnable que pour $\xi \leq A$, 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 :

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

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 $A > 0$, $N \geq 1$, $k \geq 2$. On a alors :

$$(\forall x \in \mathbb{R}) \quad |\hat{f}(x) - f(x)| \leq \alpha_1 + \alpha_2 + \alpha_3$$

avec :

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

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 $\varphi_S(\xi)$:

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

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

$$|f(x) - \hat{f}(x)| = |a_1 + a_2 + a_3| \leq |a_1| + |a_2| + |a_3|$$

où

- $|a_1|$ est l'erreur que l'on commet en omettant de calculer $\varphi_S(\xi)$ pour $\xi > A$ (erreur de régularisation).

$$|a_1| = \frac{1}{2\pi} \left| \int_{|\xi| > A} \varphi_S(\xi) e^{-ix\xi} d\xi \right|$$

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

$$\begin{aligned} |a_1| &= \left| \int_{|\xi|>A} \frac{1}{(i\xi)^k} \mathcal{F}^{-1}(f^{(k)})(\xi) e^{-ix\xi} d\xi \right| \\ &\leq \int_{|\xi|>A} \frac{1}{|\xi|^k} \underbrace{|\mathcal{F}^{-1}(f^{(k)})(\xi)|}_{\leq \|f^{(k)}\|_1} d\xi \\ &\leq 2\|f^{(k)}\|_1 \int_{\xi>A} 1/(\xi^k) d\xi \end{aligned}$$

On a donc, pour tout $k \geq 2$:

$$|a_1| \leq \alpha_1 = \frac{2\|f^{(k)}\|_1}{(k-1)A^{k-1}}$$

- $|a_2|$ est l'erreur que l'on commet en omettant dans notre calcul les moments d'ordre plus grand que $N+1$:

$$|a_2| = \frac{1}{2\pi} \left| \int_{|\xi|\leq A} \sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!} \mathbb{E}[S^k] e^{-ix\xi} d\xi \right|$$

On a :

$$\begin{aligned} 2\pi|a_2| &\leq \int_{-A}^A \left| \sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!} \mathbb{E}[S^k] \right| d\xi \\ &\leq \int_{-A}^A \sum_{k=N+1}^{+\infty} \frac{|\xi|^k}{k!} \mathbb{E}[S^k] d\xi = 2 \int_0^A \sum_{k=N+1}^{+\infty} \frac{\xi^k}{k!} \mathbb{E}[S^k] d\xi \\ &\leq 2 \sum_{k=N+1}^{+\infty} \frac{A^{k+1}}{(k+1)!} \mathbb{E}[S^k] = 2\mathbb{E}\left[\frac{1}{S} \sum_{k=N+1}^{+\infty} \frac{(AS)^{k+1}}{(k+1)!}\right] \end{aligned}$$

D'après la formule de Taylor avec reste intégral :

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

d'où :

$$|a_2| \leq \alpha_2 = 2\mathbb{E}\left[\frac{(AS)^{N+1}(e^{AS}-1)}{2\pi S(N+1)!}\right] = \frac{A^{N+1}}{\pi(N+1)!} \mathbb{E}[S^N(e^{AS}-1)]$$

— $|a_3|$ est l'erreur que l'on commet qui provient des erreurs sur le calcul des moments.

$$|a_3| = \frac{1}{2\pi} \left| \int_{|\xi| \leq A} \sum_{k=0}^N \frac{(i\xi)^k}{k!} (\mathbb{E}[S^k] - m_k) e^{-ix\xi} d\xi \right|$$

On a :

$$\begin{aligned} 2\pi |a_3| &\leq \int_{-A}^A \sum_{k=0}^N \left| \frac{(i\xi)^k}{k!} (\mathbb{E}[S^k] - m_k) e^{-i\xi x} \right| d\xi \\ &\leq 2\|\varepsilon\|_\infty \int_0^A \sum_{k=0}^N \xi^k / (k!) d\xi \\ &\leq 2\|\varepsilon\|_\infty \int_0^A e^\xi d\xi \end{aligned}$$

Avec :

$$\|\varepsilon\|_\infty = \max_{k=0,\dots,N} |\mathbb{E}[S^k] - m_k|$$

On a donc :

$$|a_3| \leq \alpha_3 = \frac{\|\varepsilon\|_\infty(e^A - 1)}{\pi}$$

□

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 (dans un cas très favorable).

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 $x \in \mathbb{R}$:

$$|f(x) - \hat{f}(x)| = O\left(\left|\frac{2d_k}{\ln^k(\varepsilon)}\right|\right) \quad \text{quand } \varepsilon \rightarrow 0$$

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

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

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

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

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 :

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

On obtient alors le paramètre A qui minimise la borne :

$$A = 2W\left(\frac{\sqrt{x/y}}{2}\right) = 2W\left(\sqrt{\frac{2\pi d_2}{4\varepsilon}}\right)$$

où W est la fonction W de Lambert, qui vérifie par définition : $z = W(z)e^{W(z)}$.

2. Pour k général, avec les mêmes calculs, on trouve que $\alpha'(A) = -\frac{2d_k}{A^k} + ye^A$, ce qui permet de déduire l'expression du A optimal :

$$A = kW\left(\frac{1}{k}\left(\frac{2\pi d_k}{\varepsilon}\right)^{1/k}\right)$$

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

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

on a, quand $\varepsilon \rightarrow 0$:

$$\begin{aligned} A_\varepsilon &= k \ln\left(\frac{1}{k}\left(\frac{2\pi d_k}{\varepsilon}\right)^{1/k}\right) - k \ln \ln\left(\frac{1}{k}\left(\frac{2\pi d_k}{\varepsilon}\right)^{1/k}\right) + o(1) \\ &= -k \ln(k) + \ln\left(\frac{2\pi d_k}{\varepsilon}\right) - k \ln\left(\ln\left(\frac{1}{k}\right) + \frac{1}{k} \ln\left(\frac{2\pi d_k}{\varepsilon}\right)\right) + o(1) \end{aligned}$$

d'où :

$$\begin{aligned} \alpha(A_\varepsilon) &= \frac{\varepsilon}{\pi}(e^{A_\varepsilon} - 1) + \frac{2d_k}{A_\varepsilon^{k-1}} \\ &= \frac{\varepsilon}{\pi} \times \frac{1}{k^k} \times \frac{2\pi d_k}{\varepsilon} \times \frac{e^{o(1)}}{\left(\ln\left(\frac{1}{k}\right) + \frac{1}{k} \ln\left(\frac{2\pi d_k}{\varepsilon}\right)\right)^k} - \frac{\varepsilon}{\pi} \\ &\quad + \frac{2d_k}{\left(k \ln\left(\frac{1}{k}\left(\frac{2\pi d_k}{\varepsilon}\right)^{1/k}\right) + o(\ln(x))\right)^{k-1}} \\ &= \frac{2d_k e^{o(1)}}{k^k \left(\ln(1/k) + \frac{1}{k} \ln\left(\frac{2\pi d_k}{\varepsilon}\right)\right)^k} - \frac{\varepsilon}{\pi} + \frac{2d_k}{\left(k \ln\left(\frac{1}{k}\left(\frac{2\pi d_k}{\varepsilon}\right)^{1/k}\right) + o(\ln(x))\right)^{k-1}} \end{aligned}$$

donc, comme le premier terme de la somme est dominant quand $\varepsilon \rightarrow 0$:

$$\alpha(A_\varepsilon) \sim \frac{2d_k}{\left(k \ln\left(\frac{1}{k}\right) + \ln(2\pi d_k) + |\ln(\varepsilon)|\right)^k}$$

d'où, quand $\varepsilon \rightarrow 0$:

$$\alpha(A_\varepsilon) \sim \left| \frac{2d_k}{\ln(\varepsilon)^k} \right|$$

ce qui permet de conclure la proposition.

□

Commentaires 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 borne que l'on obtient sur la norme infinie se comporte comme $\left|\frac{1}{\ln^k \varepsilon}\right|$, 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 tous 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 L^1 entre les fonctions de répartition, ou bien la divergence de Kullback-Leibler.

L'objectif de la partie 4 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 $\ln W_t$. Le but est de modéliser un processus limite en grande population. Toutefois, nous ne démontrons aucun résultat de convergence.

Nous commencerons par introduire (5), puis en déduirons une expression explicite pour la loi de $\ln(1 - S)$; nous finirons par faire le lien avec un problème de fragmentation.

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 :

$$m(t)\delta_{-\infty} + u(t, \cdot) \quad (3)$$

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

2. La loi de $\ln(1 - S)$ peut s'écrire :

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

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

Remarquons tout de suite que l'on peut déduire la DFE à partir de la donnée de f et de la fraction de mutations létale, mesurée directement dans l'expérience MV .

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

$$\partial_t u(t, x) = \lambda \left(\int_{\mathbb{R}} f(x - y) u(t, y) dy - \int_{\mathbb{R}} f(y) u(t, x) dy \right) - \lambda \mu u(t, x) \quad (4)$$

où μ est la proportion de mutations létale.

L'expression (4) introduit $u(t, x)$ qui peut, de manière assez naturelle, être interprétée comme la fonction telle que, pour chaque $t \geq 0$, $u(t, \cdot) \in C^\infty(\mathbb{R})$ soit la loi de $\ln W_t$ (sans compter les cellules mortes), comme on l'a définie dans (3). En effet, l'expression (4) peut se comprendre ainsi :

$$\begin{aligned} & \text{changement de densité de fitness entre } t \text{ et } t + dt \\ &= \text{taux de mutations} \times (\text{bactéries qui arrivent sur ma fitness} - \text{bactéries qui partent de ma fitness}) \\ &\quad - \text{bactéries qui meurent} \end{aligned}$$

Faisons quelques transformations pour simplifier (4). Comme $\int_{\mathbb{R}} f(y) dy = 1 - \mu$:

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

soit :

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

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

$$\partial_t u_t(x) = \lambda(f * u_t)(x) - \lambda u_t(x) \quad (5)$$

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 $\exp(-\lambda\mu t)$:

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

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 (5) est vérifiée : en quelque sorte, on connaît la solution mais on ne connaît pas le problème. On a la proposition suivante :

Proposition 4.1. Supposons que $f \in L^2(\mathbb{R})$. Soit $u \in C^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ une solution classique de (5). Supposons que, pour tout $t \geq 0$, $u_t \in \mathcal{S}(\mathbb{R})$ et que $\mathcal{F}u_t$ ne s'annule pas.

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

$$f(x) = \mathcal{F}^{-1} \left(\xi \mapsto \frac{\partial_t (\mathcal{F}u_t(\xi))}{\lambda \mathcal{F}u_t(\xi)} + 1 \right) \quad (6)$$

Démonstration. En prenant la transformée de Fourier des deux côtés dans (5), on a :

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

Comme $u_t \in \mathcal{S}(\mathbb{R})$, on a $\partial_t \mathcal{F}(u_t) = \mathcal{F}(\partial_t u_t)$, donc :

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

et, comme $\mathcal{F}u_t$ ne s'annule pas :

$$\mathcal{F}f(\xi) = \frac{\partial_t \mathcal{F}u_t(\xi)}{\lambda \mathcal{F}u_t(\xi)} + 1$$

Enfin, comme $f \in L^2(\mathbb{R})$, on obtient (6).

□

4.2 Détermination de la DFE

L'expression (6) 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 :

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

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

$$\ln |\mathcal{F}u_t(\xi)| = \ln |\mathcal{F}u_0(\xi)| + t \times \Re(a_\xi) \tag{7}$$

$$\arg(\mathcal{F}u_t(\xi)) = \arg \mathcal{F}u_0(\xi) + t \times \Im(a_\xi) \tag{8}$$

Vérification Dans l'annexe B et le notebook *IV2_DFE-EDP-Fourier_Verifications*⁶, nous avons tracé les fonctions de t (7) et (8), afin de vérifier qu'il s'agit de fonctions affines. Pour $\xi \leq 10$, on obtient des fonctions affines dont les pentes donnent respectivement la partie réelle et la partie imaginaire de a_ξ . Cependant, pour $\xi > 20$, on n'obtient plus de droites. Ainsi, pour l'estimation de la DFE, nous n'avons considéré que les valeurs de ξ inférieures à une certaine valeur ξ_{max} . Cela revient à tronquer la transformée de Fourier, et donc nous donne une DFE « trop régulière ». Cette régularisation est de toute manière nécessaire pour éviter tout phénomène de sur-apprentissage.

La figure 10 donne, pour chaque valeur de ξ comprise entre 0 et 20, le coefficient de détermination R^2 , qui mesure la qualité de la régression linéaire effectuée. Ainsi, quand R^2 est proche de 1, la régression linéaire est de bonne qualité ; quand R^2 est proche de 0, la régression linéaire est de mauvaise qualité. Pour des valeurs de ξ suffisamment petites, on obtient de très bonnes régressions linéaires, et donc, en principe, de très bonnes estimations de a_ξ et donc de $\mathcal{F}f(\xi) = a_\xi/\lambda + 1$. Cette méthode peut nous permettre d'obtenir la valeur de ξ maximale que l'on peut considérer. Ajouter du bruit à la mesure des taux de croissance oblige à réduire la valeur de ce ξ_{max} . La figure 11 montre une procédure qui permet de choisir manuellement de bons paramètres.

⁶. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/IV2_DFE-EDP-Fourier_Verifications.ipynb

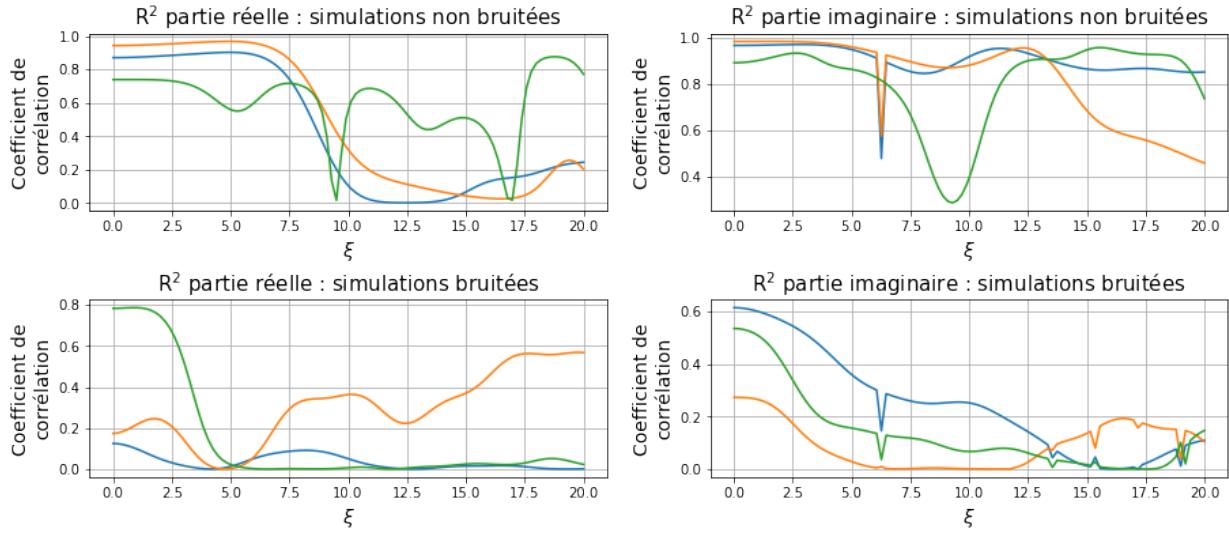


FIGURE 10 – Comparaison, selon ξ , des qualité d'ajustement d'un modèle affine pour les parties réelles (gauche) ou imaginaires (droite) de a_ξ , sur des simulations bruitées ou non. Chaque couleur correspond à une simulation indépendante de l'expérience.

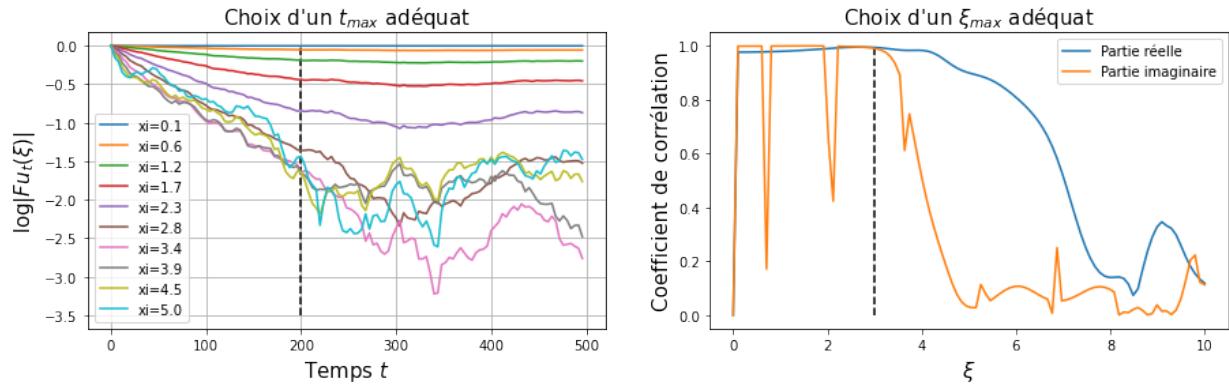


FIGURE 11 – Étapes manuelles de la procédure d'estimation de la DFE. À gauche, choix d'un temps maximal à partir duquel les fonctions ne semblent plus affines. Ensuite, à droite, choix d'un ξ maximal à partir duquel la qualité d'ajustement d'un modèle affine (coefficient de régression R^2) chute.

Estimation de la DFE Nous avons tenté d’appliquer notre méthode avec des données simulées non bruitées. L’implémentation est réalisée dans le notebook *IV2_DFE-EDP-Fourier_Estimation-complete*⁷. Nous constatons que les résultats ne sont pas satisfaisants : la figure 12 compare, pour différents paramètres (α, β) , les DFE calculées et la DFE utilisée (qui correspond à une loi Beta de paramètres (α, β)). Outre les erreurs de programmation, nous avons essentiellement deux explications pour expliquer la grande différence :

- Les DFE considérées sont discontinues en 0 et donc leur transformée de Fourier décroît lentement vers 0 quand $\xi \rightarrow \infty$: voir la figure 13. Comme nous sommes contraints de rogner la transformée de Fourier, nous perdons beaucoup d’information ;
- Nous calculons f , qui est la densité de la loi de $\ln(1 - S)$; pour afficher la densité de la loi de S , nous affichons

$$g(x) = \frac{f(\ln(1-x))}{1-x} \quad x > 0$$

ce qui fait que, pour x proche de 1, l’erreur est très grande : cela peut expliquer pourquoi l’on obtient systématiquement un pic en 1.

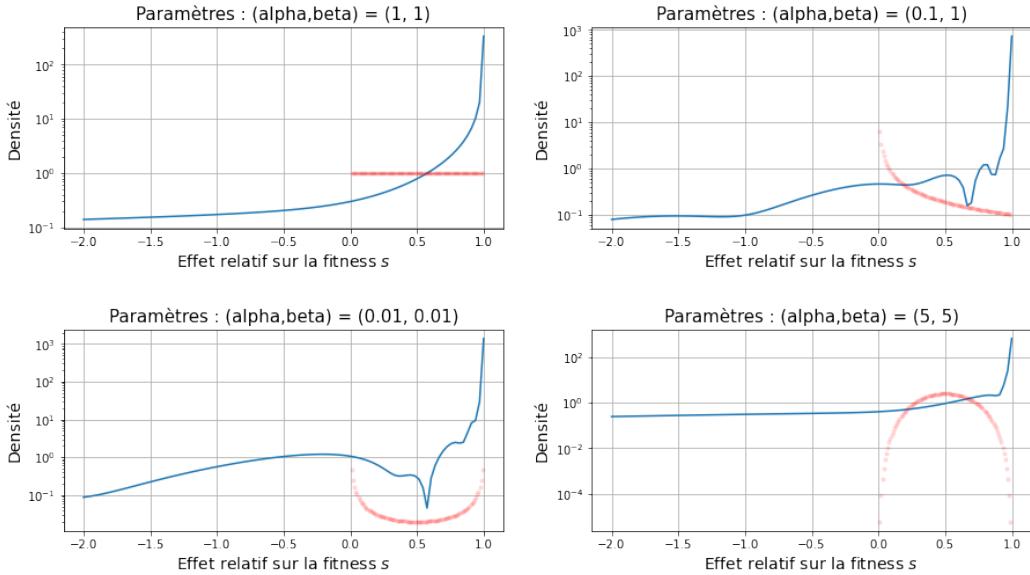


FIGURE 12 – Comparaison entre les DFE inférées (en bleu) et les lois Beta choisies pour effectuer les simulations (en rouge). Les DFE inférées sont affichées sur une fenêtre plus large, car l’on s’attendrait à ce qu’elles soient nulles en dehors des supports des DFE dont elles sont issues.

Nous avons remarqué, au cours de nos tests, que plus le nombre de cellules augmentait, plus nous pouvions considérer des ξ grands. Par exemple, pour 60 000 cellules, on pouvait prendre jusqu’à $\xi \simeq 20$. Cela est à mettre en relation avec les résultats de la figure 10, qui ont été réalisés avec des expériences de 1000 cellules.

7. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/IV2_DFE-EDP-Fourier_Estimation-complete.ipynb

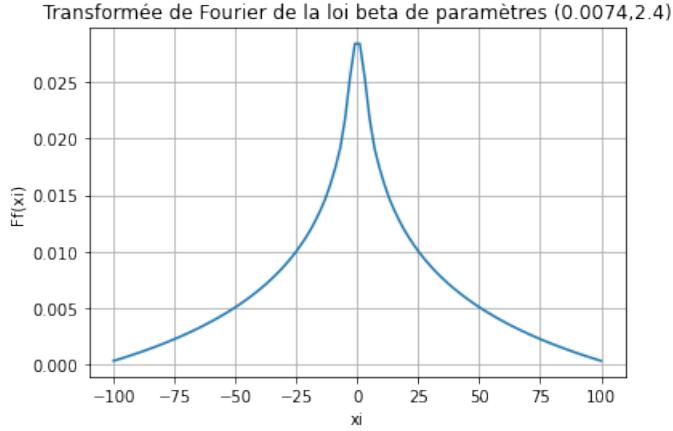


FIGURE 13 – La transformée de Fourier de la DFE inférée dans [2] décroît assez lentement vers 0 quand $\xi \rightarrow +\infty$: rogner pour $|\xi| \geq 10$ implique de perdre beaucoup d'information.

4.3 Lien avec un problème de fragmentation

Présentation du problème de fragmentation Nous présentons ici une transformation de (5) 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, modélisées comme des segments, 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 longueur x de se diviser en une cellule de longueur y et une cellule de longueur $x - y$. Le noyau de fragmentation k vérifie alors

$$k(x, y) = k(x, x - y) \quad \int_0^x k(x, y) dy = 1$$

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 changeaient 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 : en effet, d'après [2], il est tout à fait légitime de supposer que l'immense majorité des mutations sont délétères, ce qui revient à négliger les augmentations de taux de croissance.

Mise en équation Un cas particulier de l'équation étudiée dans [3], [4] est :

$$\partial_t v(t, x) = \int_x^{+\infty} k\left(\frac{x}{y}\right) v(t, y) dy - v(t, x) \tag{9}$$

Comparaison avec notre modèle Maintenant, nous allons montrer que (9) est une version multiplicative de (5). 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 :

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

donc

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

ce qui est équivalent à (9) avec $\lambda = 1$.

Remarque. Finalement, on trouve que (5) est une somme car on a transformé le produit (1) en somme en prenant le logarithme.

Résultats 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 notre simulation 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 aimerais 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.

5 Conclusion et perspectives

Plusieurs approches ont été testées au cours de ce projet afin de résoudre le problème inverse sévèrement mal posé que représente l'estimation de la distribution des effets des mutations sur la valeur sélective à partir de données bruitées de croissance bactérienne. Les premières analyses les plus naïves ayant confirmé leur insuffisance, nous avons exploré des méthodes analytiques plus poussées, sous l'angle du problème des moments ou d'une équation aux dérivées partielles retranscrivant la dynamique du système. Dans les deux cas, des résultats théoriques ont pu

être obtenus, mais en pratique les méthodes développées ne sont pas encore à même d'extraire suffisamment d'information de ces données expérimentales.

Cependant, les outils que nous avons proposés nous semblent toujours pertinents si des données plus nombreuses ou moins bruitées sont un jour acquises. Nous avons ainsi tenu à les présenter sous la forme de notebooks⁸ soigneusement annotés de manière à faciliter la reproductibilité de nos résultats et leur réutilisation.

Enfin, de nombreuses améliorations de nos méthodes et implémentations sont envisageables, par exemple, en comblant les lacunes de l'approche par EDP grâce à l'estimation des moments, en explorant plus en détail les similitudes avec les problèmes de fragmentation, ou bien en prenant en compte les propriétés de la DFE qui ont été ignorées jusqu'à présent (positivité, intégrale égale à 1). À cet égard, la possibilité de réaliser très efficacement des simulations réalistes, avec des DFE arbitraires et incorporant le bruit des mesures fidèlement à l'expérience, permettrait de tester la fiabilité de potentielles nouvelles approches. Avec, à terme, l'espoir de pouvoir obtenir une estimation robuste et entièrement non-paramétrique de la DFE, ouvrant la voie à une meilleure compréhension des conséquences phénotypiques des mutations.

Remerciements Nous tenons à remercier chaleureusement nos deux encadrantes : Marie Doumic, qui nous a accompagnés avec soin et bienveillance dans notre travail, et Lydia Robert, qui a répondu à toutes nos interrogations sur les expériences.

8. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE

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 de [2]

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

Commentaires sur les figures La figure 14, 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 15 pour un nombre plus grand de souches et de moments. Enfin, le tableau 16 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 de chaque souche : 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.

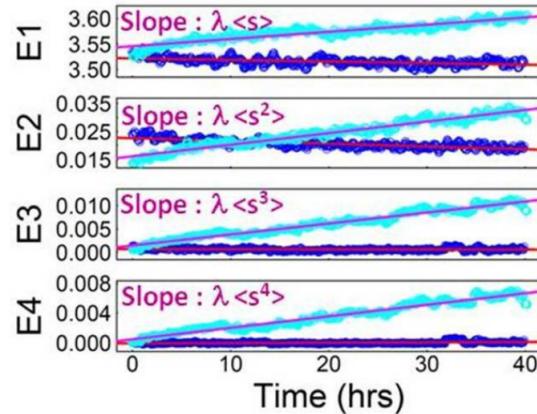


FIGURE 14 – Graphes de $t \mapsto E_n(t)$ pour $n = 1, 2, 3, 4$, obtenus dans [2], pour les souches WT (en bleu foncé) et *MutH* (en bleu clair). Les droites correspondent aux régressions linéaires effectuées.

Traitements initial des données Les données obtenues dans l'expérience μ MA sont très bruitées. Nous indiquons ici les transformations effectuées sur ces données afin d'obtenir des résultats corrects :

- Sélection des cellules qui sont encore vivantes à la 44^e heure. Cela se justifie par le fait que 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 ne meurt pas au cours de l'expérience ») n'induit pas de biais majeur dans l'estimation des moments ;

	<i>mutH</i> Exp1	<i>mutH</i> Exp2	<i>mutH</i> Exp3	<i>mutT</i>	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 15 – Pentes obtenues dans [2] pour les 10 premiers moments sur les souches *mutH* (trois expériences), *mutT*, **MF1**, **WT** (trois expériences).

	<i>mutH</i> Exp1	<i>mutH</i> Exp2	<i>mutH</i> Exp3	<i>mutT</i>	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 16 – 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**)

- 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 par les auteurs ;
- 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% simultanément de la médiane des valeurs prises avant le temps t , et de la médiane des celles prises après le temps t , alors on supprime cette valeur.

Prise en compte du bruit multiplicatif Les auteurs de [2] ont montré que, en supposant que le bruit sur la mesure des taux de croissance est multiplicatif, 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 temps 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 :

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

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 : Vérification des résultats de 4.2

La figure 17 montre les graphes des fonctions suivantes, trouvées dans 4.2, pour différentes valeurs de ξ :

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

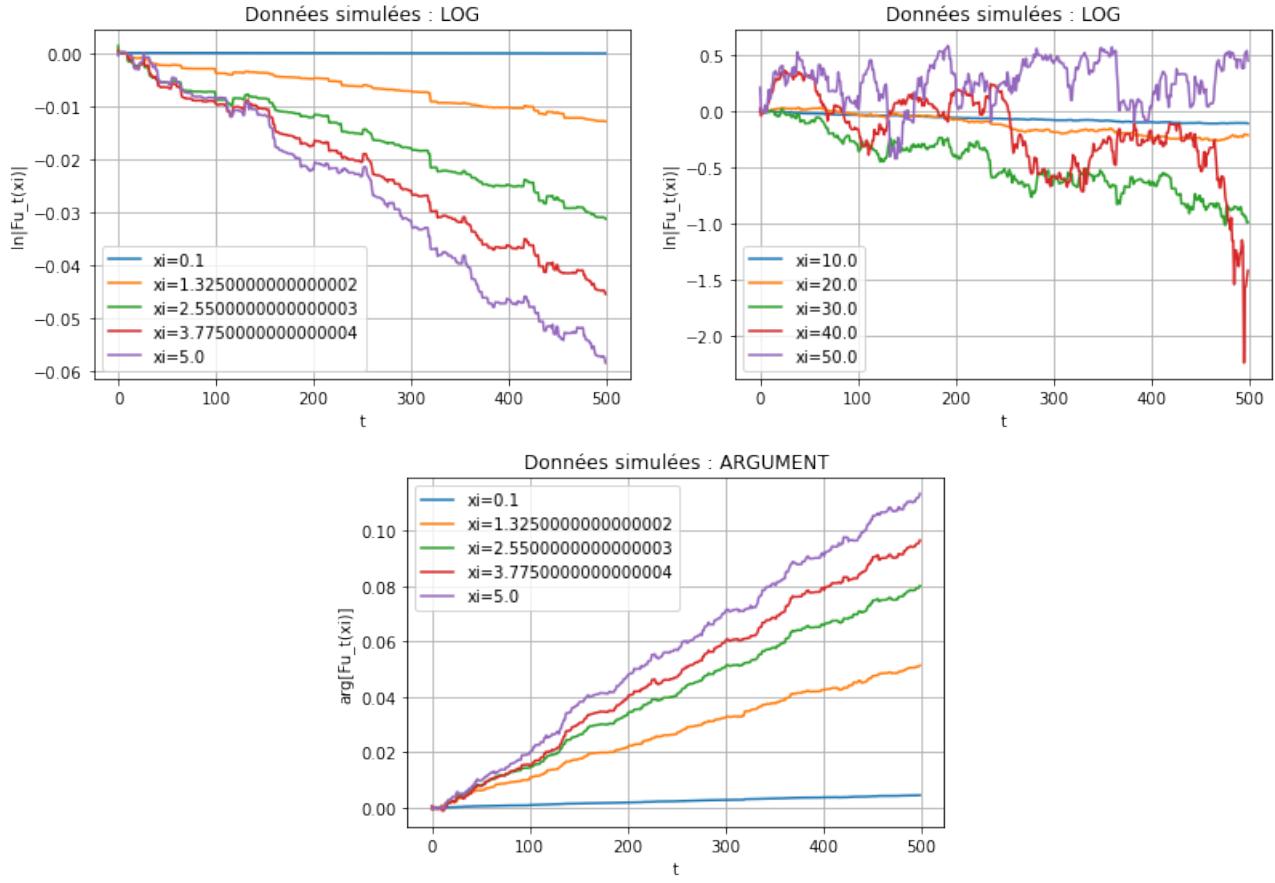


FIGURE 17 – Premier graphique : droites dont les pentes doivent donner a_ξ (petites valeurs de ξ) ; deuxième graphique : pour de grandes valeurs de ξ , on n'obtient pas de droites : cela est peut-être dû à l'overfitting ; troisième graphique : droites dont les pentes doivent donner a_ξ (petites valeurs de ξ).

C Annexe : Animations issues des simulations

La figure 18 montre des animations tirées des simulations présentées dans la partie 2. Les animations ne fonctionnent pas sur tous les lecteurs de pdf; elles restent disponibles sur le notebook *II_Simulations*⁹.



FIGURE 18 – Animations montrant l'évolution de la distribution des taux de croissance au cours du temps dans les simulations et l'expérience réelle. La seconde animation affiche un intervalle des valeurs simulées sur 100 réplications.

9. https://github.com/Jeremy-Andreoletti/MSV_Project_DFE/blob/master/II_Simulations.ipynb