Notes sur le projet

Jérémy Andréoletti et Nathanaël Boutillon

15 novembre 2020

1 Résumé de l'article

« Mutation dynamics and fitness effects followed in single cells », Lydia Robert *et al.*, Science

1.1 Texte principal

Contexte de la recherche On voudrait réussir à comprendre l'impact des mutations sur la fitness. Mais on n'a jamais pu suivre directement la dynamique d'apparition des mutations. Les expériences précédentes d'accumulation de mutations, qui moyennent les résultats sur plusieurs générations de colonies bactériennes, sont biaisée par l'effet de la sélection naturelle sur les mutations trop délétères/létales et par le nombre limité de lignées pouvant être suivies.

Questions à résoudre

- Question 1 : Peut-on retrouver la distribution complète des effets des mutations sur la valeur sélective ? (DFE = Distribution of Fitness Effects) ;
- Question 2 : Est-ce que les mutations sont ponctuelles (poissonniennes) ou arrivent groupées ?
- Question 3 : Est-ce que les baisses soudaines du taux de croissance sont dues à l'occurrence de mutations délétères isolées ou bien à une accumulation de mutations en interaction ?
- Question 4 : Quelle est la dynamique d'apparition de mutations létales ?

Méthodes développées

- **MV = Mutation Visualization experiment** : Détection en temps réel des mismatchs dus à des erreurs de réplication de l'ADN (YFP-MutL), sur des mutants (*MutH*, *MutT*) dont le système de réparation est dysfonctionnel, de manière à ce que tout mismatch donne une mutation ;
- **μMA = microfluidic Mutation Accumulation experiment** : Suivi de la fitness (taux de croissance en longueur) de bactéries piégées dans 1000 micro-canaux d'une puce microfluidique. Dispositif éliminant totalement la Sélection Naturelle puisque même les *cellules mères* mortes sont conservées.

Résultats principaux :

- Réponse 1 : Caractérisation non-paramétrique des **moments de la DFE** ;
- Réponse 2 : Distributions exponentielles et indépendantes des temps entre 2 nouvelles mutations. Le comportement est donc **poissonnien** ;
- Réponse 3 : Les mutations fortement délétères (baisse du taux de croissance supérieure à 30%) ont également une dynamique poissonnienne, avec un taux beaucoup plus faible : 0.3% de l'ensemble des mutations ;
- Réponse 4 : Décroissance exponentielle en temps du taux de survie (avant sénescence). La mort des cellules est donc principalement causée par des mutations létales uniques, de dynamique poissonnienne. Elles représentent 1% des mutations. Ainsi, il y a plus de mutations létales que de mutations fortement délétères : c'est qu'il est plus facile de tuer que de fragiliser gravement ;
- Taux de mutation (avant correction éventuelle) non-affecté par le niveau de stress endogène.

1.2 Supplementary Materials

Certaines mutations ne mènent pas à un mutant Lorsque une mutation apparaît pendant la réplication, la mutation n'est pas transmise à toute la descendance. Comme on laisse s'échapper une cellule fille à chaque division cellulaire, on ne peut pas garantir que toutes les mutations détectées par l'expérience MV mèneront à des cellules mutantes. Il faut donc remplacer le nombre de mutations détectées par le nombre de mutations transmises dans les cellule observées.

Pour estimer le nombre de mutations transmise par rapport au nombre de mutations détectées, on étudie le nombre de fourches de réplication lors de la division cellulaire. Le taux de mutations transmises est estimé à 12%.

2 Problème à résoudre

Expérimentalement, on est capable de mesurer approximativement la distribution de la fitness au temps t, W_t, et ce pour tout $t \geqslant 0$. On est également capable de montrer que le nombre de mutations N_t est un processus de Poisson. Nous voulons déduire de ces deux informations la distribution des effets des mutations sur la fitness.

On note t_i l'instant de la mutation i, et on définit l'effet relatif s_i de la mutation i ainsi :

$$s_i = \frac{W_{t_{i-1}} - W_{t_i}}{W_{t_{i-1}}}$$

Ainsi, $s_i < 0$ si la mutation est bénéfique, $s_i > 0$ si la mutation est délétère et $s_i = 1$ si la mutation est létale. On fait l'hypothèse que les mutations sont iid, ce qui est justifié par une expérience. On supposera également, sans perte de généralité, que $W_0 = 1$.

On a, pour tout $t \geqslant 0$:

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

On aimerait, avec cette expression, sachant la loi de W_t et la loi de N_t, trouver la loi des s_i.

2.1 Discussion sur les hypothèses effectuées

Indépendance des mutations Les résultats de l'article montrent que les baisses subites et importantes de la fitness, ainsi que les morts non dues à la sénescence, suivent une loi de Poisson. Cela montre une indépendance entre les mutations fortement délétères (qui causent une baisse importante de la fitness) et les mutations létales d'une part, et les mutations relativement neutres acquises précédemment d'autre part.

Cette remarque permet de justifier l'hypothèse de l'indépendance des mutations.

2.2 Résolution par les moments

On a, par indépendance, pour $k \geqslant 1$:

$$\mathbb{E}\left[W_t^k\right] = \mathbb{E}\left[\prod_{i=1}^{N_t} (1 - s_i)^k\right] = e^{\lambda t\left(\mathbb{E}\left[(1-s)^k\right]-1\right)}$$

donc :

$$\mathbb{E}\left[(1-s)^k\right] = \frac{1}{\lambda t}\left(1 + \ln \mathbb{E}\left[W_t^k\right]\right)$$

Comme nous sommes capables d'estimer les $\mathbb{E}\left[W_t^k\right]$, on peut estimer les $\mathbb{E}\left[(1-s)^k\right]$. A priori, cela nous permet de trouver la distribution f de $X = 1 - s$. Cependant, les approximations nécessaires peuvent se révéler très embêtantes.

Méthode par la fonction caractéristique À partir de tous les moments, on peut calculer la fonction caractéristique $\varphi_X(\xi) = \mathbb{E}\left[e^{i\xi X}\right]$. À partir de la fonction caractéristique φ_X, on peut trouver la distribution de X, f. Seulement, on ne connaît que les $N > 0$ premiers moments, chacun avec une certaine erreur $(\varepsilon_k)_{1 \leqslant k \leqslant N}$. Ainsi, on ne va calculer $\varphi_X(\xi)$ 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. Nos calculs montrent que, que pour $x \in \mathbb{R}$:

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

où

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

$$\alpha_1 \leqslant \int_{|x|>A} \left|\varphi_X(\xi) e^{ix\xi}\right| \mathrm{d}\xi = \int_{|x|>A} |\varphi_X(\xi)| \, \mathrm{d}\xi$$

Or, pour tout $k \geqslant 1$: $\left(\mathcal{F}(f^{(k)})\right)(\xi) = (i\xi)^k \varphi(\xi)$ donc :

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

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

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

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 $N+1$. On a :

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

D'après la formule de Taylor avec reste intégral :

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

d'où :

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

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. On a :

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

On a donc :

$$\alpha_3 \leqslant 2\|\varepsilon\|_\infty (e^A - 1)$$

C'est une borne qui est assez mauvaise ; cependant, avec des hypothèses sur l'erreur ε, on pourrait sans doute être plus précis.

Méthode par la transformée de Mellin À partir de tous les moments, on peut calculer la transformée de Mellin

$$\varphi_X(\xi) := \mathbb{E}\left[e^{i\xi X}\right] = \mathbb{E}\left[\sum_{k=0}^{+\infty} \frac{(iX\xi)^k}{k!}\right] = \underbrace{\sum_{k=0}^{N} \frac{(i\xi)^k}{k!} \mathbb{E}\left[X^k\right]}_{\hat{\varphi}_X(\xi)} + \mathbb{E}\left[\sum_{k=N+1}^{+\infty} \frac{(iX\xi)^k}{k!}\right]$$

À partir de la fonction caractéristique φ_X, on peut trouver la distribution de X, f. Seulement, on ne connaît que les $N > 0$ premiers moments, chacun avec une certaine erreur $(\varepsilon_k)_{1\leqslant k\leqslant N}$. Ainsi, on ne va calculer $\varphi_X(\xi)$ 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. On a :

$$\hat{f}(x) = \int_{-A}^{A} \left(\sum_{k=0}^{N} \frac{(i\xi)^k}{k!} m_k \right) e^{i\xi x} \,\mathrm{d}\xi =$$

pour $A > 0$, m_k notre estimation du moment d'ordre k, N l'ordre du moment maximal qu'on peut calculer. et donc :

$$\begin{aligned}\left| f(x) - \hat{f}(x) \right| &= \left| \int_{\mathbb{R}} \left(\sum_{k=0}^{+\infty} \frac{(i\xi)^k}{k!} \mathbb{E}\left[X^k\right] e^{i\xi x} \right) \mathrm{d}\xi - \int_{-A}^{A} \left(\sum_{k=0}^{N} \frac{(i\xi)^k}{k!} m_k \right) e^{i\xi x} \,\mathrm{d}\xi \right| \\ &\leqslant \underbrace{\left| \int_{|x|>A} \left(\sum_{k=0}^{+\infty} \frac{(i\xi)^k}{k!} \mathbb{E}\left[X^k\right] e^{i\xi x} \right) \mathrm{d}\xi \right|}_{\alpha_1} + \underbrace{\left| \int_{-A}^{A} \left(\sum_{k=N+1}^{+\infty} \frac{(i\xi)^k}{k!} \mathbb{E}\left[X^k\right] \right) e^{i\xi x} \,\mathrm{d}\xi \right|}_{\alpha_2} \\ &\quad + \underbrace{\left| \int_{-A}^{A} \left(\sum_{k=0}^{N} \frac{(i\xi)^k}{k!} \left| \mathbb{E}\left[X^k\right] - m_k \right| \right) e^{i\xi x} \,\mathrm{d}\xi \right|}_{\alpha_3}\end{aligned}$$

Par calcul, on peut montrer que, pour $x \in \mathbb{R}$:

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

où

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

$$\alpha_1 \leqslant \int_{|x|>A} \left|\varphi_X(\xi) e^{ix\xi}\right| \mathrm{d}\xi = \int_{|x|>A} |\varphi_X(\xi)| \, \mathrm{d}\xi$$

Or, pour tout $k \geqslant 1$: $\left(\mathcal{F}(f^{(k)})\right)(\xi) = (i\xi)^k \varphi(\xi)$ donc :

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

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

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

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 $N+1$. On a :

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

D'après la formule de Taylor avec reste intégral :

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

d'où :

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

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. On a :

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

On a donc :

$$\alpha_3 \leqslant 2\|\varepsilon\|_\infty (e^A - 1)$$

C'est une borne qui est assez mauvaise ; cependant, avec des hypothèses sur l'erreur ε, on pourrait sans doute être plus précis.

Ils s'agit de majorations assez grossières et que l'on peut sans doute largement affiner.
Dans tous les cas, il reste à estimer, à partir des données que l'on a :

— les erreurs ε_k ;

— les hypothèses que l'on pourrait faire sur f et ses dérivées ;

— $\mathbb{E}\left[X^N e^{AX}\right]$.

2.3 Choses à faire pour la suite

— Tenter de peaufiner les estimations que l'on a sur l'erreur commise lors de l'estimation par les moments ;

— Faire un programme en Python pour tester les résultats que l'on obtiendra en partant d'une distribution fixée pour les s_i ;

— Pour l'étude des mutations les plus délétères, on peut supposer qu'elles sont rares et qu'il est possible d'évaluer leur effet directement en regardant l'évolution de la fitness qu'elles ont causées (ainsi, on suppose qu'une seule mutation fortement délétère est apparue lors de la division, et que l'effet de cette mutation est grand devant les effets des autres mutations) ;

— Utiliser le fait que $\ln W_t$ est une somme de variables aléatoires iid (on peut déduire la loi qui nous intéresse depuis la loi des termes de cette somme). Comme on est capable d'estimer la loi de $\ln W_t$, pour tout t, il peut être intéressant de chercher à déduire la loi qui nous intéresse.