# Impacts de la dominance dans le modèle de fixation de mutations de Kimura

Genome, Populations, Species - Master EvoGEM

Anis Toussirt

## 1 Introduction

En 1866, le botaniste Gregor Mendel établit pour la première fois la loi de la dominance en génétique. Celle-ci affirme qu'un individu reçoit deux variantes d'un même caractère, et que seulement l'une d'entre elles est exprimée : c'est l'allèle dominant. Celui-ci s'exprime au détriment de l'allèle appelé récessif. Cette loi est alors la première étape dans la compréhension des interactions entre allèles. Le concept de dominance s'étendra plus tard aux sous-concepts de co-dominance ou encore de sur ou sous-dominance, qui supposent que les allèles peuvent entrer dans des formes de collaborations, quitte à parfois donner naissance à un caractère nouveau. Ainsi, la dominance a un rôle extrêmement important dans la diffusion d'un caractère au sein d'une population, dans la mesure où la propagation (ou l'extinction) d'un caractère sera conditionnée à son expression. On étudiera ici ce rôle de la dominance, particulièrement son impact dans le cas de fixation d'un gène, c'est-à-dire de diffusion complète au sein d'une population. Pour ce, on s'intéressera au modèle de probabilité de fixation d'un gène mutant de Motoo Kimura, un théoricien de l'évolution dont le modèle probabiliste de diffusion de gènes est peut-être le plus renommé au sein de la communauté scientifique. Dans un premier temps, on présentera ce modèle et on en détaillera la démonstration. Celle-ci est bien souvent intégrée et confondue avec des preuves plus générales de modèles de diffusion, donc il semblait pertinent de l'expliciter ici pour donner une intuition sur son fonctionnement. On effectuera par la suite plusieurs simulations qui ont semblé pertinentes en se plaçant dans différents cadres de dominance et en variant certains paramètres du modèle. Ces simulations ont été réalisées sur Python et sont disponibles en annexe du rapport.

## 2 Modèle de Kimura : Probabilité de fixation d'un gène mutant dans une population

### 2.1 Présentation du modèle

On définit la probabilité $u$ de fixation d'un gène mutant au temps $t$ en fonction de sa fréquence initiale $p$ :

$$\begin{gathered} u : [0,1] \times \mathbb{R}^+ \to [0,1], \\ (p,t) \mapsto u(p,t) \end{gathered}$$

Le modèle que l'on étudiera ici vise à mesurer la probabilité de fixation d'un gène mutant lorsque $t \to \infty$. On notera tout le long de ce rapport la fonction $u$ tel que :

$$u(p) = \lim_{t \to +\infty} u(p,t)$$

On note $M_{\delta x}$ et $V_{\delta x}$ respectivement la moyenne et la variance de changement de fréquence d'un gène dans un temps très faible $\delta$.

Alors le modèle étudié sera le suivant :

$$u(p) = \frac{\int_0^p G(x)\,\mathrm{d}x}{\int_0^1 G(x)\,\mathrm{d}x}$$

avec

$$G(x) = exp(-\int \frac{2M_{\delta x}}{V_{\delta x}}\mathrm{d}x)$$

## 2.2 Démonstration du modèle

Il convient dans un premier temps de comprendre l'origine de ce modèle pour nous permettre d'avoir une intuition sur son fonctionnement. La démonstration de ce modèle part d'une idée plus générale de passage d'une fréquence $p$ à une fréquence $x$ au temps $t$. On commence alors par noter $u(p, x, t)$ cette probabilité. On définit par la suite $g(\delta p, p; \delta t)$ la densité de probabilité pour un changement de fréquence de $p$ à $p+\delta p$ dans un intervalle de temps de $\delta t$. On notera que le temps est mesuré ici en générations : $\delta t$ correspond à une variation d'une génération.

Alors on a (avec $x$ fixé) :

$$u(p, x; t + \delta t) = \int g(\delta p, p; \delta t) u(p + \delta p, x; t) \, \mathrm{d}(\delta \mathrm{p})$$

En effectuant un développement de Taylor de $u(p + \delta p, x; t)$ sur $\delta p$ on obtient :

$$u(p + \delta p, x; t) = u + \delta p \frac{d}{dp}(u) + \frac{\delta p^2}{2!} \frac{d^2}{dp^2}(u) + \frac{\delta p^3}{3!} \frac{d^3}{dp^3}(u) ... \text{ pour } u = u(p, x, t)$$

On a donc en multipliant par $g = g(\delta p, p; \delta t)$ et en intégrant :

$$u(p, x; t + \delta t) = u \int g \, \mathrm{d}(\delta \mathrm{p}) + \frac{d}{dp}(u \int (\delta p) g \, \mathrm{d}(\delta \mathrm{p})) + \frac{1}{2} \frac{d^2}{dp^2}(u \int (\delta p)^2 g \, \mathrm{d}(\delta \mathrm{p})) ...$$

On sait que :

$$\int g \, \mathrm{d}(\delta \mathrm{p}) = 1$$

Alors en divisant l'expression par $\delta t$ :

$$\frac{u(p, x; t + \delta t) - u(p, x; t)}{\delta t} = \frac{d}{dp}(u \frac{1}{\delta t} \int (\delta p) g \, \mathrm{d}(\delta \mathrm{p})) + \frac{1}{2} \frac{d^2}{dp^2}(u \frac{1}{\delta t} \int (\delta p)^2 g \, \mathrm{d}(\delta \mathrm{p}))$$

Ici $\lim\limits_{\delta t \to 0} \frac{1}{\delta t} \int (\delta p) g \, \mathrm{d}(\delta \mathrm{p}))$ et $\lim\limits_{\delta t \to 0} \frac{1}{\delta t} \int (\delta p)^2 g \, \mathrm{d}(\delta \mathrm{p}))$ correspondent respectivement aux moments d'ordre 1 et 2 de $\delta p$ dans l'intervalle de temps $\delta t$. De plus, on admettera que l'auteur décide d'approximer la moyenne du changement de fréquence en une génération par le moment d'ordre 1 et la variance par le moment d'ordre 2.

On notera alors ces deux expressions respectivement $M_{\delta p}$ et $V_{\delta p}$.

En considérant un développement de Taylor de degré 2, on obtient par passage à la limite du taux d'accroissement :

$$\frac{du}{dt} = \frac{V_{\delta p}}{2} \frac{d^2 u}{dp^2} + M_{\delta p} \frac{du}{dp}$$

On rappelle que notre intérêt ici est la fixation d'une mutation lorsque $t \to \infty$, et que l'étude se concentre sur $u(p)$, qui considère donc implicitement que $x = 1$. On a donc $\frac{du(p)}{dt} = 0$.
L'étude se réduit alors à :

$$\frac{V_{\delta p}}{2} \frac{d^2 u(p)}{dp^2} + M_{\delta p} \frac{du(p)}{dp} = 0 \text{ avec les conditions } u(0) = 0 \text{ et } u(1) = 1$$

Il suffit finalement de résoudre cette équation différentielle :

On pose $U(p) = \frac{du(p)}{dp}$ :

$$\text{On a } U'(p) + \frac{2M_{\delta p}}{V_{\delta p}} U(p) = 0 \iff U(p) = \frac{du(p)}{dp} = C_1 e^{-\int \frac{2M_{\delta p}}{V_{\delta p}} \mathrm{dp}}$$

$$\text{Donc : } u(p) = \int_0^p C_1 e^{-\int \frac{2M_{\delta x}}{V_{\delta x}} \mathrm{dx}} \mathrm{dx} + C_2 \text{ et on notera } G(x) = e^{-\int \frac{2M_{\delta x}}{V_{\delta x}} \mathrm{dx}}$$

$C_1$ et $C_2$ deux constantes d'intégration. Sachant les conditions à $p = 0$ et $p = 1$, on a $C_2 = 0$ et $C_1 = \frac{1}{\int_0^1 G(x)\,\mathrm{dx}}$

On obtient donc bien la probabilité $u$ de fixation d'un gène mutant de fréquence initiale $p$ :

$$u(p) = \frac{\int_0^p e^{-\int \frac{2M_{\delta x}}{V_{\delta x}}\,\mathrm{dx}}\mathrm{dx}}{\int_0^1 e^{-\int \frac{2M_{\delta x}}{V_{\delta x}}\,\mathrm{dx}}\mathrm{dx}}$$

# 3 Application à un cas général de sélection

## 3.1 Probabilité de fixation dans le cas d'une dominance intermédiaire

### 3.1.1 Hypothèses

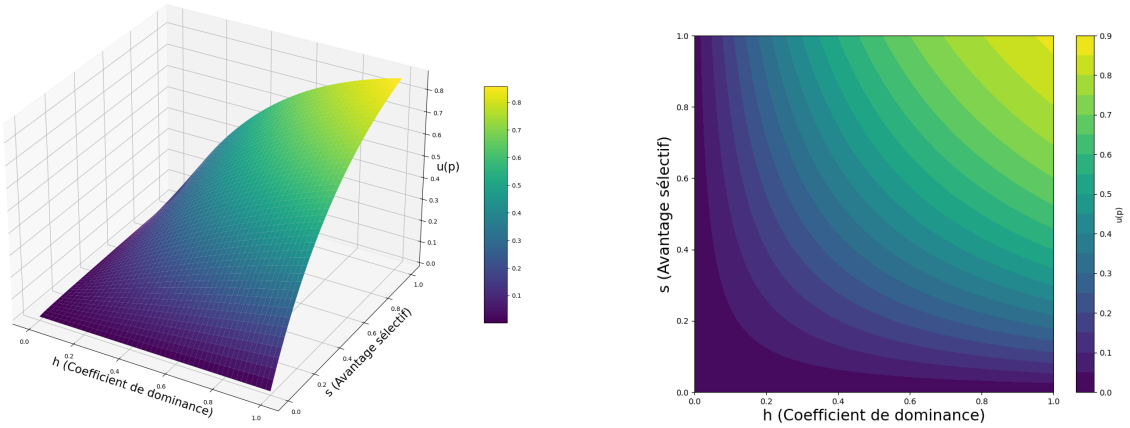
On se place dans un cas général d'apparition d'une mutation, où l'on aura $AA$ le mutant homozygote et $Aa$ le mutant hétérozygote. On leur attribue des coefficients d'avantage, respectivement $s$ et $sh$. Ici $s$ sera le paramètre déterminant le bénéfice phénotypique de la mutation, et $h$ sera le coefficient de dominance. En principe, $s \in [-1, 1]$, avec -1 une mutation complétement délétère et 1 une mutation parfaitement avantageuse. Néanmoins, nous considérerons dans un premier temps que $s \in [0, 1]$, c'est-à-dire que la mutation est neutre dans le pire des cas.
Le coefficient de dominance quant à lui est tel que $h \in [0, 1]$, dans ce cadre de dominance intermédiaire.
$h = 0 \rightarrow$ la mutation est récessive
$h = 1 \rightarrow$ la mutation est dominante
$h \in ]0, 1[\rightarrow$ on a une dominance partielle (potentiellement une codominance si $h = 0.5$). Le coefficient d'avantage $sh$ pour le mutant hétérozygote s'explique donc par le fait que son avantage $s$ est conditionnée à l'expression du gène. Ainsi si l'avantage est parfait (ie $s = 1$), mais que la mutation est récessive, alors l'avantage total de l'hétérozygote sera nul.

Dans ce cadre là, il reste à expliciter $M_{\delta x}$ et $V_{\delta x}$ pour pouvoir faire nos premières simulations.

Dans un premier temps, on a $M_{\delta x} = sx(1 - x)[h + (1 - 2h)x]$
Cette expression représente l'effet moyen de la sélection en fonction des fréquences de $A$ et $a$ (respectivement $x$ et $(1-x)$, et des avantages décrits plus haut. On remarque que si $x$ est très proche de 0, l'impact du coefficient de dominance est beaucoup plus important ($(1 - 2h)x \rightarrow 0$), car la mutation est présente principalement en hétérozygote. De plus, le terme $x(1 - x)$ garantit que le changement de fréquence de l'allèle sera d'autant plus important que sa fréquence $x$ est éloignée des bornes 0 et 1.

Enfin, l'effet de la dérive génétique est détérminé par la variance tel que : $V_{\delta x} = \frac{x(1-x)}{2Ne}$ avec $Ne$ la taille de population effective. Cette formule est en fait le rapport entre la variabilité génétique comme utilisée dans $M_{\delta x}$ et le terme $2Ne$ qui résulte de la définition de dérive génétique : celle-ci est inversement proportionnelle à la taille de population. Donc plus une population est grande (ou plus la variabilité génétique est faible), plus l'effet de la dérive génétique est faible.

### 3.1.2 Impacts de la dominance sur la fixation d'une mutation



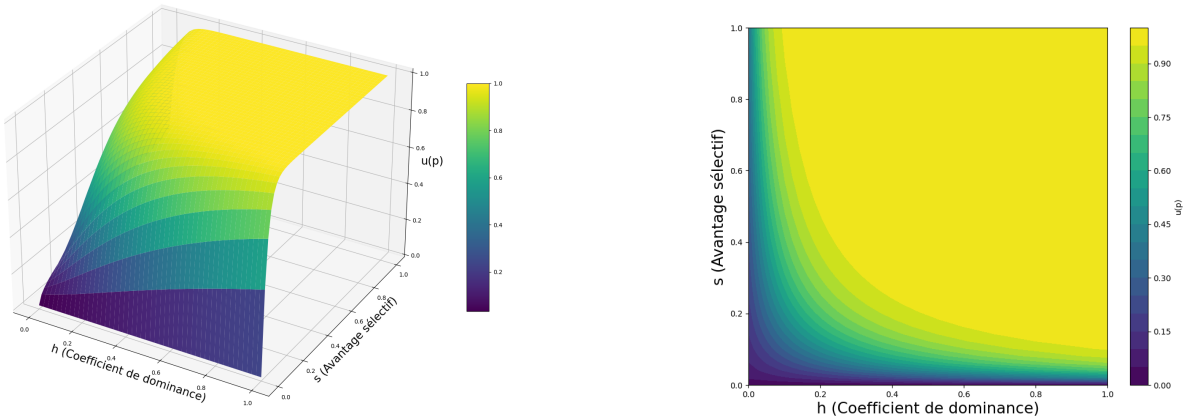

(a) Représentation 3D

(b) Carte thermique

FIGURE 1 – Probabilité de fixation d'un gène mutant ponctuel en fonction de son niveau de dominance et de son avantage selectif - Taille de population : 1000

On constate dans ce cadre, pour un gène mutant complètement récessif qui apparaît ponctuellement, que la probabilité de fixation est forcément nulle, quel que soit son avantage sélectif. Mais en réalité, il est possible d'observer des mutations récessives se fixer dans des populations sous certaines conditions. Alors, quel(s) paramètres ont manqués ici pour envisager cette possibilité ? En fait, même si la mutation est porteuse d'avantages, ils ne s'expriment que lorsque le porteur est homozygote. Or, l'apparition étant ponctuelle, la propagation à court terme est principalement sous forme hétérozygote, donc la sélection n'opère quasiment pas. Ainsi, la fixation de la mutation est impossible.

Pour permettre la fixation d'une mutation récessive, il faudrait alors peut-être une quantité initiale importante de porteurs homozygotes. Donc, plus généralement, une mutation qui apparaît de manière "groupée" dans la population. Un tel événement est par ailleurs réaliste, par exemple en conséquence d'un phénomène de migration de population.




(a) Représentation 3D

(b) Carte thermique

FIGURE 2 – Probabilité de fixation d'un gène mutant présent dans 1% d'une population en fonction de son niveau de dominance et de son avantage selectif - Taille de population effective : 1000

En choisissant une fréquence initiale à $p = 1/100$, on remarque en effet qu'une mutation récessive peut obtenir une probabilité significative de fixation. La difficulté pour une telle mutation est en fait le premier "gap" de propagation que la sélection ne permet pas de franchir. Il lui faut soit compter sur une dérive génétique favorable, soit sur un évènement favorisant comme une migration ou un goulot d'étranglement.

On remarque néanmoins un point commun entre les deux simulations précédentes : une mutation ayant un avantage sélectif nul ne semble pas pouvoir se fixer. Si on s'intéresse à l'impact mathématique du paramètre $s$ sur le modèle, on remarque que $s = 0 \iff M_{\delta x} = 0$. En effet, si une mutation a un avantage sélectif neutre, alors elle n'est pas soumise à la sélection naturelle. Plus précisement, la probabilité de fixation d'une mutation neutre est égale à sa fréquence initiale : $s = 0 \iff u(p) = p$.

De manière générale - et sans surprise - il semblerait que la probabilité de fixation d'un gène mutant est croissante (plus ou moins lentement) du coefficient de dominance et de l'avantage sélectif, dans le cas d'une dominance intermédiaire. Ceci est valable pour n'importe quelle fréquence initiale.

Qu'en est-il si l'on considère des cas de sur-dominance ou de sous-dominance ?

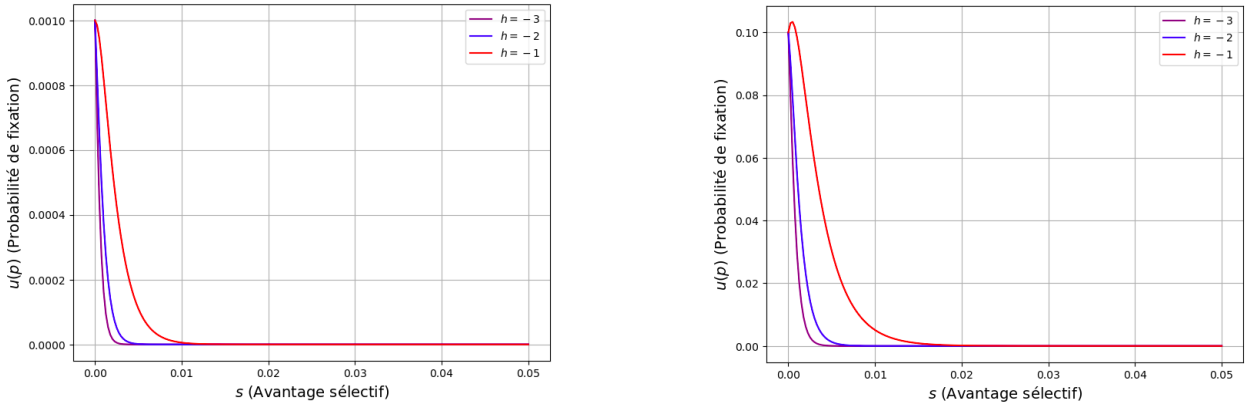
## 3.2 Probabilité de fixation dans les cas de sur/sous-dominance

Les concepts de sur et sous dominance permettent d'élargir la compréhension des interactions entre allèles. Dans le cas des dominances intermédiaires, la valeur sélective de l'hétérozygote $Aa$ était comprise entre celles des homozygotes $AA$ et $aa$. La récessivité d'un allèle la rendait silencieuse face à l'allèle dominant. Ceci permettait d'obtenir exactement la même fitness entre un hétérozygote et un homozygote de l'allèle dominant. Mais dans un cas de sous-dominance, la valeur sélective de l'hétérozygote $Aa$ est inférieure à celle des deux homozygotes $AA$ et $aa$. De manière équivalente, la sur-dominance confère une valeur sélective supérieure à l'hétérozygote face aux homozygotes. L'exemple le plus classique de sur-dominance est celui de l'allèle responsable de la drépanocytose. Cet allèle récessif provoque la mort prématurée des individus porteurs homozygotes, mais permet la résistance au paludisme des porteurs hétérozygotes. Ainsi, l'hétérozygote porte le double avantage d'être résistant au paludisme et de ne pas développer la drépanocytose. Sa valeur sélective est donc supérieure à celle des homozygotes.

### 3.2.1 Hypothèses

Pour revenir au modèle de Kimura, rappelons que les avantages sélectifs sont définis comme ceci : $s$ est l'avantage sélectif du mutant homozygote $AA$ et $sh$ est l'avantage sélectif du mutant hétérozygote $Aa$, avec le coefficient $h$ qui détermine le niveau de dominance de l'allèle mutant.
Pour simuler une sur-dominance, on pourrait donc choisir la combinaison $s < 0$ et $h < 0$ ou $s > 0$ et $h > 1$ et de manière symétrique, choisir $s > 0$ et $h < 0$ ou $s < 0$ et $h > 1$ pour la sous-dominance.

### 3.2.2 Cas de sous-dominance




(a) Fréquence initiale $p = 1/2Ne$

(b) Fréquence initiale $p = 1/10$

FIGURE 3 – Probabilité de fixation d'un gène mutant en fonction de l'avantage sélectif $s$ pour des cas de sous-dominance avec $s > 0$ et $h < 0$ - Taille de population effective : 1000

On remarque que la probabilité de fixation chute en augmentant l'avantage sélectif $s$ (c'est-à-dire celui du mutant homozygote). En effet, on accroit l'écart avec le mutant hétérozygote qui devient de plus en plus délétère. Or pour de faibles fréquences initiales, le mutant est présent majoritairement en hétérozygote et est donc fortement contre-sélectionné. De plus, on remarque un défaut de monotonie de la probabilité de fixation

dans la figure 3*b* : pour les paramètres $p = 1/10$ et $h = -1$, la probabilité de fixation trouve un maximum en $s \approx 0$

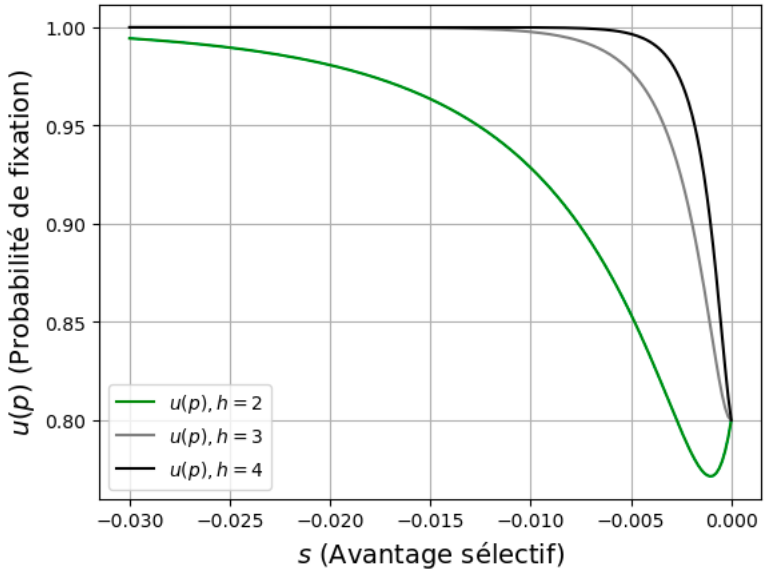
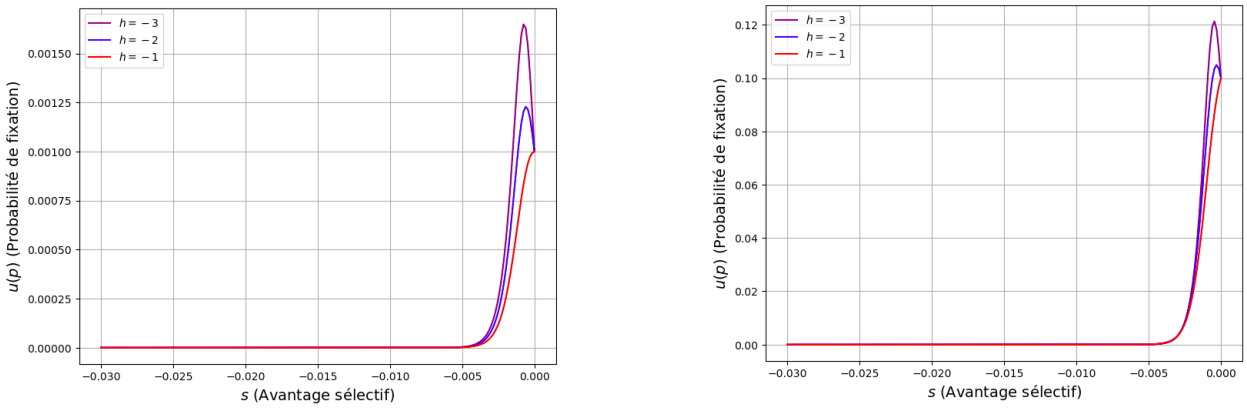



FIGURE 4 – Probabilité de fixation d'un gène mutant en fonction de l'avantage sélectif s pour des cas de sous-dominance avec $s < 0$ et $h > 1$ - Taille de population effective : 1000

Ce défaut de monotonie existe également dans le cas d'une forte fréquence initiale. En effet, pour $p = 0.8$ dans un cas de sous-dominance modélisé par $s < 0$ et $h > 1$, on observe que la probabilité de fixation est globalement décroissante en $s$, mais que le cas $h = 2$ atteint un minimum (toujours pour $s \approx 0$) avant de remonter. De plus on remarque que la probabilité de fixation est plus forte lorsque l'avantage sélectif de l'hétérozygote diminue (c'est-à-dire que $s$ décroit). Ceci pourrait être dû à la faible fréquence de l'allèle ancestral $a$ : en effet, l'hétérozygote est fortement contre-sélectionné et donc l'allèle $a$ également. Ses chances de propagation diminuent ce qui augmente la probabilité de fixation de l'allèle mutant.

On observe ici des résultats beaucoup moins intuitifs que dans le cadre de dominance intermédiaire. Des résultats du même ordre devraient confirmer ceci dans un cadre de sur-dominance.

### 3.2.3 Cas de sur-dominance




(a) Fréquence initiale $p = 1/2Ne$

(b) Fréquence initiale $p = 1/10$

FIGURE 5 – Probabilité de fixation d'un gène mutant présent dans une population en fonction de l'avantage sélectif $s$ pour des cas de sur-dominance avec $s < 0$ et $h < 0$ - Taille de population effective : 1000

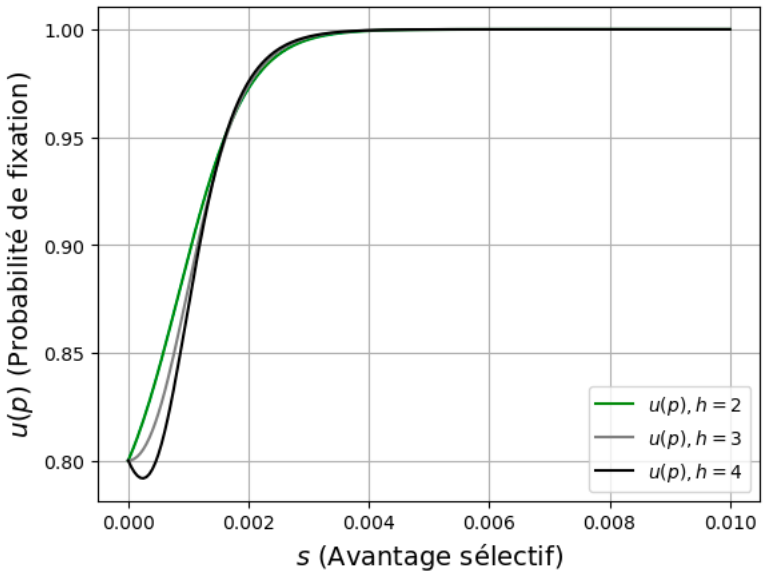



FIGURE 6 – Probabilité de fixation d'un gène mutant en fonction de l'avantage sélectif s pour des cas de sur-dominance avec $s > 0$ et $h > 1$ - Taille de population effective : 1000

Les simulations dans des cas de sur-dominance nous permettent de confirmer que la probabilité de fixation n'est pas monotone de l'avantage sélectif $s$.

De plus, on remarque que la corrélation entre la fitness de l'hétérozygote et la probabilité de fixation n'est pas la même en fonction de la fréquence initiale du gène mutant et du "type" de sur-dominance que l'on applique. En diminuant la valeur de $h$ dans les figures 5a et 5b, ou en augmentant sa valeur dans la Fig 6, on augmente la fitness des hétérozygotes. Or, d'une part, si le gène mutant $A$ est présent en très faible quantité, il le sera principalement en hétérozygote. Donc en augmentant sa fitness, on favorise sa fixation. Ainsi, la probabilité de fixation est inversement proportionnelle à la valeur de $h$ lorsque $p$ est faible. D'autre part, si la fréquence initiale du gène mutant est élevée, alors favoriser les hétérozygotes permet l'augmentation de la fréquence du gène ancestral $a$ et donc une diminution de la capacité du gène mutant à se fixer. Donc pour le cas de la Figure 6, l'impact mathématique de $h$ sera le même mais le sens biologique sera inversé.

# 4 Conclusion

On a voulu s'intéresser ici au modèle mathématique de probabilité de fixation d'une mutation de Kimura, probablement le modèle le plus classique d'évolution de fréquences de gènes. Celui-ci est construit sur une approximation de diffusion qui modélise de petits changements, sur plusieurs générations, dans la fréquence d'un allèle. En évaluant ce modèle sous le prisme de la dominance, on a pu faire des simulations qui permettent d'illustrer l'impact des différents paramètres du modèle en fonction du type de dominance que l'on étudie. Ce travail met en évidence l'importance des interactions complexes entre sélection naturelle, dérive génétique et dominance dans les dynamiques évolutives des mutations. On a montré que le niveau de dominance influence fortement la probabilité de fixation, en interaction avec d'autres paramètres tels que la fréquence initiale et les coefficients sélectifs. Les cas spécifiques de sur-dominance et de sous-dominance offrent des perspectives intéressantes pour comprendre l'évolution de traits complexes dans les populations naturelles.

## 4.1 Discussion

Ce qu'on peut retenir de cette courte étude est la non-trivialité des résultats lors de certaines variations de paramètres. En effet, si les résultats du modèle sont peu surprenants dans le cadre d'une dominance intermédiaire, la sur/sous-dominance réduit fortement nos capacités intuitives à anticiper l'évolution de la probabilité de fixation. Le principal exemple de cette conclusion est la non-monotonie du modèle. Il serait intéressant, dans une étude plus détaillée, de chercher les origines de ces maximums et minimums atteints par la fonction, aussi bien d'un point de vue mathématique que biologique. Plus largement, il serait intéressant de constituer un répertoire complet des combinaisons de paramètres et des impacts sur la probabilité de fixation, et de comparer plusieurs répertoires similaires pour différents modèles mathématiques. La taille de population a par exemple

été fixée tout le long de ce travail, alors que son impact est crucial sur la diffusion d'une mutation, tant en étant fixée à différents ordres de grandeur, qu'en évoluant au cours du temps.

## Code informatique utilisé

Lien GitHub : https://github.com/AnisToussirt/Kimura-Model.git

## Références


[1] KIMURA, M. "On the probability of fixation of mutant genes in a population." Genetics vol. 47,6 (1962) : 713-9. doi :10.1093/genetics/47.6.713

[2] Chen, Christina T L et al. "Effects of dominance on the probability of fixation of a mutant allele." Journal of mathematical biology vol. 56,3 (2008) : 413-34. doi :10.1007/s00285-007-0121-7

[3] Kimura, Motoo. "Diffusion models in population genetics." Journal of Applied Probability 1 (1964) : 177 - 232.

[4] Otto, S.P. and Whitlock, M.C. (2013). Fixation Probabilities and Times. In eLS, (Ed.).