

Loi de Hardy-Weinberg & dérive génique

Élaboration du modèle

Dans les cas simples, un gène peut prendre deux formes (ou allèles) A et a et un individu peut avoir l'un des trois génotypes suivants : A//A, A//a, a//a.

On va modéliser l'évolution de la répartition de ce gène au sein de la population, mais on ne va s'intéresser ni aux phénotypes directement observables, ni aux génotypes précédents : on va considérer la fréquence des allèles dans une population. On imagine une sorte d'urne des gamètes : les deux allèles d'un même gène qui déterminent un caractère se séparent de manière aléatoire lors de la méiose au moment de la formation des gamètes. Chaque gamète ne transporte donc qu'un seul allèle.

Cette urne des gamètes existe de fait chez les poissons (qui déversent leurs gamètes dans l'eau). On peut alors considérer que les rencontres se font au hasard. On va supposer (H1) de plus que ces rencontres ont une même chance de se produire : les gamètes portant tel ou tel allèle n'ont pas moins de chance de féconder, ne sont pas moins fertiles, et les accouplements se font sans choix préférentiel.

Exercice 1 On note p et q les proportions respectives des allèles A et a. On a donc $p+q=1$. Montrer que les génotypes obéissent à la loi de probabilité suivante

Génotype	A//A	A//a	a//a
Probabilité	p^2	$2pq$	q^2

Exercice 2 À l'inverse, si on observe que les fréquences des trois génotypes A//A, A//a et a//a sont respectivement u , v et w , avec $u+v+w=1$, retrouver la loi de probabilité des allèles.

Question de modélisation : la taille de la population (i.e. le nombre d'individus) est-elle une donnée pertinente ? Influence-t-elle les calculs ? Quelle hypothèse fait-on tacitement ? On note (H2) cette hypothèse.

On va maintenant étudier l'évolution des fréquences des allèles et/ou des génotypes au cours des générations. Les données précédentes sont donc celles de la génération de départ, ou génération 0. On note p_0 et q_0 les fréquences alléliques de cette génération et u_0 , v_0 et w_0 les fréquences génotypiques. Plus généralement on note p_n et q_n les fréquences alléliques de $n^{\text{ème}}$ génération et u_n , v_n et w_n les fréquences génotypiques.

On va faire encore quelques hypothèses de modélisation. On suppose qu'il n'y a pas :

- (H3) de croisement entre les générations ;
- (H4) de migrations des adultes ;
- (H5) de mutation des gènes ;
- (H6) d'effet de sélection, i.e. pas d'avantage ou d'inconvénient à être hétérozygote.

Exercice 3 Construire un arbre de probabilité pour calculer u_1 , v_1 et w_1 et retrouver les résultats de l'exercice 1. En déduire que les suites (p_n) et (q_n) sont constantes et que les suites (u_n) , (v_n) et (w_n) sont stationnaires à partir du rang 1, i.e. qu'elles prennent la même valeur à partir du rang 1.

Ce résultat constitue la loi de Hardy-Weinberg.

Lorsqu'une population présente une telle stabilité des fréquences alléliques, on parle d'équilibre de Hardy-Weinberg.

Étude d'exemples

Exercice 4 - Applications de la loi de Hardy-Weinberg

1. Chez l'Homme le gène A de la pigmentation de la peau est dominant par rapport à celui de l'albinisme a. Des recherches montrent que la fréquence relative de l'albinisme a//a vaut 0,005%. Quelle est la fréquence relative des hétérozygotes A//a, porteur du caractère d'albinisme ?
2. Si, comme dans le cas des groupes sanguins AB, A, B et O, on a trois allèles possibles A, B et O, comment adapter les calculs sachant que le groupe A correspond aux génotypes A//A et A//O, le groupe B à B//B et B//O, le groupe AB à A//B et le groupe O à O//O ? Montrer que la répartition O 45%, A 42%, B 9,5% et AB 3,5% est un équilibre de Hardy-Weinberg.
3. La mucoviscidose est une maladie qui frappe un enfant sur 2500. L'étude de sa transmission a montré qu'elle est due à l'état homozygote d'un certain gène, que l'on désigne par m. Les individus hétérozygotes N//m, où N désigne l'allèle normal sont sains et il est même impossible de déceler chez eux l'allèle pathogène, totalement camouflé par l'allèle normal. Quelle est la fréquence de l'allèle pathogène et celle des porteurs sains ?
4. Comparer l'étude de la mucoviscidose avec celle de la phénylcétonurie qui a le même mécanisme, mais dont la fréquence des enfants atteints est de un pour 11 000.

Question de modélisation Dans une population où l'état homozygote d'un gène récessif provoque une maladie mortelle et où les individus atteints ont peu de chance de se reproduire, peut-on vraiment considérer que l'on peut être dans un état d'équilibre de Hardy-Weinberg ? Quelle hypothèse n'est pas vérifiée ?

Exercice 5 - Illustration géométrique

1. On note α la quantité $u-w$. Montrer qu'elle est constante et qu'à partir du rang 1, u_n , v_n et w_n s'expriment comme des fonctions quadratiques de α .
2. Montrer que la loi de Hardy-Weinberg régit cette répartition des fréquences génotypiques si et seulement si $v_n^2 = 4u_n w_n$.
3. Montrer que si loi de Hardy-Weinberg s'applique alors la moitié au plus des individus sont hétérozygotes.
4. Dans un triangle équilatéral (ABC) de hauteur 1 (et donc de côté $2\sqrt{3}/3$), on considère un point M et on appelle H, I et K ses projetés orthogonaux sur les côtés (AB), (BC) et (CA) respectivement. Montrer $MH+MI+MK=1$ et que, réciproquement, si u , v et w vérifiant $u+v+w=1$ sont donnés alors il existe un point M tel que $MH=u$, $MI=v$ et $MK=w$.
5. Montrer $BI/CI=p/q$ où p et q correspondent à u , v et w comme dans l'exercice 2.
6. En déduire que les points correspondant à des équilibres de Hardy-Weinberg sont situés sur une parabole passant par B et C.

Étude d'une maladie invalidante ou mortelle

On considère que seule une proportion k des homozygotes m//m parvient à avoir une descendance. Autrement dit : $q_{n+1} = q_n - (1-k)q_n^2$.

Exercice 6 - Première approximation (pour k petit)

1. On pose $a_n = 1/q_n$. Calculez l'accroissement $\Delta a_n = a_{n+1} - a_n$.
2. Montrer que l'on peut considérer que la suite (a_n) est arithmétique de raison $1-k$.
3. Calculer a_n puis q_n pour tout entier n et chercher pour quel n on a $q_n = q_0/2$.

Complément 1 – Argument pour le calcul de Δq_n

1. On peut faire un calcul exact des proportions : on part des résultats de l'exercice 1, à savoir $u=p^2$, $v=2pq$ et $w=q^2$ et on considère la population entière, de taille P . On a donc uP individus homozygotes N/N , vP porteurs sains hétérozygotes N/m et wP individus malades m/m . Si on ne considère que la partie de la population ayant une descendance, on obtient respectivement uP , vP et kwP individus de chaque type. On a

$$uP+vP+kwP=(u+v+w)P-(1-k)wP=(1-(1-k)w)P.$$

Montrer que, en proportion, la fréquence de l'allèle m est à la génération suivante est $q(kq+1-q)/(1-(1-k)q^2)$.

2. On rappelle que la somme d'une suite géométrique de raison t est donnée par $1+t+t^2+\dots+t^n=(1-t^{n+1})/(1-t)$. En déduire que si t est petit, on peut approcher $1/(1-t)$ par $1+t$ (ou par $1+t+t^2$).
3. En déduire $q(kq+1-q)/(1-(1-k)q^2)\approx q(1-(1-k)q)(1+(1-k)q^2)\approx q(1-(1-k)q+(1-k)q^2)\approx q(1-(1-k)q)$.

Complément 2 - Écriture différentielle

1. On peut récrire ce qui précède sous la forme $-\Delta q/q^2=1-k$. Plutôt que d'écrire q comme une fonction de n , la génération, on l'écrit en fonction du temps t . Montrer qu'alors on peut écrire l'équation précédente sous la forme $(1/q)'=1-k$. (**Attention** : $t=1$ pour une génération, i.e. l'unité de temps est environ 25 ans.)
2. En déduire $1/q-1/q_0=(1-k)t$ puis calculer la durée de réduction de moitié de l'allèle m .
3. Pour la phénylcétonurie on peut prendre $k=0$ et $q_0=1/105$. Montrer que la durée de division par deux est alors 52,5 générations. Qu'en est-il de la mucoviscidose ($k=0$ et $q_0=1/50$) ?

Le paradoxe de la mucoviscidose

Lorsque la fréquence de la maladie dans une population semble constante, comme pour la mucoviscidose, la diminution de p entraînée par la mort d'une partie des personnes atteintes est compensée par d'autres causes, mutations ou avantage sélectif des hétérozygotes par exemple, i.e. on remet en cause les hypothèses (H5) ou (H6).

La guérison de la maladie ne modifiera pas ces autres causes et à l'équilibre succèdera un accroissement progressif de la fréquence du gène. Cet accroissement sera approximativement égal à l'opposé de la diminution que nous venons de calculer. Ainsi, si on parvenait à guérir la mucoviscidose, on pourrait penser qu'en 25 générations la fréquence de l'allèle m serait doublée.

La mucoviscidose n'est fréquente qu'en Europe et est due principalement à la mutation principale du gène CFTR (Cystic Fibrosis Trans-membrane conductance Regulator) sur le chromosome 7, la mutation $\Delta F508$. Sans cette mutation la maladie serait dix fois plus rare. Toujours sans elle l'équilibre pourrait être maintenu en autorisant une mutation de ce gène (remise en cause de (H5)) de $3,6 \cdot 10^{-5}$, ce qui est élevé mais acceptable : c'est certainement cette situation qui prévaut en dehors de l'Europe.

L'avantage aux hétérozygotes a été montré notamment vis-à-vis de la résistance au choléra. Mais le choléra est aussi important en Afrique, en Asie ou en Europe. Par ailleurs l'augmentation de la fréquence des allèles par avantage sélectif (remise en cause de (H6)) est très très lent. À cette échelle de temps, la population humaine est encore très jeune !

Il est plus probable que la mutation $\Delta F508$ n'est survenue qu'une fois dans l'histoire génétique de l'Europe et est suffisamment ancienne (20 000 ou 50 000 ans) pour être survenue dans une population mère de toutes les populations européennes. Ceci expliquerait bien sa présence en Europe et non ailleurs. Ce phénomène est dû à la petitesse de la population : c'est l'hypothèse (H2) qu'il faut remettre en question.

Étude d'un avantage sélectif

Comme précédemment, on note p_n et q_n les fréquences alléliques (allèles A et a) de $n^{\text{ème}}$ génération et u_n , v_n et w_n les fréquences génotypiques (A//A, A//a, a//a). On suppose que les individus dont le génotype est a//a ne peuvent pas se reproduire. Pour le génotype A//A on suppose que le taux de reproduction est t , et pour A//a on suppose qu'il est égal à s . Les taux t et s sont des proportions, donc des nombres entre 0 et 1.

1. Quelle proportion de la population de la $n^{\text{ème}}$ génération participe à la reproduction ?
2. En déduire $p_{n+1} = (tp_n + sq_n) / (tp_n + 2sq_n)$.
3. On note $r = t/s$ et on suppose $0 < r \leq 1$. On définit les fonctions f_r sur $[0;1]$ par la formule $f_r(x) = ((r-1)x+1) / ((r-2)x+2)$. Montrer qu'on a $p_{n+1} = f_r(p_n)$.
4. Cas $r=1$.
 - Tracer la courbe représentative (C_l) de f_1 et étudier la position relative de (C_l) par rapport à la première bissectrice.
 - Démontrer que la suite (p_n) est croissante, comprise entre 0 et 1 et donc convergente.
 - Donner la limite de p_n et interpréter biologiquement.
5. L'anémie falciforme est une très grave maladie génétique, souvent mortelle. Les individus hétérozygotes sont statistiquement plus résistantes au paludisme. L'anémie falciforme peut atteindre 4% des individus de certaines populations vivant dans des régions d'Afrique ou d'Asie où sévit une forme grave de paludisme. Est-il possible que l'on ait $r=1$ dans ce cas ?
6. Cas $0 < r < 1$ et $0 < p_0 < 1$.
 - Tracer la courbe représentative (C_r) de f_r pour trois valeurs de r et étudier la position relative de (C_r) par rapport à la première bissectrice.
 - Étudier les variations de f_r ainsi que les points d'intersection de (C_r) avec la première bissectrice.
 - Choisir une valeur de p_0 comprise entre 0 et $1/(2-r)$ et construire graphiquement les premiers termes de la suite (p_n) . Faire de même sur un tableur puis démontrer que (p_n) est croissante et comprise entre 0 et $1/(2-r)$.
 - Choisir une valeur de p_0 comprise entre $1/(2-r)$ et 1 et construire graphiquement les premiers termes de la suite (p_n) . Faire de même sur un tableur puis démontrer que (p_n) est décroissante et comprise entre $1/(2-r)$ et p_0 .
 - En déduire que (p_n) converge vers une limite dépendant de r et non de p_0 .
 - Pouvez-vous estimer la valeur de r pour l'anémie falciforme ?
 - Pouvez-vous estimer la vitesse de convergence de la suite (p_n) à l'aide d'un tableur ?

Étude d'un test pour la mucoviscidose

1. On reprend l'étude de l'exercice 4, question 3. Quelle est la fréquence des couples à risque (i.e. des couples formés de deux individus hétérozygotes) ? Quelle est alors la fréquence de naissance d'un enfant atteint de la mucoviscidose ?
2. Quelle est le risque d'avoir un enfant atteint pour un couple : ayant déjà eu un enfant atteint de mucoviscidose ? pour un couple dont l'un des deux a un frère ou une sœur ayant eu un enfant atteint de mucoviscidose ? Comparer avec le calcul sans avoir de connaissances sur les risques supplémentaires.

Le gène CFTR code pour une chaîne peptidique de 1480 acides aminés. Plus de 500 mutations du gène CFTR sont responsables de la mucoviscidose (caractérisée par une dégénérescence fibreuse du pancréas et des poumons). Leur étude est très longue et très chère. Si on teste uniquement les plus courantes, l'étude reste longue et chère, et de plus sans garantie de passer à côté d'une mutation non recherchée.

Or, en amont du gène CFTR existent deux sites facultatifs constituant un polymorphisme de l'ADN de type RFLP (Polymorphisme de Longueur de Fragments de Restriction) car le nombre et la longueur des fragments de restriction, générés à cet endroit du génôme, sera fonction de l'absence ou de la présence de ces sites facultatifs. Selon qu'à cet emplacement du chromosome 7 les sites sont présents (+) ou absents (-), on peut définir quatre associations appelées haplotypes. Un chromosome 7 porteur de la mutation $\Delta F508$ est aussi porteur dans 91,4% des cas de l'haplotype B (-+), alors que dans le cas inverse il ne l'est que dans 19,4%. Il s'agit donc d'une association physique ou cartographique, et non pas fonctionnelle, entre l'haplotype B et la mutation $\Delta F508$.

On pense que cette association provient du fait que la mutation est ancienne (environ 40 000 ans) et est survenue sur un chromosome porteur de l'haplotype B.

On peut pratiquer un test sur cet haplotype pour décider de rechercher les risques de mutation $\Delta F508$, ceci évitant des coûts de recherche, mais on arrive alors face à des problèmes de faux négatifs et de faux positifs.

TaqI	PstI	Haplotype	Fréquence générale	Fréquence relative sur un chromosome non muté	Fréquence relative sur un chromosome muté
-	-	A	32,60%	33,80%	1,30%
-	+	B	22,40%	19,40%	91,40%
+	-	C	33,50%	34,80%	0,50%
+	+	D	11,40%	12,10%	6,80%

Comparer le test sur l'haplotype B et celui sur B/D, i.e. celui uniquement sur le site de restriction PstI.

Étude de la dérive génique

On considère maintenant que la taille de la population est fixée, disons N .

D'une génération à l'autre, les N individus suivants sont obtenus à partir des N précédents aléatoirement, ce que l'on peut modéliser par N expériences aléatoires où l'on choisit parmi les $2N$ allèles présents dans la génération précédente deux des allèles. On peut simplifier encore en considérant, pour étudier uniquement la fréquence allélique et non la fréquence génotypique, $2N$ expériences aléatoires de type Bernoulli de paramètre p . Par conséquent p_{n+1} est égal à $X_n/2N$ où X_n suit une loi binomiale de paramètres $2N$ et p_n .

On a donc $E(p_{n+1})=p_n$ et $Var(p_{n+1})=p_n(1-p_n)/2N$. On retrouve l'équilibre de Hardy-Weinberg en ce sens que p_{n+1} et p_n sont en moyenne égaux, mais on voit l'influence de la taille de la population sur les fluctuations autour de cette moyenne. On peut également calculer la probabilité d'extinction d'un gène dans ce modèle.

On peut affiner le modèle en donnant un certain nombre de descendants à chaque individu (selon une loi de Poisson par exemple) et simuler la descendance pour tester l'extinction d'un gène.

Étude de la panmixie – consanguinité

Les populations ne sont pas souvent panmictiques : les individus se reproduisent avec ceux qui sont proches d'eux, voire avec qui ils sont apparentés.

Dans le cas extrême d'auto-fécondation, les descendants d'un hétérozygote ont une probabilité $\frac{1}{2}$ d'être homozygotes et ainsi la proportion d'homozygotes va augmenter, jusqu'à ce qu'il ne reste plus que de tels individus. Néanmoins, a priori, les fréquences alléliques ne vont pas être modifiées si la population est suffisamment grande.

D'une façon générale, s'il y a consanguinité la proportion d'hétérozygotes diminue au fil des générations.

Si, de plus, la population est petite, il y a un risque d'extinction d'une des lignées homozygotes. C'est ce genre de phénomène qui est à l'origine (probablement) de l'importance de la mucoviscidose en Europe.

Il y a aussi un risque d'effet fondateur par la remise en cause de l'hypothèse (H4). Si un petit groupe de migrants fonde une nouvelle communauté et que les fréquences alléliques y sont déséquilibrées par rapport à la population mère, il y a très vite une dérive génique et certains gènes seront éliminés alors que d'autres, pourtant très rares dans la population mère, se fixent définitivement.

Un exemple plus simple est celui de la disparition d'un élément présent sur le chromosome Y. On simule la dynamique du phénomène en donnant une distribution initiale des allèles A et a dans la population mâle. Puis on attribue à chaque mâle une descendance mâle selon une loi de Poisson (par exemple de paramètre 1) et on étudie au bout de combien de temps l'un des deux allèles disparaît.

Glossaire

La synthèse des **protéines** nécessite l'existence de **gènes** pour disposer les constituants élémentaires des protéines (les acides aminés) linéairement dans le bon ordre. Les protéines sont soit des enzymes, des hormones, des protéines de structure des parois des cellules, des protéines de régulation de l'expression d'autres gènes etc.

Un **gène** est un fragment de la molécule d'ADN. Ils sont présents dans les **chromosomes**, mais comme ceux-ci vont par paires, les gènes sont répétés deux fois. On les nomme les **allèles** du gène. Chez un individu donné il n'y a que deux allèles, mais dans une population il peut y en avoir jusqu'à 20 et plus.

Lors de la méiose, au moment de la formation des gamètes, les paires de **chromosomes** se séparent de manière aléatoire, chaque gamète emportant ainsi un seul des **allèles** du **gène**. Parmi les chromosomes, il faut distinguer les chromosomes **sexuels** (X et Y) et les autres (**autosomes**).

Si les deux allèles sont identiques, l'individu est dit **homozygote**. Sinon il est dit **hétérozygote**. Si les deux allèles sont différents, ils peuvent s'exprimer tous les deux (**co-dominants**) ou bien l'un est **dominant** et l'autre est **récessif** (souvent, il ne s'exprime pas).

L'apparence visible d'une espèce est son **phénotype**. Les gènes réellement présents ne s'expriment pas tous, il faut donc distinguer le **génotype** du phénotype.

L'expression d'un allèle peut aussi dépendre de divers facteurs du milieu comme la température.

Les maladies qui ne dépendent que d'un seul gène sont dites **monogéniques**.