Génétique des populations et évolution moléculaire

Groupe de travail, 23 mars 2005

Laboratoire de Biodiversité-Hydrobiologie

I - Introduction : la loi de Hardy-Weinberg

BUT : étude de l'évolution du patrimoine génétique d'une population

On considère un gène avec 2 allèles $\mathbf{A_1}$ et $\mathbf{A_2}$

$\rightsquigarrow$ fréquence des génotypes $\mathbf{A_1A_1}$ (p_1), $\mathbf{A_1A_2}$ (2q) et $\mathbf{A_2A_2}$ (p_2)

Fréquence des génotypes aux générations suivantes ?

<u>génération 1</u> :	<u>génération 2</u> :
$\mathbf{A_1A_1}$ $\left((p_1+q)^2 \right)$	$\mathbf{A_1A_1}$ $\left((p_1+q)^2 \right)$
$\mathbf{A_1A_2}$ $\left(2(p_1+q)(q+p_2) \right)$	$\mathbf{A_1A_2}$ $\left(2(p_1+q)(q+p_2) \right)$
$\mathbf{A_2A_2}$ $\left((q+p_2)^2 \right)$	$\mathbf{A_2A_2}$ $\left((q+p_2)^2 \right)$

Loi de Hardy-Weinberg

Stabilité de la répartition des génotypes dès la $2^{\text{ème}}$ génération

$\rightsquigarrow \mathbf{A_1A_1}\ (\ \alpha^2\ ),\ \mathbf{A_1A_2}\ (\ 2\alpha(1-\alpha)\ )$ et $\mathbf{A_2A_2}\ (\ (1-\alpha)^2\ )$

Remarque : fortes hypothèses sur la population

- panmixie
- population de grande taille
- population isolée (ni migrations, ni mutations)

II - Modèle de Wright-Fisher

1 - Hypothèses

On suppose : - population diploïde de taille constante N

- fécondations au hasard et indépendantes
- population isolée (pas de migrations)
- ni mutation, ni sélection

On note X_t le nombre d'allèles $\mathbf{A_1}$ portés par la génération t, $t \in \mathbb{N}$.

$\rightsquigarrow$ génération initiale : N individus et X_0 allèles $\mathbf{A_1}$

$\rightsquigarrow$ génération t : N individus et X_t allèles $\mathbf{A_1}$

⚠ X_t est une variable aléatoire !

On connaît l'ensemble des valeurs possibles pour X_t $(0,\ 1,\ 2 \cdots$ ou $2N)$,
ainsi que les fréquences respectives.
Mais la valeur prise va dépendre de l'expérience.

2 - Loi de X_t

BUT : calculer la loi de X_{t+1} connaissant X_t

$\rightsquigarrow$ Effectif des génotypes à la génération t :

$\mathbf{A_1A_1}$ (n_1), $\mathbf{A_1A_2}$ $(2m)$ et $\mathbf{A_2A_2}$ (n_2) $\quad\Longrightarrow\quad X_t = 2n_1 + 2m$

$\rightsquigarrow$ Fréquence des génotypes à la génération t :

$\mathbf{A_1A_1}$ $(\frac{n_1}{N})$, $\mathbf{A_1A_2}$ $(\frac{2m}{N})$ et $\mathbf{A_2A_2}$ $(\frac{n_2}{N})$

$\rightsquigarrow$ Fréquence des génotypes à la génération suivante ?

Soit Z_k = nb d'allèles $\mathbf{A_1}$ du k$^{\text{ème}}$ individu de la génération $t+1$

$$\begin{array}{lrcl}
\mathbf{A_1A_1}: & \mathbb{P}(Z_k = 2) & = \left(\frac{n_1}{N} + \frac{m}{N}\right)^2 & = \left(\frac{X_t}{2N}\right)^2 \\
\mathbf{A_1A_2}: & \mathbb{P}(Z_k = 1) & = 2\left(\frac{n_1}{N} + \frac{m}{N}\right)\left(\frac{m}{N} + \frac{n_2}{N}\right) & = 2\,\frac{X_t}{2N}\left(1 - \frac{X_t}{2N}\right) \\
\mathbf{A_2A_2}: & \mathbb{P}(Z_k = 0) & = \left(\frac{m}{N} + \frac{n_2}{N}\right)^2 & = \left(1 - \frac{X_t}{2N}\right)^2
\end{array}$$

Rappelons que X_t est connu. On pose $X_t = i$.

$$\begin{array}{rlll} \text{Comme} & \mathbb{P}(Z_k = 2 | X_t = i) & = & \left(\frac{i}{2N}\right)^2 \\ & \mathbb{P}(Z_k = 1 | X_t = i) & = & 2\,\frac{i}{2N}\left(1 - \frac{i}{2N}\right) \\ & \mathbb{P}(Z_k = 0 | X_t = i) & = & \left(1 - \frac{i}{2N}\right)^2 \end{array}$$

alors la loi de Z_k sachant $X_t = i$ est une loi binomiale $\mathcal{B}\left(2, \frac{i}{2N}\right)$

On répète une expérience aléatoire plusieurs fois (on examine chaque chromosome du k$^{\text{ème}}$ individu) et on compte le nombre de fois où un critère particulier c apparaît (c = allèle $\mathbf{A_1}$).
Les résultats doivent être indépendants
(le fait que le 2$^{\text{ème}}$ chromosome porte ou non $\mathbf{A_1}$ ne dépend pas de l'état du 1$^{\text{er}}$).
On obtient une loi binomiale $\mathcal{B}$(nb de répétitions de l'expériences , probabilité que c apparaisse).

On veut calculer $X_{t+1} = Z_1 + Z_2 + \cdots + Z_N$

$\Rightarrow$ la loi de X_{t+1} sachant $X_t = i$ est une loi binomiale $\mathcal{B}\left(2N, \frac{i}{2N}\right)$

(on considère les 2N allèles des individus de la génération t+1 et on compte le nombre de fois où l'allèle $\mathbf{A_1}$ apparaît)

3 - Stabilité moyenne de la fréquence de l'allèle A_1

Comme X_{t+1} sachant $X_t = i$ suit la loi binomiale $\mathcal{B}\left(2N, \frac{i}{2N}\right)$, alors

a) $\mathbb{P}(X_{t+1} = j | X_t = i) = C_{2N}^{j} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{2N-j}$

$\rightsquigarrow$ $(X_t)_t$ est une chaîne de Markov de matrice de transition

$$P = (P_{ij})_{i,j} \quad \text{avec} \quad P_{ij} = C_{2N}^{j} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{2N-j}$$

La valeur future de X_{t+1} ne dépend du passé que par la valeur présente de X_t.

b) $\mathbb{E}(X_{t+1} \,|\, X_t = i) = 2N \times \frac{i}{2N} = i$

$\rightsquigarrow$ effectif constant en moyenne ; stabilité de Hardy-Weinberg ?

Le nombre moyen d'apparitions du critère c au cours des expériences indépendantes est :

nombre de répétitions de l'expérience $\times$ probabilité d'apparition de c

c) $Var(X_{t+1} \,|\, X_t = i) = 2N \times \frac{i}{2N} \times \left(1 - \frac{i}{2N}\right) = i\left(1 - \frac{i}{2N}\right)$

4 - Perte définitive d'un allèle

La chaîne de Markov a 2 états absorbants : 0 et $2N$

Tous les autres états sont transients.

$\Rightarrow$ perte définitive d'un allèle.

a) Combien de temps avant le dérive génique ?

On note T l'instant de perte d'un allèle.

T est un temps d'arrêt.

L'instant d'arrêt ne dépend que du passé et du présent.

$$\mathbb{E}(T) \;=\; -4N\left[\,\left(1-\tfrac{X_0}{2N}\right) ln\left(1-\tfrac{X_0}{2N}\right) \;+\; \tfrac{X_0}{2N}\, ln\left(\tfrac{X_0}{2N}\right)\,\right]$$

b) Quel allèle reste ?

$$\mathbb{E}(X_{t+1}\,|\mathcal{F}_t) \;=\; \mathbb{E}(X_{t+1}\,|X_t) \text{ car } (X_t)_t \text{ est une chaîne de Markov}$$

$$= \sum_{i=0}^{2N} \mathbb{E}(\,X_{t+1}|X_t = i\,)\,\mathbb{1}_{\{X_t=i\}} \;=\; \sum_{i=0}^{2N} i\,\mathbb{1}_{\{X_t=i\}} \;=\; X_t$$

$\rightsquigarrow$ $(X_t)_t$ est une martingale

Connaissant le comportement passé de $(X_t)_t$: $X_0, X_1, \cdots X_t$,

la meilleure prédiction qu'on puisse faire pour la valeur future de X_{t+1} est X_t.

Par conséquent, $\forall\, t \in \mathbb{N},\ \mathbb{E}(X_t) = X_0$. En moyenne, une martingale est constante.

Et aussi, $\mathbb{E}(X_T) = X_0$

avec $\mathbb{E}(X_T) = 0 \times \mathbb{P}(X_T = 0) + 2N\,\mathbb{P}(X_T = 2N)$

$\Longrightarrow$ $\mathbb{P}(\mathbf{A_2}$ disparaît $) = \frac{X_0}{2N}$

III - Application aux mutations

1 - Probabilité de fixation d'une mutation

instant initial = apparition d'une mutation

On note $\mathbf{A_1}$ l'allèle muté et $\mathbf{A_2}$ l'ensemble des autres allèles.

Soit X_t le nombre d'allèles $\mathbf{A_1}$ portés par la génération t, $t \in \mathbb{N}$.
Remarque : $X_0 = 1$

D'après précédemment,
$\mathbb{P}$(fixation de la mutation) = $\mathbb{P}(\mathbf{A_2}$ disparaît) = $\frac{X_0}{2N} = \frac{1}{2N}$

2 - Théorie de l'horloge moléculaire

Le taux d'apparition des mutations est constant
(indépendant de t et du locus considéré).

Notons μ le taux de mutation par locus, par génération
et f_0 la proportion de mutations neutres (non létales).

$\rightsquigarrow$ nombre moyen de mutations neutres par locus, par génération
$= 2Nf_0\mu$. On note $\frac{\theta}{2} = 2Nf_0\mu$.

Soit S_t = nombre de mutations neutres apparues en t générations.
Alors $(S_t)_{t\geq 0}$ est un processus de Poisson de taux $\frac{\theta}{2}$.

On observe l'occurence d'événements qui surviennent avec un taux constant ;
les nombres d'occurences sur deux intervalles disjoints sont indépendants ;
il ne se produit pas 2 occurrences de l'événement simultanément.
Le temps d'attente entre deux occurrences successives est sans mémoire.

Remarque : pour des considérations phylogéniques, on mesure le
temps en unités de générations ou de $2N$ générations.

$\rightsquigarrow$ taux de mutation neutre $= 2N f_0 \mu \times \mathbb{P}(\text{ fixation d'une mutation })$.

$\Longrightarrow$ taux d'évolution $k_0 = f_0 \mu$.

Soit K = nombre de mutations neutres apparues en T générations susceptibles d'être fixées.

Alors K suit une loi de Poisson $\mathcal{P}(k_0 T)$.

Remarque : $\mathbb{E}(K)$ est indépendant de la taille de la population !

On observe l'occurence d'événements indépendants qui surviennent avec un taux constant ; le temps entre deux occurences successives est sans mémoire.

IV - Différences entre séquences d'ADN

1 - Hypothèses

On suppose : - population diploïde de taille constante N

- fécondations au hasard et indépendantes
- population isolée (pas de migrations)
- pas de sélection

$\mathbb{P}\Big($ 2 chromosomes de la génération t+1 soient issus du même chromosome de la génération t $\Big) = \frac{1}{2N}$

Soit M = nombre de générations écoulées depuis l'ancêtre commun.

Alors M suit une loi géométrique $\mathcal{G}\left(\frac{1}{2N}\right)$.

On répète une expérience aléatoire plusieurs fois (on examine les générations antérieures)
et on compte le nombre de fois où on répète l'expérience avant d'obtenir un succès (ancêtre commun).

$\rightsquigarrow$ temps moyen d'attente $= \mathbb{E}(M) = 2N$ générations.

2 - Temps de coalescence

On s'intéresse au temps de coalescence de 2 chromosomes.

$T_2 =$ temps de coalescence $= \frac{M}{2N}$: unité de $2N$ générations.

Comme $\mathbb{P}(M = m) = \left(1 - \frac{1}{2N}\right)^{m-1} \left(\frac{1}{2N}\right)$,

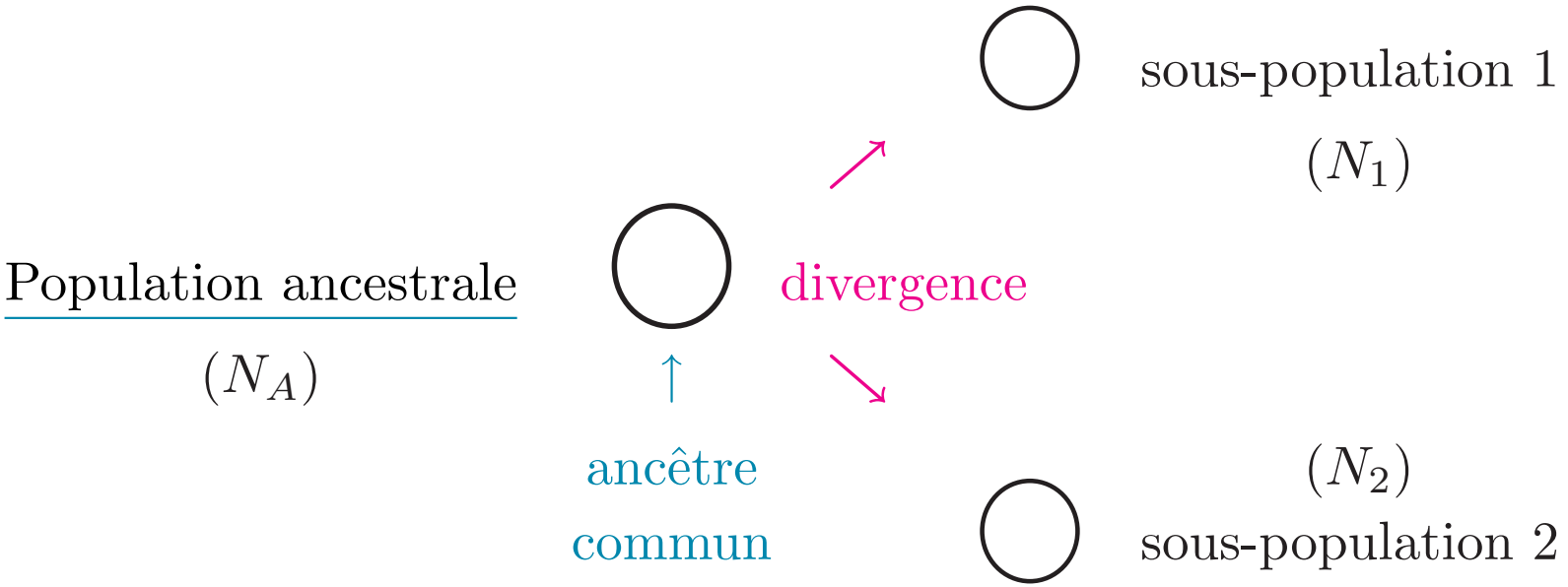
alors $\mathbb{P}(T_2 \leq x) = \sum_{m=1}^{2Nx} \mathbb{P}(M = m) = 1 - \left(1 - \frac{1}{2N}\right)^{2Nx}$

On se place dans le cadre d'une population de grande taille.
alors $\mathbb{P}(T_2 \leq x) \simeq 1 - e^{-x}$
donc le temps de coalescence T_2 suit une loi exponentielle $\mathcal{E}(1)$.

De même que la loi géométrique, la loi exponentielle est sans mémoire.

3 - Nombre de mutations depuis la divergence

Population ancestrale composée de N_A individus.
Il y a T générations, divergence en 2 sous-populations.
Considérons 1 séquence d'ADN dans chaque sous-population.



Depuis l'instant de divergence, chaque séquence a subi
indépendamment une évolution due à des mutations.

Soit K_i = nombre de mutations neutres apparues en T générations sur la séquence i, susceptibles d'être fixées.

horloge moléculaire $\Longrightarrow$ K_i suit une loi de Poisson $\mathcal{P}(k_0 T)$.

Remarque : Le taux d'évolution $k_0 = f_0 \mu$ ne dépend pas de l'effectif des populations N_1 et N_2.

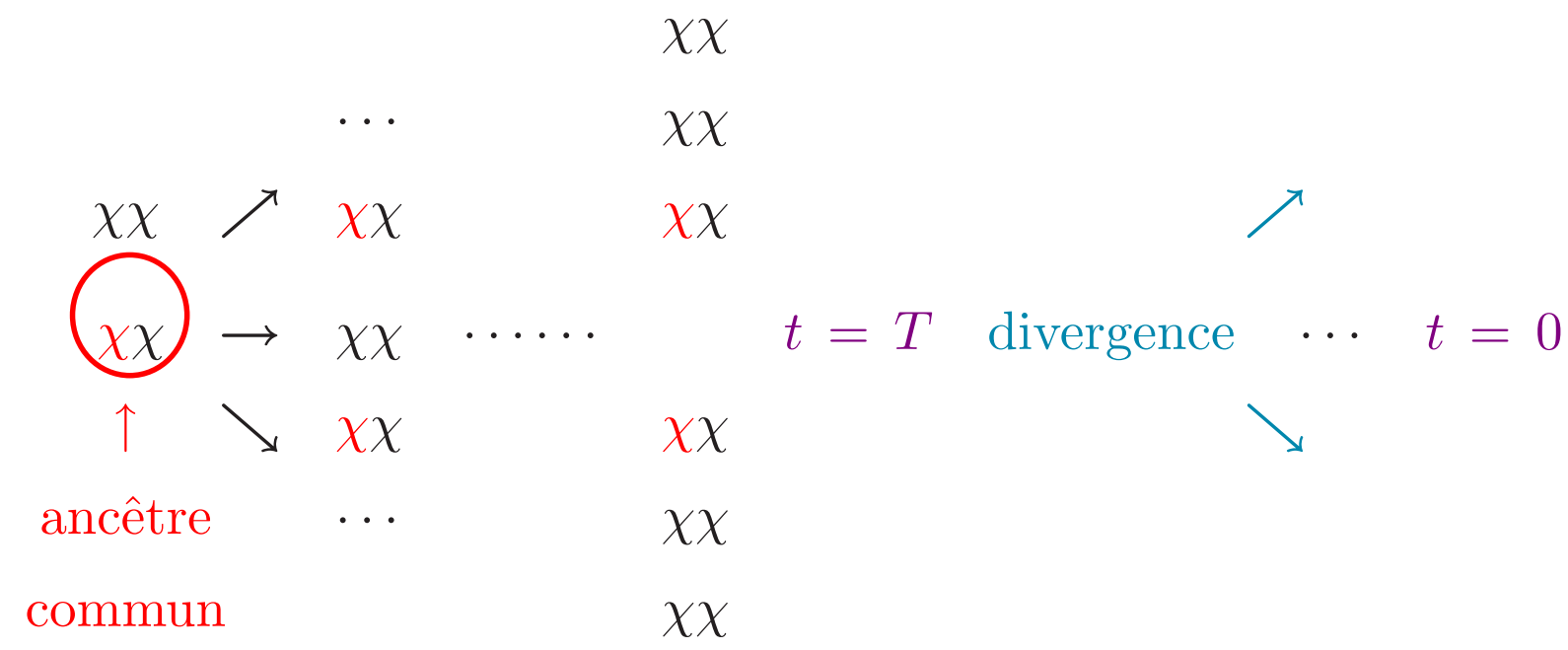
Ainsi, $K_1 + K_2$ suit une loi de Poisson $\mathcal{P}(2k_0 T)$ ($k_0 = f_0 \mu$).

$\Longrightarrow$ $\mathbb{E}(K_1 + K_2) = 2 f_0 \mu T$ et $Var(K_1 + K_2) = 2 f_0 \mu T$.

4 - Nombre de mutations ancestrales

On note χ le chromosome transmis par l'ancêtre commun.

Population ancestrale ($2N_A$ chromosomes) $\rightsquigarrow$ 2 sous-populations



⚠ La mesure du temps est rétrograde !

Depuis l'instant de coalescence, chaque séquence a subi indépendamment une évolution due à des mutations.

Soit K_t^i = nombre de mutations neutres apparues en t générations sur la séquence i, susceptibles d'être fixées.

horloge moléculaire $\Longrightarrow$ $(K_t^i)_{t\geq 0}$ processus de Poisson $(f_0\mu)$

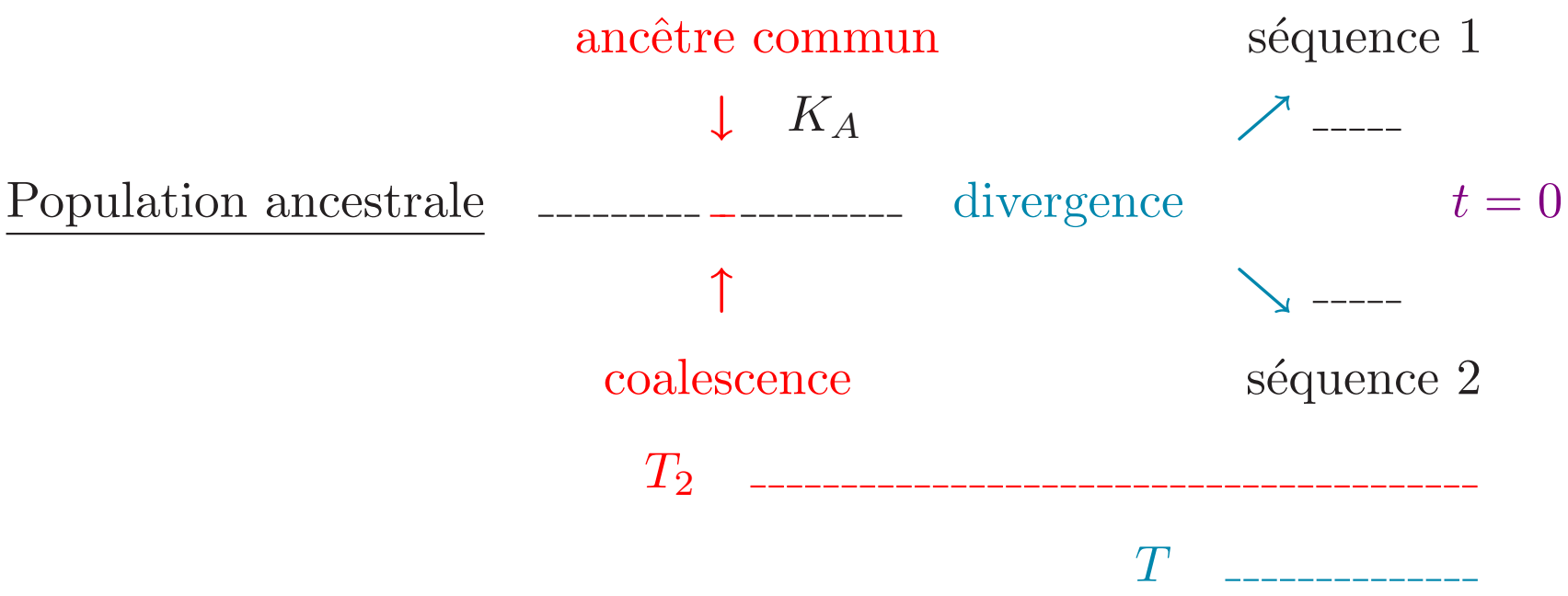
Ainsi, $(K_t^1 + K_t^2)_{t\geq 0}$ est un processus de Poisson, de paramètre $2f_0\mu$.

$\Longrightarrow$ t_M = temps d'attente entre deux mutations susceptibles d'être fixées sur l'une ou l'autre des séquences,

$\sim$ loi exponentielle $\mathcal{E}\big(2f_0\mu\big)$

$\Longrightarrow$ T_M = temps d'attente en unités de $2N_A$ générations

$\sim$ loi exponentielle $\mathcal{E}\big(\theta_A\big)$ $\quad\big(\theta_A = 4N_A f_0\mu\big)$



Les séquences 1 et 2 sont issues de populations qui ont divergé.
L'ancêtre commun se trouve donc dans la population ancestrale.
D'où $T_2 \geq T$.

T_2 est sans mémoire $\Longrightarrow$ $\mathbb{P}(T_2 \geq T + s | T_2 \geq T) = \mathbb{P}(T_2 \geq s)$.

Le temps d'attente de la coalescence, depuis l'instant de divergence suit une loi $\mathcal{E}(1)$.

T_2 et T_M sont indépendantes, avec $T_2 \sim \mathcal{E}(1)$ et $T_M \sim \mathcal{E}(\theta_A)$

Le temps d'attente avant la coalescence ne dépend pas du nombre de mutations opérées sur le chromosome.

donc $\mathbb{P}(T_M < T_2) = \frac{\theta_A}{1+\theta_A}$

K_A = nombre de mutations ancestrales (entre la divergence et la coalescence)

L'événement $\{T_M < T_2\}$ est un échec répété k fois avant le premier succès $\{T_M \geq T_2\}$.

On répète donc l'expérience $k+1$ fois avant d'obtenir le premier succès de probabilité $\frac{1}{1+\theta_A}$.

Ainsi, $K_A+1 \sim$ loi géométrique $\mathcal{G}\left(\frac{1}{1+\theta_A}\right)$

Alors, $\mathbb{P}(K_A = k) = \left(\frac{\theta_A}{1+\theta_A}\right)^k \frac{1}{1+\theta_A}$

et $\mathbb{E}(K_A) = \theta_A, \quad Var(K_A) = \theta_A(1+\theta_A)$

La translation des données translate le moyenne, mais le coefficient de dispersion est identique.

Autre point de vue : file M/M/1

Temps de service $=\ T_2\ \sim \mathcal{E}(1)$

Processus arrivée clients (mutation susceptibles d'être fixées)
$=\ (S_t)_{t\geq 0}$ processus de Poisson de paramètre θ_A.

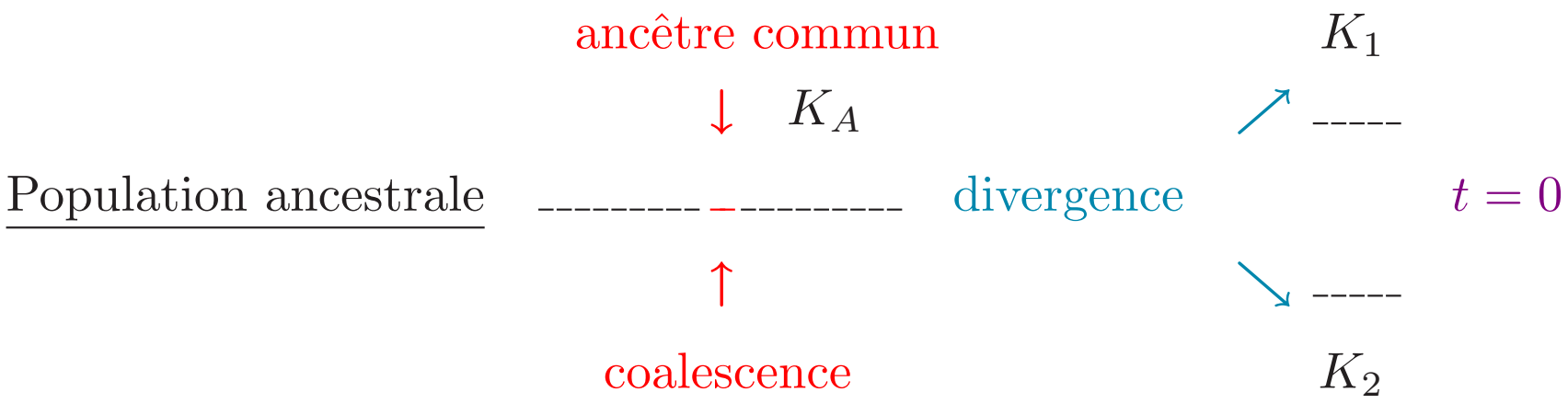
T_2 est indépendant du processus d'arrivée des mutations.

On compte le nombre de clients K_A qui arrivent pendant le temps de service d'un client :

$$\begin{aligned}
\mathbb{P}(K_A = k) &= \int_0^\infty \mathbb{P}(S_t = k | T_2 = t)\ e^{-t}\ dt \\
&= \frac{\theta_A^{k+1}}{(k+1)!} \int_0^\infty t^{k+1}\, e^{-(\theta_A+1)t} \\
&= \Big(\frac{\theta_A}{1+\theta_A}\Big)^k\ \frac{1}{1+\theta_A}
\end{aligned}$$

5 - Nombre de mutations totales

K = nombre de mutations totales (depuis la coalescence).



Alors, $K = K_A + K_1 + K_2$, somme de variables indépendantes.

$$\Longrightarrow \quad \mathbb{E}(K) = \mathbb{E}(K_A) + \mathbb{E}(K_1 + K_2) = \theta_A + 2\mu f_0 T$$

$$\text{et} \quad Var(K) = Var(K_A) + Var(K_1 + K_2) = \theta_A(1 + \theta_A) + 2\mu f_0 T$$

Conclusion :

Malgré l'hypothèse de l'horloge moléculaire, le processus du nombre de mutations susceptibles d'être fixées depuis la coalescence n'est pas un processus de Poisson si la population ancestrale a divergé.

Si c'était le cas, K suivrait une loi de Poisson,
et on aurait $\mathbb{E}(K) = Var(K)$.

$$\text{Or, } \frac{Var(K)}{\mathbb{E}(K)} = \frac{\theta_A^2 + \theta_A + 2\mu f_0 T}{\theta_A + 2\mu f_0 T} = \frac{\theta_A}{1 + \frac{T}{2N_A}} + 1 \neq 1,$$

avec $\theta_A = 4N_A f_0 \mu$.

V - Sélection naturelle

1 - Hypothèses

On suppose : - population diploïde de taille constante N

- fécondations au hasard et indépendantes
- population isolée (pas de migrations)

Coefficient de sélection s qui mesure l'aptitude des génotypes.

s tient compte de : - la probabilité de survie jusqu'à l'age adulte

- la fertilité

Coefficient de dominance h qui ajuste s pour les hétérozygotes.

Remarque s et h sont ≥ 0 ou ≤ 0 ;

s et sh sont de l'ordre du %.

Soit p la fréquence de l'allèle $\mathbf{A_1}$ pour les adultes de la génération 0.

On étudie la fréquence des génotypes à la génération suivante.
Pour l'exemple, on suppose $s = -0.05$, $h = 0$, $p = 0.9$, $N = 100$.

Génotype	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Fréquence (nais.)	p^2 (81%)	$2p(1-p)$ (18%)	$(1-p)^2$ (1%)
Aptitude	$1+2s$ (0,9)	$1+2sh$ (1)	1 (1)
Nombre (adulte)	$Np^2(1+2s)$	$2Np(1-p)(1+2sh)$	$N(1-p)^2$
	(73)	(18)	(1)
Fréquence (adulte)	$\frac{p^2\,(1+2s)}{\overline{w}}$	$\frac{2p(1-p)\,(1+2sh)}{\overline{w}}$	$\frac{(1-p)^2}{\overline{w}}$
	(79%)	(20%)	(1%)

avec $\overline{w} = p^2\,(1+2s) + 2p(1-p)\,(1+2sh) + (1-p)^2$

2 - Evolution de la fréquence de l'allèle $\mathbf{A_1}$

X_t = nombre d'allèles $\mathbf{A_1}$ portés par la génération adulte t,

(après sélection) avec $t \in \mathbb{N}$.

Alors X_{t+1} sachant $X_t = i$ suit la loi binomiale $\mathcal{B}\Big(2N, \psi_i\Big)$.

$$\mathbb{P}(\, X_{t+1} = j | X_t = i \,) \;=\; \mathrm{C}_{2N}^{j}\, {\psi_i}^{j} \left(1 - \psi_i\right)^{2N-j}$$

$$\text{avec} \quad \psi_i \;=\; \frac{\frac{i}{2N}^2 (1+2s) \;+\; \frac{i}{2N}(1-\frac{i}{2N})\,(1+2sh)}{\frac{i}{2N}^2 (1+2s) \;+\; 2\frac{i}{2N}(1-\frac{i}{2N})\,(1+2sh) \;+\; (1-\frac{i}{2N})^2}$$

$\rightsquigarrow$ $(X_t)_t$ est une chaîne de Markov de matrice de transition

$$P = (P_{ij})_{i,j} \quad \text{avec} \quad P_{ij} = \mathrm{C}_{2N}^{j}\, \psi^{j} \left(1 - \psi\right)^{2N-j}$$

La chaîne de Markov a 2 états absorbants : 0 et $2N$

x_t = fréquence de l'allèle $\mathbf{A_1}$ dans la génération adulte t,
(après sélection) avec $t \in \mathbb{N}$.

$$\text{Alors} \quad x_{t+1} = \frac{{x_t}^2\,(1+2s) + x_t(1-x_t)\,(1+2sh)}{{x_t}^2\,(1+2s) + 2x_t(1-x_t)\,(1+2sh) + (1-x_t)^2}$$

Les fréquences $x_\infty = 0$ et 1 sont les seuls points d'équilibre.

Si $s < 0 \implies \mathbf{A_1}$ s'éteint ; si $s > 0 \implies \mathbf{A_2}$ s'éteint.

Si $s = 0 \implies \mathbb{P}(\text{ fixation de l'allèle } \mathbf{A_1}\ ) = \frac{X_0}{2N}$.

3 - Approximation par une diffusion

Si l'effectif de la population est grand,
on approche la chaîne de Markov par une diffusion.

a) Moyenne et variance de l'accroissement de la fréquence

Notons M_t l'accroissement moyen de la fréquence entre t et $t+1$.

$$\begin{aligned} M_t &= \mathbb{E}\left(\frac{X_{t+1}}{2N} - \frac{X_t}{2N} \,\middle|\, X_t = i \right) \\ &= \frac{1}{2N}\, \mathbb{E}\left(X_{t+1} - i \,\middle|\, X_t = i \right) \\ &= \frac{1}{2N} \times 2N\psi_i \;-\; \frac{i}{2N} \;=\; \psi_i \;-\; \frac{i}{2N} \end{aligned}$$

X_{t+1} sachant $X_t = i$ suit la loi $\mathcal{B}(2N, \psi_i)$.

Et aussi V_t = variance de l'accroissement de la fréquence.

$$
\begin{aligned}
V_t &= Var\left(\tfrac{X_{t+1}}{2N} - \tfrac{X_t}{2N} \,\middle|\, X_t = i \right) \\
&= \tfrac{1}{4N^2}\, \mathbb{E}\left(\left(X_{t+1} - i\right)^2 \,\middle|\, X_t = i \right) - \tfrac{1}{4N^2} \left(\mathbb{E}\left(X_{t+1} - i \,\middle|\, X_t = i \right) \right)^2 \\
&= \tfrac{1}{4N^2}\, \mathbb{E}\left({X_{t+1}}^2 \,|\, X_t = i \right) + \tfrac{i^2}{4N^2} - \tfrac{i}{2N^2}\, \mathbb{E}\left(X_{t+1} \,|\, X_t = i \right) \\
&\quad - \left(\psi_i - \tfrac{i}{2N} \right)^2 \\
&= \frac{\psi_i (1 - \psi_i)}{2N}
\end{aligned}
$$

Remarque : M_t et V_t ne dépendent pas de la génération t considérée, mais seulement de i et de N.

b) Calcul sur des exemples

On note $x = \frac{i}{2N}$, $x \in [0,1]$.

Cas d'une population de très grande taille.

On cherche un équivalent lorsque $x \to 0$; on les note $M_{\delta x}$ et $V_{\delta x}$.

Rappel : $M_t = \psi_i - x$ et $V_t = \frac{\psi_i(1-\psi_i)}{2N}$.

- ni mutation, ni sélection : $\psi_i = x \implies M_{\delta x} = 0$ et $V_{\delta x} = \frac{x(1-x)}{2N}$
- sélection, dominance : $\psi_i = \frac{x^2\,(1+2s) + x(1-x)\,(1+2sh)}{x^2\,(1+2s) + 2x(1-x)\,(1+2sh) + (1-x)^2}$

  $\implies M_{\delta x} \approx 2sx(1-x)\big((1-2h)x + h\big)$ et $V_{\delta x} \approx \frac{x(1-x)}{2N}$
- mutation réciproque $\mathbf{A_1} \xrightarrow{\mu} \mathbf{A_2}$ et $\mathbf{A_2} \xrightarrow{\nu} \mathbf{A_1}$:

  $\psi_i = (1-x)\nu + x(1-\mu)$

  $\implies M_{\delta x} = (1-x)\nu - x\mu$ et $V_{\delta x} \approx \frac{x(1-x)}{2N}$

c) Introduction d'une EDS

Exemple : évolution dynamique d'une population isolée.

χ_t = effectif de la population à l'instant t.

- $\chi_t = \chi_0 + \int_0^t \rho\,\chi_r \left(1 - \frac{\chi_r}{k}\right) dr.$

$\rightsquigarrow \chi_t = \chi_0 + \int_0^t m(\chi_r)\,dr$, avec $m(\mathrm{x}) = \rho\,\mathrm{x}\left(1 - \frac{\mathrm{x}}{k}\right)$.

- Modélisation de l'aléa par le mouvement brownien $(W_t)_{t\geq 0}$.

EDS : $\boldsymbol{\chi_t = \chi_0 + \int_0^t m(\chi_r)\,dr + \int_0^t \sigma(\chi_r)\,dW_r}$

Remarque : χ_t est une variable aléatoire !

Equation Différentielle Stochastique (EDS)

$$\boldsymbol{\chi_t = \chi_0 + \int_0^t m(\chi_r)\, dr + \int_0^t \sigma(\chi_r)\, dW_r}$$

D'après le calcul stochastique,

$$m(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}(\chi_{r+\epsilon} - \chi_r \,|\, \chi_r = x) = \text{accroissement infinitésimal}$$

$$\sigma^2(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}\Big((\chi_{r+\epsilon} - \chi_r)^2 \,\Big|\, \chi_r = x \Big)$$

On mesure le temps par échelle de $2N$ générations.

Ici, $\chi_r = \frac{X_{E(2Nr)}}{2N}$ la fréquence à la génération $E(2Nr)$, et $\epsilon = \frac{1}{2N}$.

donc $m(x) = \lim_{N \to +\infty} 2N\, M_{\delta x}$ et $\sigma^2(x) = \lim_{N \to +\infty} 2N \left(V_{\delta x} + {M_{\delta x}}^2 \right)$

Rappel : $m(x) = \lim_{N\to+\infty} 2NM_{\delta x}$ et $\sigma^2(x) = \lim_{N\to+\infty} 2N\left(V_{\delta x} + {M_{\delta x}}^2\right)$

- ni mutation, ni sélection : $M_{\delta x} = 0$ et $V_{\delta x} = \frac{x(1-x)}{2N}$

$$\Longrightarrow \quad \chi_t = \chi_0 + \int_0^t \sqrt{\chi_r(1-\chi_r)}\, dW_r$$

- sélection, dominance : $M_{\delta x} \approx 2sx\,(1-x)\big((1-2h)x + h\big)$

pour $h = \frac{1}{2}$, on a $M_{\delta x} \approx sx(1-x)$ et $V_{\delta x} \approx \frac{x(1-x)}{2N}$

$$\Longrightarrow \quad \chi_t = \chi_0 + \int_0^t \alpha\,\chi_r\,(1-\chi_r)\,dr + \int_0^t \sqrt{\chi_r(1-\chi_r)}\, dW_r$$

Le coefficient de sélection s est de l'ordre de $\frac{1}{N}$.

On pose $\alpha = 2sN$; s et s^2N sont négligeables.

- mutation réciproque $\mathbf{A_1} \xrightarrow{\mu} \mathbf{A_2}$ et $\mathbf{A_2} \xrightarrow{\nu} \mathbf{A_1}$:

$M_{\delta x} = \nu(1-x) - \mu x$ et $V_{\delta x} \approx \frac{x(1-x)}{2N}$

$$\Longrightarrow \quad \chi_t \;=\; \chi_0 + \int_0^t \big(\,\gamma_1(1-\chi_r) - \gamma_2\chi_r\,\big)\,dr + \int_0^t \sqrt{\chi_r(1-\chi_r)}\,dW_r$$

Les coefficients de mutation μ et ν sont de l'ordre de $\frac{1}{N}$.

On pose $\gamma_1 = \nu N$ et $\gamma_2 = \nu N$;

$2N\,\mu^2$, $2N\,\nu^2$ et $2N\,\mu\,\nu$ sont négligeables.

4 - Lien avec les EDP

a) Formule de Feynman-Kac

Rappel : $\chi_r = \frac{X_{E(2Nr)}}{2N}$ la fréquence à la génération $E(2Nr)$

$$\boldsymbol{\chi_t = \chi_0 + \int_0^t m(\chi_r)\, dr + \int_0^t \sigma(\chi_r)\, dW_r}, \quad (0 \leq \chi_t \leq 1).$$

On note $\tau_x = \inf \{t \geq 0 \,|\, \chi_t = 0 \text{ ou } 1 \text{ sachant que } \chi_0 = x\}$

- $\chi_{\tau_x} = 0$ ou $1 \quad \Longrightarrow \mathbb{E}(\chi_{\tau_x}) = \mathbb{P}(\chi_{\tau_x} = 1)$, noté $u(x)$

est solution de $\begin{cases} m(x)\, u'(x) + \frac{1}{2}\sigma^2(x)\, u''(x) = 0\,, \\ u(0) = 0 \text{ et } u(1) = 1 \end{cases}$

- $\mathbb{E}(\tau_x)$, noté $v(x)$ est solution de

$\begin{cases} m(x)\, v'(x) + \frac{1}{2}\sigma^2(x)\, v''(x) = -1 \\ v(0) = 0 \text{ et } v(1) = 0 \end{cases}$

b) Probabilité de fixation

$\rightsquigarrow$ $\mathbb{P}(\,\chi_s$ atteigne la valeur 1 avant la valeur $0 \,|\, \chi_0 = x\,) \;=\; u(x)$

$$\text{solution de } \begin{cases} m(x)\,u'(x) + \dfrac{1}{2}\,\sigma^2(x)\,u''(x) \;=\; 0\,, & \text{si } 0 < x < 1 \\ u(0) \;=\; 0 \;\text{ et }\; u(1) \;=\; 1 \end{cases}$$

$$\Longrightarrow \quad u(x) \;=\; \frac{\int_0^x e^{-2\int \frac{m(z)}{\sigma^2(z)}\,dz}\,dx}{\int_0^1 e^{-2\int \frac{m(z)}{\sigma^2(z)}\,dz}\,dx}$$

- ni mutation, ni sélection : $m(x) = 0$ et $\sigma^2(x) = x(1-x) \Longrightarrow u(x) = x$

$\rightsquigarrow \mathbb{P}(\,$ fixation allèle $\mathbf{A_1} \,|\, X_0 = x\,) = x$ et $\mathbb{P}(\,$ fixation mutation $\,) = \frac{1}{2N}$

- sélection, dominance $(h = \frac{1}{2})$: $m(x) \approx \alpha x(1-x)$ et $\sigma^2(x) \approx x(1-x)$

$$\Longrightarrow \quad u(x) \;=\; \tfrac{1-e^{-2\alpha x}}{1-e^{-2\alpha}}, \qquad \text{avec } \alpha = 2Ns.$$

$\rightsquigarrow \mathbb{P}(\,$ fixation mutation $\,) = \frac{1-e^{-2s}}{1-e^{-4Ns}} \approx \frac{2s}{1-e^{-4Ns}}$

Importance de la sélection : $N = 10^5$, $x_1 = \frac{1}{2}$ et $x_2 = \frac{1}{2N}$.

Génotype	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Aptitude	$1+2s$	$1+2sh$	1

Absence de sélection ($s = 0$)

$$u\left(\tfrac{1}{2}\right) = 0,5$$
$$u\left(\tfrac{1}{2N}\right) = 5 \times 10^{-6}$$

Sélection faible ($s = 10^{-5}$)

$$u\left(\tfrac{1}{2}\right) = \frac{1-e^{-2}}{1-e^{-4}} \simeq 0,88$$
$$u\left(\tfrac{1}{2N}\right) \simeq 2 \times 10^{-5}$$

Ainsi, même si s est faible, l'influence sur la probabilité de fixation d'un allèle ou d'une mutation est grande
(car à l'échelle de nombreuses générations).

c) Temps de fixation

$\rightsquigarrow$ $\mathbb{E}\big(\text{ temps d'atteinte d'une des valeurs 0 ou 1} \,\big|\, \chi_0 = x\,\big) \;=\; v(x)$

solution de $\left\{ \begin{array}{l} m(x)\, v'(x) + \dfrac{1}{2}\,\sigma^2(x)\, v''(x) \;=\; -1\,, \quad \text{si } 0 < x < 1 \\ v(0) \;=\; 0 \quad \text{et} \quad v(1) \;=\; 0 \end{array} \right.$

- ni mutation, ni sélection : $m(x) = 0$ et $\sigma^2(x) = x(1-x)$

$\Longrightarrow \; v(x) \;=\; -2\,\big(\, x\,\ln x + (1-x)\,\ln(1-x)\,\big) \geq 0$ car $0 < x < 1$.

$\rightsquigarrow$ $\mathbb{E}(\,$ temps de fixation d'un des allèles $\mathbf{A_1}$ ou $\mathbf{A_2}$ $\,|\, X_0 = x\,)$

$= -\,4N\,\big(x\,\ln x + (1-x)\,\ln(1-x)\big)$ générations.

Si $x = \frac{1}{2}$, on obtient le temps maximal d'attente :
temps moyen de fixation d'un des allèles $\simeq 2,8N$ générations.

d) Evolution de la fréquence en fonction du temps

Notons $\phi(t,x) = \mathbb{P}\left(\frac{X_t}{2N} \leq x \,|\, \frac{X_0}{2N} = p \right)$

ϕ est solution de l'équation de Kolmogorov backward

$$\frac{\partial \phi}{\partial t}(t,x) - \mu(x) \frac{\partial \phi}{\partial x}(t,x) - \frac{1}{2}\sigma^2(x) \frac{\partial^2 \phi}{\partial x^2}(t,x) = 0$$

e) Conclusion sur la sélection

théorie de l'horloge moléculaire

$\rightsquigarrow$ détection de mutations sélectionnées

K = nombre de mutations neutres sélectionnées apparues en t générations et susceptibles d'être fixées.

Alors, $\mathbb{E}(K) = 2N\mu f_0 t \ \frac{2s}{1-e^{-2Ns}}$

On a supposé $s = \text{cste}$,

on peut supposer que s suit une loi de densité f donnée (exponentielle, normale...)

Dans ce cas, $\mathbb{E}(K) = 2N\mu f_0 t \int_{-\infty}^{+\infty} \frac{2s}{1-e^{-2Ns}} f(s)\, ds.$