INTRODUCTION A LA COALESCENCE

–

Cours introductif en français à la théorie de la coalescence

–

Guillaume Achaz

Janvier 2016 (version 2.2 –qqes nouvelles erreurs fixées–)

1 Introduction

COALESCENCE subst. fém.
(source *http ://www.lexilogos.com/francais_langue_dictionnaires.htm*)

1. **Biologie**. Réunion normale ou pathologique de tissus voisins.
2. **Physique-chimie**. Réunion de particules en suspension.
3. **Linguistique**. Aspect de la contraction, qui consiste dans la fusion de deux voyelles voisines en une voyelle nouvelle : ae > e, au > o. Diphtongue par coalescence.

En **génétique des populations**, on fait référence à la coalescence pour décrire la réunion des lignées phylogénétiques de séquences orthologues dans une population. Comme une généalogie de séquences est décrite dans un arbre, les événements de coalescence font référence aux noeuds des arbres. Ces événements de coalescence sont la représentation d'un ancêtre commun entre deux séquences dans le passé.

Imaginons que nous ayons à notre disposition un échantillon de quelques séquences orthologues d'un ou plusieurs loci provenant de plusieurs individus d'une même espèce. La généalogie de ces loci est appelée arbre de coalescence. En théorie de la coalescence, on s'intéresse surtout aux généalogies au sein d'une population, mais on verra que l'on peut étendre les résultats à des généalogies contenant plusieurs espèces. On ne cherche pas nécessairement à reconstruire l'arbre "vrai" de ces séquences, mais à comprendre à travers leur généalogie les forces sélectives qui les font évoluer.

1.1 Quelques questions

Voici quelques questions auxquelles nous allons essayer de répondre au cours de ce cours introductif. Ces questions guideront les choix des aspects présentés et développés par la suite. De plus, elles illustrent comment la théorie de la coalescence peut apporter des outils pour résoudre des problèmes concrets rencontrés par des biologistes.

Quelle diversité dans ma population ?
On peut définir la *diversité* d'une espèce/population comme le nombre moyen de différences entre plusieurs séquences orthologues d'un même locus. A quoi doit-on comparer la diversité observée pour une espèce donnée ? Quelles informations clefs sont contenues dans cette diversité ? Quels sont les paramètres qui peuvent faire varier cette diversité ?

Sous quel régime évolue ma séquence ?
Comment peut-on, à partir d'un échantillon de séquences, tester si les polymorphismes observés évoluent par simple dérive neutre ou si l'on doit envisager des scénarii plus complexes pour expliquer leur évolution ? (Sélection, structuration géographique, etc.)

Comment expliquer certaines incohérences phylogénétiques ?
Attend-on les même arbres phylogénétiques pour plusieurs gènes orthologues entre plusieurs génomes ? Quelles sont les incertitudes attendues pour chaque noeud de l'arbre phylogénétique ? Si plusieurs gènes orthologues de génomes proches ne présentent pas le même arbre phylogénétique, doit-on nécéssairement conclure à des tansferts horizontaux ?

1.2 Coalescence *versus* phylogénie moléculaire

Ce chapitre a pour but d'illustrer les ressemblances et les différences qui existent entre deux aspects de la généalogie des séquences : la "phylogénie moléculaire" et la "théorie de la coalescence". En effet, bien qu'il existe de nombreux aspects communs entre les deux disciplines, il existe aussi des différences importantes. Aussi bien en phylogénie moléculaire qu'en coalescence, on s'intéresse à la généalogie des séquences (*i.e.* locus, gènes), néanmoins :

1. En phylogénie moléculaire, on s'intéresse surtout à la généalogie de loci homologues entre les espèces. En coalescence, on s'intéresse, *a contrario*, à la généalogie de ces loci au sein de chaque espèce.
2. En phylogénie moléculaire, on cherche surtout à reconstruire l'arbre "vrai" de loci homologues. En coalescence, on ne cherche pas à reconstruire l'arbre "vrai" ; on cherche les forces évolutives (décrites en génétique des populations) qui sont les plus compatibles avec la généalogie observée des séquences étudiées.

1.3 Esprit de l'approche "coalescence"

En coalescence, on s'interesse à déterminer l'ensemble des arbres généalogiques et/ou phylogénétiques qu'il est possible d'observer pour un modèle de génétique des populations fixé. A l'inverse, si on connaît tous les arbres possibles pour un modèle donné, on peut trouver l'ensemble des paramètres de ce modèle qui sont le plus compatible avec l'arbre observé. Les paramètres sont typiquement des valeurs pertinentes en génétique des populations (taille de la population, taux de migration, taux de croissance de la population, force de la sélection etc.). En recherchant le jeu de paramètres le plus "compatible" avec l'arbre généalogique que l'on observe, on pourra inférer l'importance des différentes forces évolutives qui sont "responsables" de l'arbre observé.

1.4 Nécessité de la formalisation probabilistique

Les notions d'arbres "possibles", paramètres "compatibles" sont formalisable (et doivent l'être pour être utilisable). Pour ce formalisme, on fait appel à des notions de probabilités. L'approche coalescence repose sur des notions plus ou moins simples de probabilité. On ne cherche pas si un arbre est "possible", mais quelle est la probabilité associée à cet arbre sous un certain modèle. Dans ce cours d'introduction, nous utiliserons surtout des notions simples de probabilités basées sur la loi géométrique (et par extension sur la loi exponentielle) et la loi de Poisson.

Quelques définitions clefs pour comprendre la suite.

1. Si $P(x)$ est une fonction de probabilité associée à la variable aléatoire x, alors on a nécessairement $\sum_x P(x) = 1$. La somme des probabilités associées à l'ensemble des valeurs de x est 1.
2. On définit la moyenne de $P(x)$ comme $E[x] = \sum_x xP(x)$. La moyenne est donc l'ensemble des valeurs de x pondéré par leur probabilité.
3. On définit la variance de $P(x)$ comme $Var[x] = \sum_x (x - E[x])^2 P(x)$. La variance mesure la dispersion des valeurs de x autour de la moyenne $E[x]$. On peut montrer assez simplement que la variance s'exprime aussi comme $Var(x) = E[x^2] - E[x]^2$.
4. Comme la variance est la somme des différences au carré, on définit également l'écart type d'une fonction comme $stdev[x] = \sqrt{Var[x]}$, qui mesure la dispersion des valeurs sur la même échelle que la moyenne.

2 Ancêtre commun et coalescence

Le terme "coalescence" fait référence aux lignées des gènes échantillonnés qui fusionnent dans le passé. Une coalescence de deux lignées est donc dû à l'existence d'un ancêtre commun entre ces deux séquences. Sachant que toutes les séquences actuelles descendent d'un séquence ancêtre, les lignées de ces séquences doivent coalescer à un moment dans le passé. Bien souvent, on ne doit pas remonter à LUCA (Last Universal Common Ancestor) pour observer des événements de coalescence. En fait, on recherche donc à estimer combien il faut remonter de générations dans le passé pour observer les évènements de coalescence (qui forment les noeuds de l'arbre). La coalescence est donc décrite dans des modèles idéalisés de génétique des populations. Un de ces modèles est celui de Wright-Fisher.

2.1 Le modèle de Wright-Fisher

Le modèle le plus "simple" en génétique des populations est vraisemblablement celui développé par Fisher (1929 et 1930) et Wright (1931). Dans ce modèle (voir figure 1), la taille de la population (N ou 2N pour les diploides) est constante et les générations ne sont pas chevauchantes. Au contraire, à chaque génération, tous les individus meurent et sont remplacés par de nouveaux individus dont le génotype est tiré aléatoirement (avec remise) parmi les individus de la génération précédente. La justification est qu'il existe un pool de gamète "infini" à chaque génération et que seule la taille de la population (*i.e.* contraintes environnementales) dicte le nombre de descendants.

Dans la suite de ce cours d'introduction, nous ne ferons référence qu'à ce modèle, mais il est nécessaire de mentionner qu'il en existe plusieurs autres et notamment le modèle de Moran (1958) où à chaque génération un seul individu meure et est remplacé. On démontre aisément que la coalescence est quasi-identique dans ce second modèle. D'autres modèles plus complexes existent également où la coalescence est similaire.

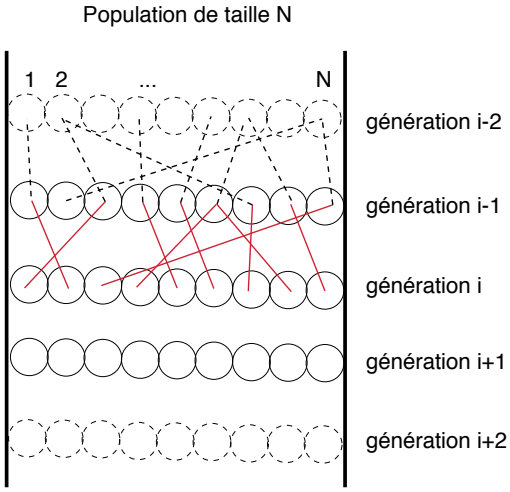

FIGURE 1 – Modèle de Wright-Fisher

2.2 Coalescence de 2 lignées

Dans le cas le plus simple, on recherche combien il faut remonter de générations en arrière dans le temps pour trouver un ancêtre commun à deux séquences d'une population idéalisée dans le modèle de Wright-Fisher.

2.2.1 Le cas des générations discrètes

Pour deux séquences portées par deux individus différents (deux lignées), on peut à la génération précédente observer un événement de coalescence ou non. La probabilité d'observer un tel événement est de $1/N$. Notons qu'une population de séquences diploïdes est de taille $2N$, car N est le nombre d'individus de la population. Tous les résultats qui vont suivre sont démontrés pour une population haploïde mais, en remplaçant N par $2N$, on obtient les résultats pour les diploïdes.

Ainsi donc, les deux séquences ont une probabilité $1/N$ de coalescer à la génération précédente et $(1 - 1/N)$ de ne pas coalescer. Sachant que $1/N$ est une petite valeur (car bien souvent la taille de la population est grande, $N \gg 1$), il y a une forte probabilité pour que les deux lignées ne coalescent pas. Dans ce cas, il existe une nouvelle chance de coalescer une génération encore auparavant (deux générations dans le passé par rapport à la génération présente). Comme chaque génération est indépendante l'une de l'autre, on peut calculer que la probabilité d'avoir un ancêtre commun deux générations dans le passé est $(1 - 1/N) \times 1/N$ et celle d'avoir un ancêtre commun ni à la première ni à la seconde est $(1 - 1/N)^2$.

Plus généralement, la probabilité d'avoir un ancêtre t générations auparavant (ni avant ni après) est de :

$$P(t_2) = 1/N \times (1 - 1/N)^{t-1} \tag{1}$$

Cette distribution de probabilité est bien connue, il s'agit d'une distribution géométrique qui, sous sa forme générale est $P(n) = p \times (1-p)^{n-1}$. On peut montrer que la moyenne de cette distribution est $E[n] = 1/p$ et sa variance $V[n] = (1-p)/p^2$. Intuitivement, on comprend que si il y a 1/6 de chance de sortir un 4 au dé, on attend en moyenne 6 coups de dé pour faire le premier 4.

Appliqué à l'équation (1), on s'attend, en moyenne, à remonter $E[t_2] = N$ générations ($E[t_2] = 2N$ pour les diploïdes) pour trouver le premier ancêtre commun à deux loci pris au hasard dans la population. Il faut tout de suite mentionner que la variance associée à cette moyenne est très grande. Elle est de $Var(t_2) = N(N-1)$, c'est à dire environ de N^2 (car $N \gg 1$).

2.2.2 Approximation au cas continu

Par la suite, on considére qu'il faut remonter toujours beaucoup de générations en arrière pour trouver l'ancêtre commun à deux ou plusieurs loci ($t \gg 1$). Ainsi, on approxime le nombre de générations à une variable continue et non plus discrète. Pour cela, on utilise l'approximation : $(1-x)^y \simeq e^{-xt}$, (quand $x \ll 1$). Ainsi, si t est le nombre de générations avant la première coalescence, on peut réarranger l'équation (1) en :

$$f_{t_2}(t) = 1/N \times e^{-t/N} \tag{2}$$

Ici, on approxime donc la loi géométrique par une loi continue de probabilité, la loi exponentielle. Cette loi ($f(x) = \lambda e^{-\lambda x}$) a pour moyenne E[x]=$1/\lambda$ et pour variance $Var[x] = 1/\lambda^2$. Appliqué au cas particulier de l'équation (2), on estime que $E[t_2] = N$ et $Var[t_2] = N^2$, c'est à dire les même résultats que pour le cas discret (ouf!). L'intérêt principal de cette aproximation est que la manipulation de l'équation (2) est bien plus simple que celle de l'équation (1).

2.3 Coalescence de 3, puis n lignées

Si l'on cherche l'ancêtre commun de trois individus (et non plus de 2), il faut considérer de nouvelles possibilités. A chaque génération, on considère trois cas : $(3 \to 1)$ les trois lignées coalescent dans un même individu, $(3 \to 2)$ deux des trois lignées coalescent (réduction de 3 à 2 lignées) et enfin $(3 \to 3)$ aucune lignée ne coalesce.

A l'aide de la figure 2, on calcule aisément leurs probabilités respectives ($P_{3\to1} = 1/N^2$, $P_{3\to2} = 1/N \times (1-1/N) + (1-1/N) \times 2/N$ et $P_{3\to3} = (1-1/N) \times (1-2/N)$ (on peut vérifier que la somme des trois probabilités est bien de 1). On note également que les probabilités associées aux trois cas peuvent se réécrire comme des polynômes du second

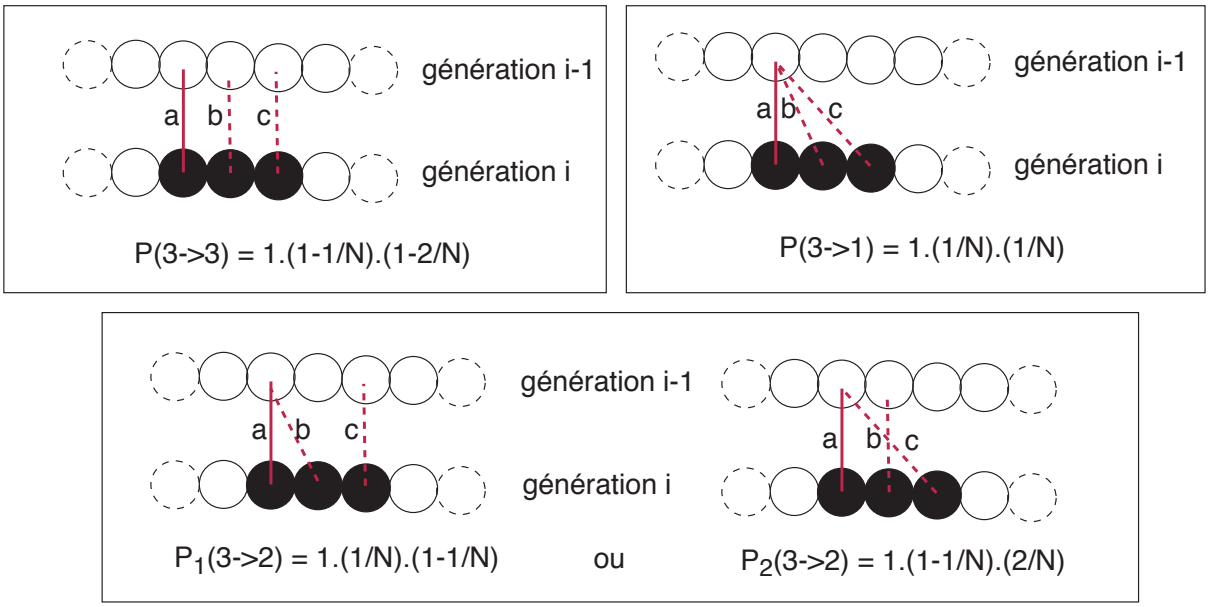

FIGURE 2 – Coalescence de 3 lignées dans une population de taille N

degré $(ax^2 + bx + c)$. De plus comme on a N très grand $(N \gg 1)$, on négligera les termes en $1/N^2$ devant les termes en $1/N$.

Ainsi, les trois probabilités deviennent :

$$3 \text{ lignées } \begin{cases} P_{3\to 1} = 1/N^2 & \approx 0 \\ P_{3\to 2} = 3/N - 3/N^2 & \approx 3/N \\ P_{3\to 3} = 1 - 3/N + 2/N^2 & \approx 1 - 3/N \end{cases}$$

En clair, dans le cas de trois lignées indépendantes, on ne considère que deux types d'événements, la coalescence de deux lignées ($3 \to 2$) ou l'absence de coalescence ($3 \to 3$) avec les probabilités décrites ci-dessus. On peut facilement extrapoler de 3 à i lignées en ne considérant que les événements impliquant 1 seul événement de coalescence pour une génération donnée. On considère que les événements impliquant i coalescences simultanées ont une probabilité faible (en $O(1/N^i)$) et sont négligeables devant les événements n'impliquant qu'une seule coalescence. On ne retient donc que deux possibilités par génération : les événements ($i \to i$) et les événements ($i \to i-1$).

On peut calculer la probabilité d'un événement de type $(i \to i-1)$ comme :

$$\begin{aligned} P_{i\to i-1} &= \frac{1}{N}(1-\frac{2}{N})\cdots(1-\frac{i-1}{N}) + (1-\frac{1}{N})\frac{2}{N}\cdots(1-\frac{i-1}{N}) + \\ &\quad \cdots + (1-\frac{1}{N})(1-\frac{2}{N})\cdots\frac{i-1}{N} \\ &= \frac{1}{N}[1+2+\cdots+(i-1)] + O(\frac{1}{N^2}) \\ &\approx \frac{1}{N}\times\frac{i(i-1)}{2} \end{aligned}$$

Comme la probabilité d'un événement $(i \to i)$ est $1 - P_{i\to i-1}$ (ce qui manque pour sommer à 1), on a $P_{i\to i-1} = \binom{i}{2}/N$ et $P_{i\to i} = 1 - \binom{i}{2}/N$. Comme dans le cas de deux lignées de l'équation 2, on peut calculer la fonction de densité de probabilité d'avoir un premier ancêtre commun t générations auparavant :

$$f_{t_i}(t) = \frac{\binom{i}{2}}{N} \times e^{-t\frac{\binom{i}{2}}{N}} \tag{3}$$

Afin de s'affranchir de la dépendance entre les temps de coalescence et la taille de population, on redéfini des temps de coalescence (T_i) exprimés en N générations $(T_i = t_i/N)$. Dans cette nouvelle échelle de temps, on peut réécrire l'équation 3 en :

$$f_{T_i}(t) = \binom{i}{2} \times e^{-t\binom{i}{2}} \tag{4}$$

Cette dernière probabilité a pour moyenne et variance :

$$\begin{aligned} E[T_i] &= \frac{2}{i(i-1)} \\ Var[T_i] &= \frac{4}{i^2(i-1)^2} \end{aligned} \tag{5}$$

En observant l'équation 5, on montre que les événements de coalescence sont d'autant plus rapides que le nombre de loci est important. C'est à dire que le temps attendu entre l'état ou l'arbre à i séquences jusqu'à celui où il a $i-1$ séquences est d'autant plus petit que i est grand. Par ailleurs, l'examen de la variance de T_i montre que la dispersion de temps de coalescence autour de la moyenne est très grande. Ceci implique qu'entre deux loci échantillonnés dans de deux populations différentes de même taille, on ne s'attend pas à trouver des T_i similaires.

2.4 Premier ancêtre commun de n lignées

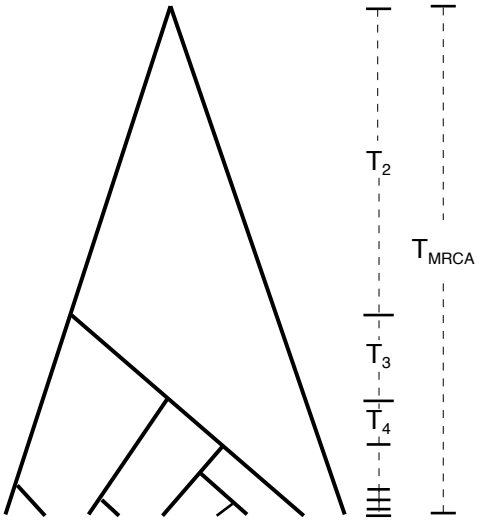

FIGURE 3 – Arbre de coalescence "moyen" illustrant les différents temps de coalescence

Nous ne cherchons plus seulement la distribution du temps de la première coalescence pour i lignées, qui représente à l'événement $(i \rightarrow i-1)$. Nous cherchons à étudier la distribution du nombre de générations qu'il faut remonter dans le temps pour trouver l'ancêtre commun de toutes les i lignées. Pour cela, on utilise deux valeurs importantes des arbres de coalescence. La première est le temps nécessaire à trouver un ancêtre commun à tous les n loci de l'échantillon (voir figure 3) : $T_{MRCA} = \sum_{i=2}^{n} T_i$. On note que *MRCA* signifie *Most Recent Common Ancestor*. La seconde est la longueur totale de toutes les branches de l'arbre de coalescence : $T_{total} = \sum_{i=2}^{n} iT_n$. Même s'il n'est pas aisé d'obtenir la distribution complète de ces deux valeurs, on peut calculer assez simplement leur moyenne et leur variance.

2.4.1 Moyennes des T_{total} et T_{MRCA}

Le calcul de la moyenne de T_{total} peut se faire en utilisant des propriétés simples des moyennes de loi de probabilités.

$$
\begin{aligned}
E[T_{total}] &= \sum_{i=2}^{n} iE[T_i] \\
&= \sum_{i=2}^{n} \frac{2}{(i-1)} \\
&= 2 \times \sum_{i=1}^{n-1} \frac{1}{i} \qquad (6)
\end{aligned}
$$

Ainsi, on montre que la moyenne de T_{total} augmente indéfiniment avec le nombre de loci échantillonnés. Cependant, il faut noter que le temps total additionné est de plus en plus petit lorsque i augmente (les branches sont nettement plus grandes pour des i petits). Pour ce qui est de la moyenne de T_{MRCA} à tout l'échantillon, on a :

$$
\begin{aligned}
E[T_{MRCA}] &= \sum_{i=2}^{n} E[T_i] \\
&= \sum_{i=2}^{n} \frac{2}{i(i-1)} \\
&= 2 \times \sum_{i=2}^{n} (\frac{1}{(i-1)} - \frac{1}{i}) \\
&= 2 \times (1 - \frac{1}{n}) \qquad (7)
\end{aligned}
$$

A l'inverse de $E[T_{total}]$ (équation 6), la moyenne de T_{MRCA} de n lignées ne grandit pas à l'infini, elle tend vers 2 lorsque $n \gg 1$. Cela signifie que si le T_{MRCA} moyen de quelques individus pris au hasard dans une population est le même (à epsilon près) que celui de la population prise dans son entier (les séquences de tous les individus de la population). Par ailleurs, on sait que $E[T_2] = 1$ (voir équation 5) ; le temps moyen nécessaire pour trouver l'ancêtre commun de 2 individus est de N générations. Ceci signifie qu'en moyenne, plus de la moitié du temps nécessaire pour trouver l'ancêtre commun est dû au dernier événement de coalescence (voir figure 3).

2.4.2 Variances des T_{total} et T_{MRCA}

On peut également estimer les variances de ces deux mesures de l'arbre de coalescence. En utilisant le fait que tous les événements de coalescence sont indépendants (les covariances sont nulles) conjointement avec le même type de notation que pour le calcul de la moyenne, on montre assez simplement que la variance de T_{total} peut se calculer comme suit :

$$
\begin{aligned}
Var[T_{total}] &= Var[2T_2 + 3T_3 + \cdots + nT_n] \\
&= \sum_{i=2}^{n} i^2 Var[T_i] + \sum_{i=2}^{n} \sum_{j \neq i} Cov[iT_i, jT_j] \\
&= \sum_{i=2}^{n} 4 \frac{1}{(i-1)^2} \\
&= 4 \times \sum_{i=1}^{n-1} \frac{1}{i^2} \qquad (8)
\end{aligned}
$$

De même, la variance du T_{MRCA} est :

$$
\begin{aligned}
Var[T_{MRCA}] &= Var[T_2 + T_3 + \cdots + T_n] \\
&= \sum_{i=2}^{n} Var[T_i] + \sum_{i=2}^{n} \sum_{j \neq i} Cov[T_i, T_j] \\
&= \sum_{i=2}^{n} 4 \times (\frac{1}{i-1} - \frac{1}{i})^2 \\
&= 8 \sum_{i=2}^{n} \frac{1}{i^2} - 4\Big(1 - \frac{1}{n}\Big)^2 \qquad (9)
\end{aligned}
$$

De ces résultats, il faut garder à l'esprit que les deux variances des temps de coalescence sont grandes (voir très grandes). Ainsi, on ne peut pas considérer l'arbre de coalescence moyen représenté sur la figure 3 comme un arbre "typique" mais plutôt la "tendance" moyenne de chaque noeud de l'arbre. Quelques simulations seront nécessaires pour vous convaincre totalement.

2.4.3 Distribution complète des T_{total} et T_{MRCA}

Pour information on peut dériver la distribution de T_{MRCA} et T_{total} par plusieurs méthodes dont l'explication dépasserait le cadre de cette introduction. Pour plus d'information se référer à des ouvrages plus complets sur le sujet tel que le livre de J Wakeley (*Coalescent theory, an introduction*), dont certains chapitres peuvent être consulter à l'adresse http ://www.roberts-publishers.com/wakeley/.

Ainsi, on peut montrer que :

$$
f_{T_{MRCA}}(t) = \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{\substack{j=2 \\ j \neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \qquad (10)
$$

$$f_{T_{total}}(t) = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} e^{-\frac{i-1}{2}t} \tag{11}$$

3 Mutations dans les arbres de coalescence

Dans la section précédente, nous nous sommes occupés des arbres de coalescence attendus dans une population théorique. Comme en phylogénie, si on ne dispose pas de sites polymorphes dans notre échantillon de séquence, nous ne pouvons faire aucune prédiction intéressante quant à l'arbre de coalescence sous-jacent à notre échantillon de séquence (et donc aucune inférence sur la population dont est issu l'échantillon). Ainsi, toute la théorie développée ci-dessus n'est pas applicable aux séquences biologiques si les mutations ne sont pas prises en compte.

3.1 Ajout des mutations dans les arbres

Dans le cadre du modèle neutre considéré pour les arbres de séquences, toutes les mutations que nous observons dans les séquences échantillonnées sont neutres (nous aborderons le cas où les mutations ne sont pas neutres plus loin). En d'autres termes, les mutations qui se produisent dans les séquences n'affectent en rien la généalogie de ces séquences. C'est à dire que le processus de mutation est indépendant du processus généalogique.

Ainsi, l'astuce utilisée en coalescence standard pour faire des prédictions sur les fréquences des polymorphismes neutres consiste à considerer les généalogies telles que nous l'avons fait précédemment et d'y ajouter des mutations (voir figure 4).

Ainsi, si on admet (*i*) que la fréquence de mutations par locus et par génération est faible, (*ii*) que le nombre de générations considéré est grand (de l'ordre de N générations, avec $N \gg 1$), le nombre de mutations attendues dans l'arbre est donné par une loi de Poisson.

3.1.1 Nombre de mutations en t générations

La loi de Poisson dérive de la loi binomiale lorsque le nombre d'essais (ici le nombre de générations) est très grand et la probabilité d'observer un succès (ici le taux de mutation) est très petite. Ainsi, on peut montrer que, sachant μ le taux de mutations par génération et par locus, le nombre k de mutations après t générations suit une distribution décrite par une loi de Poisson :

$$P(k|t) = e^{-\mu t} \frac{(\mu t)^k}{k!} \tag{12}$$

Pour une loi de Poisson $P(k) = e^{-\lambda}\lambda^k/k!$, la moyenne est égale à la variance et est égale à λ. Pour l'équation 12, on a :

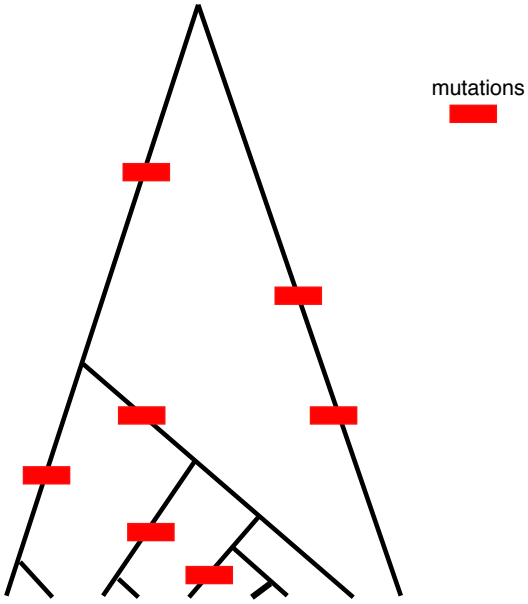

FIGURE 4 – Ajouter des mutations dans une généalogie

$$E[k|t] = Var[k|t] = \mu t \tag{13}$$

C'est à dire que le nombre moyen de mutations accumulées au cours de t générations est simplement le nombre de générations multiplié par le taux de mutation du locus.

3.1.2 En moyenne, combien de différences entre 2 loci

Comme les deux processus de mutation et de coalescence sont indépendants l'un de l'autre, on peut aisément calculer le nombre moyen de différences comme la moyenne du processus de coalescence fois le nombre moyen de mutations. On connaît le nombre moyen $E[t_2]$ de générations nécessaires pour trouver un ancêtre commun à deux loci (voir équation 2). Pour trouver le nombre de bases différentes entre deux locus, il faut prendre en compte que les mutations se sont accumulées dans les deux lignées. On a donc :

$$\begin{aligned} E(k_2) &= 2 \times E[t_2] \times \mu \\ &= 2N\mu \end{aligned} \tag{14}$$

Comme nous le verrons par la suite, ce nombre $\theta = 2N\mu$ (notons que $\theta = 4N\mu$ pour les diploïdes) est une valeur de première importance pour les processus de génétique des populations. θ est souvent assimilé à la diversité de la population, puisqu'il représente le nombre moyen de différences attendues dans deux loci échantillonnés au hasard dans la population. Comme l'on pouvait intuitivement s'en douter, cette diversité dépend du taux

de mutation (μ), mais elle dépend aussi de la taille de la population (N). Plus la taille est grande, plus on attend de diversité à l'équilibre. Intuitivement, quand la taille de la population est petite, la dérive est plus forte (perte rapide de diversité) et le nombre de mutations qui "entrent" dans la population à chaque génération est plus faible.

3.2 Mutations (suite)

Dans la partie précédente, nous avons estimé le nombre moyen de différences entre deux séquences prises au hasard dans la population. Pour cela, nous avons implicitement assumer que le nombre de mutations calculé était égal au nombre de différences. Nous avons donc implicitement opté pour le modèle dit de "sites infinis". Dans ce modèle, chaque mutation se produit à un nouveau site et le nombre de sites potentiels est infini. Cela implique que dans ce modèle, il n'existe pas de mutations doubles. C'est un des modèles les plus naïfs décrivant le processus de mutation, mais il rend les calculs plus simples à manipuler. Il n'est pas totalement absurde si le nombre de mutations est très petit devant la longueur de la séquence. Même si le nombre de sites mutés est incorrect, le nombre de mutations reste valide. Aussi, par la suite, nous garderons ce modèle de sites infinis.

3.3 Nombre de différence entre deux séquences

Si l'on prend 2 séquences aléatoirement dans une population, on attend un temps de coalescence t_2 décrite par l'équation 2. Ici on cherche à déterminer le nombre de mutations quel que soit le temps de coalescence t_2, c'est à dire que l'on veut intégrer l'équation 12 sur toutes les t_2 possibles.

Ainsi, on calcule la distribution complète de μ attendue entre 2 séquences comme suit :

$$\begin{aligned}
P_{k_2}(k) &= \int_{t=0}^{\infty} P(k|t) f_{t_2}(t) dt \\
&= \int_{t=0}^{\infty} e^{-2\mu t} \frac{(2\mu t)^k}{k!} \times \frac{1}{N} e^{-t/N} dt \\
&= (2\mu)^k \frac{1}{N} \int_{t=0}^{\infty} \frac{t^k}{k!} e^{-\frac{2N\mu+1}{N} t} dt \\
&= (2\mu)^k (\frac{N}{2N\mu+1})^{k+1} \frac{1}{N} \int_{t=0}^{\infty} \frac{t^k}{k!} (\frac{2N\mu+1}{N})^{k+1} e^{-\frac{2N\mu+1}{N} t} dt \\
&= (\frac{2N\mu}{2N\mu+1})^k \frac{1}{2N\mu+1} \\
&= (\frac{\theta}{\theta+1})^k \frac{1}{\theta+1} \qquad (15)
\end{aligned}$$

Dans le développement précédent, on utilise la loi de probabilité gamma définie comme $\Gamma(t) = \frac{t^{k-1}}{k-1!} \lambda^k e^{-\lambda t}$ pour k entier (il existe un forme plus générale pour k non-entier qu'on ne verra pas ici). La loi gamma peut s'obtenir par k convolutions successives de lois exponentielles ayant le même paramètre λ. Elle représente donc le temps qu'il faut attendre

pour que k événements de types exponentiels se produisent (par exemple, temps d'attente pour que k particules radioactives se désintègrent). Comme pour toutes les lois continues de densité de probabilités, on a $\int_{t=0}^{\infty} \Gamma(t)dt = 1$.

Comme $\frac{1}{\theta+1} \leq 1$, l'équation 15 est une loi géométrique ayant pour paramètre $p = \frac{1}{\theta+1}$. On peut donc aisément calculer sa moyenne et sa variance comme :

$$E[k_2] = \theta \quad (16)$$
$$Var[k_2] = \theta + \theta^2 \quad (17)$$

Si l'on ne considère que deux séquences, ce nombre k_2 représente le nombre moyen de différences entre les deux séquences mais aussi le nombre de sites polymorphes. Cependant, dès lors que l'on considère plus de deux séquences, il faut faire la différence entre le nombre moyen de différences entre les séquences (π) et le nombre de sites polymorphes (S). Si, par exemple, on considère trois séquences dont une seule porte un nucléotide différent, les deux mesures ne sont pas identiques ; on a $\pi = 2/3$ et $S = 1$.

3.4 Le nombre de sites polymorphes (S)

Pour calculer le nombre mutations étant apparues dans l'histoire des séquences de l'échantillon, il faut considérer la longueur totale de l'arbre de coalescence (t_{total}) et utiliser ce temps pour la loi de Poisson décrivant le nombre de mutations. Ici, nous ne calculerons que la moyenne et la variance de S. Pour cela, nous utiliserons les temps de chaque étape de la coalescence (les t_i correspondant aux étapes où il y a exactement i lignées dans l'arbre).

Dans la section précédente, nous avons utilisé la distribution de la longueur de l'arbre pour deux séquences (équation 15). Pour le cas plus général, il suffit d'opérer la même opération en utilisant la distribution de t_i (équation 3) :

$$\begin{aligned} P_{k_i}(k) &= \int_{t=0}^{\infty} P_{k_i}(k|t) f_{t_i}(t)dt \\ &= \int_{t=0}^{\infty} \frac{(i\mu t)^k}{k!} e^{-i\mu t} \times \frac{\binom{i}{2}}{N} e^{-t\frac{\binom{i}{2}}{N}} dt \\ &= \Big(\frac{iN\mu}{iN\mu + i(i-1)/2}\Big)^k \frac{i(i-1)/2}{iN\mu + i(i-1)/2} \quad (18) \\ &= \Big(\frac{\theta}{\theta + i - 1}\Big)^k \frac{i-1}{\theta + i - 1} \quad (19) \end{aligned}$$

Cette équation décrit la distribution du nombre de mutations lorsque l'arbre de coalescence a exactement i lignées. En utilisant les propriétés des lois géométriques, on montre que sa moyenne est $E[k_i] = \theta/(i-1)$ et sa variance $Var[k_i] = \theta/(i-1) + \theta^2/(i-1)^2$.

Ces résultats vont permettre de déterminer S, le nombre total de site contenu dans un arbre de coalescence ($S = k_{total}$). Ce nombre représente, sous le modèle de sites infinis,

le nombres de sites polymorphes attendus dans un échantillon de séquences. Bien que l'estimation de la distribution complète de S est délicate, les calculs de sa moyenne et sa variance sont triviaux.

En effet, comme S est simplement la somme de toutes les mutations étant apparue à toutes les étapes de l'arbre, on peut utiliser les moments de l'équation 19 pour calculer sa moyenne :

$$\begin{aligned} E[S] &= \sum_{i=2}^{n} E[k_i] \\ &= \sum_{i=2}^{n} \frac{\theta}{i-1} \\ &= \theta \sum_{i=1}^{n-1} \frac{1}{i} \end{aligned} \tag{20}$$

De même, pour sa variance on a :

$$\begin{aligned} Var[S] &= \sum_{i=2}^{n} Var[k_i] + \sum_{i=2}^{n} \sum_{j \neq i} Cov[ik_i, jk_j] \\ &= \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \end{aligned} \tag{21}$$

Comme mentionné ci-dessus, l'obtention de la distribution complète de S est complexe. Néanmoins, elle peut se faire d'au moins deux façons. Une première consiste à faire les $n-1$ convolutions successives des équations 19 avec i variant de 2 à n. Une seconde est de résoudre directement $P_S(k) = P_{k_{total}}(k) = \int_{t=0}^{\infty} P_{k_{total}}(k|t) f_{t_{total}}(t) dt$. Quelque soit la façon retenue, on peut montrer que :

$$P_S(k) = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \left(\frac{i-1}{\theta+i-1}\right) \left(\frac{\theta}{\theta+i-1}\right)^k \tag{22}$$

3.5 Le nombre moyen de différences (π)

Nous ne cherchons maintenant plus le nombre total de sites (S) qui ont muté dans l'arbre de coalescence, mais le nombre moyen de différence entre deux séquences parmi n séquences échantillonnées. Ainsi, il s'agit de déterminer parmi toutes les $\binom{n}{2}$ comparaisons deux à deux entre les n séquences de l'échantillon, combien on attend, en moyenne, de différences. Comme pour la détermination de S, on découple le processus de mutation de celui de coalescence. Le processus de mutation étant une simple loi de Poisson, il s'agira de trouver le temps moyen de coalescence entre 2 séquences parmi n échantillonnées.

Plus formellement, si on défini k_{ij}, comme le nombre de différences entre la séquence i et j et T_{ij}, le temps de coalescence (en N générations) pour que les deux séquences coalescent, on peut écrire :

$$\begin{aligned} E[\pi] &= \frac{1}{\binom{n}{2}} E\Big[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\Big] \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E[k_{ij}] \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} N\mu E[2T_{ij}] \\ &= \frac{\theta}{\binom{n}{2}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E[T_{ij}] \qquad (23) \\ &= \theta \qquad (24) \end{aligned}$$

Pour passer de l'équation 23 à l'équation 24, il est nécessaire de montrer que le temps moyen de coalescence entre deux séquences d'un arbre de coalescence contenant n séquences est de $E[T_{ij}] = 1$. Nous avons déjà montrer auparavant que si l'on ne considère que deux séquences, on attend en moyenne N générations pour trouver un ancêtre commun ($E[T_2] = 1$, voir équation 5). Ici, on cherche donc à montrer que quelque soit le nombre de séquences échantillonnées, le nombre moyen de génération reste inchangé. Autrement dit, on cherche à montrer que l'histoire généalogique de chaque paire de séquences ne dépend pas des autres séquences.

Le développement exact de la demonstration n'est pas trivial et dépasse le cadre de cette introduction. Pour plus de détails, se référer à la publication originale de Tajima (Genetics, 1983) ou bien au livre de J Wakeley (*Coalescence theory, an introduction*).

Pour la variance de π, on peut montrer que :

$$Var[\pi] = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \qquad (25)$$

4 Et dans le monde réel ?

Jusqu'ici, nous avons considéré que le monde de la coalescence était un monde théorique habité par des populations idéalisées de Wright-Fisher. Cependant, en tant que biologiste, nous ne somme pas dupes ! Les populations réelles ne sont pas celles de Wright-Fisher. Aussi, par la suite, nous tacherons de montrer comment on peut tenter de mettre en adéquation ou en inadéquation les populations réelles avec celles de Wright-Fisher.

4.1 Test de neutralité (Tajima's D)

Précédemment, nous avons montré que aussi bien le nombre site polymorphes (S) que le nombre moyen de différences entre les séquences (π) sont une fonction très simple de θ. Le corollaire de ce résultat est que l'on peut utiliser ces deux mesures pour estimer θ. Ces deux estimations seront nommées respectivement $\widehat{\theta_S}$ et $\widehat{\theta_\pi}$ et on les définit comme :

$$\widehat{\theta_S} = S/a_1, \quad \text{où } a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \tag{26}$$

$$\widehat{\theta_\pi} = \pi \tag{27}$$

Ces deux estimateurs de θ sont égaux pour une population de Wright-Fisher. Cependant, si l'évolution des séquences échantillonnées dans la nature ne peut pas être assimilé à un modèle de population neutre, panmictique et de taille constante, alors les deux estimateurs pourraient être différents. C'est dans cette idée, que Tajima (en 1989) proposa un test de neutralité basé sur ces deux estimateurs. Le test calcule une valeur D qui mesure la "déviance" par rapport à un modèle de Wright-Fisher. On définit D comme suit :

$$D = \frac{\pi - S/a_1}{\sqrt{Var(\pi - S/a_1)}} \tag{28}$$

Pour information, le dénominateur de ce test est donné par :

$$\widehat{Var}(\pi - S/a_1) = e_1 S + e_2 S(S-1)$$

ou :

$$\begin{array}{ll} e_1 = \frac{c_1}{a_1} & e_2 = \dfrac{c_2}{a_1^2 + a_2} \\ c_1 = b_1 - \frac{1}{a_1} & c_2 = b_2 - \dfrac{n+2}{n a_1} + \dfrac{a_2}{a_1^2} \\ b_1 = \frac{n+1}{3(n-1)} & b_2 = 2\dfrac{n^2+n+3}{9n(n-1)} \\ a_1 = \sum_{i=1}^{n-1} \frac{1}{i} & a_2 = \displaystyle\sum_{i=1}^{n-1} \frac{1}{i^2} \end{array}$$

Ce test mesure la différence entre π et S/a_1.

Si le nombre moyen de nucleotides différents est supérieur à celui attendu par rapport au nombre de sites polymorphes S corrigé par $a1$, on attend un $D > 0$. On attend, bien sûr, $D < 0$ dans la situation inverse.

Pour comprendre intuitivement le signe de D, on peut raisonner sur les nucléotides qui ne sont différents que dans une seule séquence de l'échantillon (les polymorphismes dits "singletons"). Chaque singleton ajoute un nouveau site polymorphe à l'échantillon

mais ne modifie que peu le nombre moyen de différences π (il l'augmente de $2/n$). Plus généralement, les polymorphismes de faible fréquence (les singletons étant le cas extrême) tendent à augmenter S sans augmenter beaucoup π. Ainsi, lorsqu'on a un excès de polymorphismes de basse fréquence, on attend un D négatif. A l'inverse, lorsque l'on a un déficit de polymorphismes de basse fréquence, on attend un D positif.

On note que puisque devant un échantillon de séquence, nous ne savons pas quelle est le nucléotide ancestral et le nucléotide muté (pour un site polymorphe), on ne ne peut pas faire la différence entre un polymorphisme de basse fréquence et un polymorphisme de haute fréquence. Ceci signifie qu'il existe pour nous que deux types de polymorphismes : ceux de basse/hautre fréquence et ceux de fréquence moyenne.

Comme nous le verrons par la suite, le $TajimaD$ peut être compatible ou incompatible avec scénarios de déviance par rapport au modèle de Wright-Fisher.

5 Déviation par rapport au modèle simple

L'effectif efficace, N_e, peut être considéré comme une taille théorique associé à une population réelle. N_e serait la taille de cette population réelle si on devait l'assimiler par une population de Wright-Fisher. Si $N = N_e$, cela signifie que notre population réelle se comporte exactement comme une population idéalisée. Bien souvent, $N_e \neq N$, ce qui signifie qu'il existe des facteurs différents entre de la population réelle et la population de Wright-Fisher. Cependant, on peut modéliser la population réelle comme une population idéale si on choisit N_e au lieu de N. Les différences entre N et N_e proviennent de violation de la population modèle. En effet, les populations modèles sont neutres, panmictiques et de taille constante. Nous disposons, en plus de la différence entre N et N_e d'une autre mesure de la déviance au modèle neutre, la valeur $TaijmaD$.

Nous allons, dans la suite, regarder intuitivement quel serait l'effet d'une violation d'une des trois hypothèses fortes du modèle de Wright-Fisher : (*i*) population de taille constante, (*ii*) population panmictique et (*iii*) absence de sélection.

5.1 Taille variable

5.1.1 Population croissante/décroissante

Pour traiter une population croissante (ou décroissante), on doit considérer que la taille de la population varie à chaque génération. Pour cela, on peut étudier (par exemple) un modèle où la taille de la population varie en croissance/décroissance exponentielle. Pour cela, on admet que la taille $N(t) = e^{ct}$.

Pour cela, on utilise une astuce qui consiste à opérer un changement d'échelle de temps (comme nous avions fait pour compter le temps en N génération de t vers T, voir équation 1 et 2). Ici on définit un nouveau temps, τ qui est proportionnel à la différence de taille entre la taille au moment de l'échantillonnage N_0 et la taille t générations auparavant $N(t)$.

Ainsi, on peut écrit :

$$
\begin{aligned}
d\tau &= N_0/N(t), \quad \text{et donc} \\
\tau &= \int_0^t d\tau \\
&= \int_0^t \frac{N_0}{N_t} \\
&= \frac{1}{c}(1 - e^{-ct})
\end{aligned}
$$

Dans cette nouvelle échelle de temps τ, le processus de coalescence peut être traité comme le processus standard. Simplement le temps T est remplacé par le temps τ dans les équations de coalescence. Ainsi, on peut traiter le problème analytiquement.

Intuitivement, on peut comprendre l'effet de la croissance et de la décroissance de la population sur le $TajimaD$.

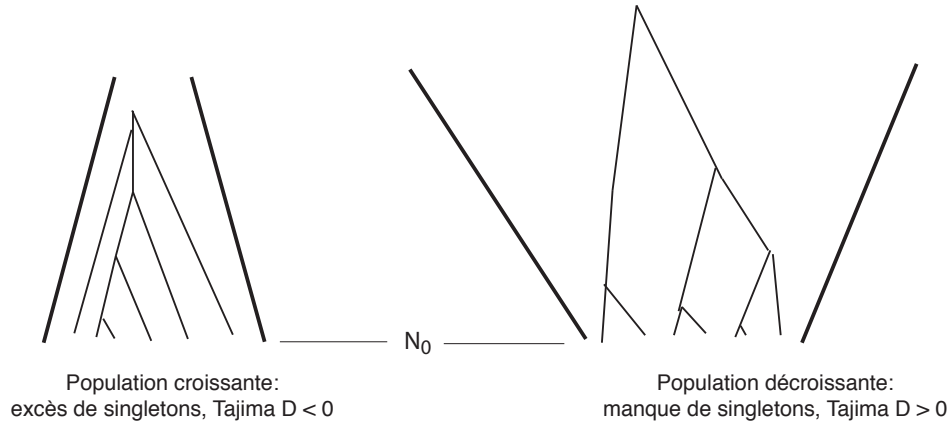

FIGURE 5 – Croissance et Décroissance de populations

Si la population est en croissance alors les événements de coalescence récent vont avoir lieu plus lentement que les événements de coalescence antérieurs. Ainsi, les branches internes de l'arbre de coalescence vont être raccourcies et les branches externes rallongées (voir figure 5). Les branches externes portent les mutations à basse fréquence (voir même les singletons) et les branches internes les mutations à fréquence moyenne. Ainsi, si la population est en croissance, on attend un excès de mutation à basse fréquence et donc un $TajimaD$ négatif (voir équation 28). A l'inverse, si la population est en décroissance, on attend un $TajimaD$ positif.

Il faut également mentionner que la taille total de l'arbre est modifiée. En effet, dans une population en croissance, l'arbre est "trop petit" par rapport à la taille observable lors de

l'échantillonnage. Dans ce cas, en plus de l'effet sur le $TajimaD$, on attend globalement une diversité appauvrie par rapport à la taille N_0 de la population. Ainsi, comme nous venons de le voir dans la section précédente, l'estimation de N_e donnera typiquement une taille efficace bien plus petite que la taille réelle de la population au moment de l'échantillonnage N_0. A l'inverse, pour une population en décroissance, on attend un excès de diversité pour le N_0 (et donc un $N_e > N_0$).

5.1.2 Population de taille fluctuante

Dans cette section, nous allons considérer une population dont la taille n'est pas constante mais peut fluctuer à chaque génération. C'est typiquement le cas des "goulots d'étranglements" qui sont souvent associé à l'un des moteurs de la diversification des espèces. Des résultats classiques de génétique des population montrent que l'on peut calculer N_e comme la moyenne harmonique des différentes tailles que prend la population à chaque génération :

$$\frac{1}{N_e} = \sum_{i}^{gen.} \frac{1}{N_i} P(i)$$

Dans cette moyenne harmonique, N_e est bien plus proche des faibles valeurs de N. Dans un état d'esprit de coalescence, on est intuitivement satisfaisant par ce résultat qui montre que lors des générations où la population est de taille faible, les événements de coalescence sont beaucoup plus fréquents et donc globalement, l'arbre de coalescence est plus petit que si la taille était à la stable à la moyenne.

Dans un exemple plus formel, on imagine une population ayant 2 tailles différentes (notées de N_1 et N_2). On fait également l'hypothèse qu'une fraction p des générations se fait alors que la population à une taille N_1. Si on considère deux séquences dans cette population, alors le taux de coalescence est donné par :

$$\begin{aligned}
P(t_2) &= p\Big[\frac{1}{N_1}(1-\frac{1}{N_1})^{pt}(1-\frac{1}{N_2})^{(1-p)t}\Big] + (1-p)\Big[\frac{1}{N_2}(1-\frac{1}{N_1})^{pt}(1-\frac{1}{N_2})^{(1-p)t}\Big] \\
&= \Big[p\frac{1}{N_1} + (1-p)\frac{1}{N_2}\Big]\Big[e^{-\frac{pt}{N_1}} e^{-\frac{(1-p)t}{N_2}}\Big] \\
&= \Big[p\frac{1}{N_1} + (1-p)\frac{1}{N_2}\Big] e^{-t\left(p\frac{1}{N_1}+(1-p)\frac{1}{N_2}\right)} \\
&= \frac{1}{N_e} e^{-\frac{t}{N_e}}
\end{aligned}$$

Bien que nous ne le ferons pas, cet exemple est aisément généralisable à plusieurs tailles différentes. Ce qu'il faut retenir, c'est que les tailles les plus plus petites ont les poids le plus importants dans les processus de coalescence.

Quant à $TajimaD$, si les changement de taille sont suffisamment fréquent, le spectre de fréquences de polymorphismes n'est pas affecté. Si, à l'inverse, les changements de taille sont rares, le $TajimaD$ peut être affecté et le sens de son affectation dépend du ratio entre

taille actuelle et la taille ancestrale (plus petite ou plus grande) et du temps écoulé depuis le dernier changement de taille. Ces cas, qui ressemblent à la croissance/décroissance (tant au niveau de la diversité qu'au niveau de $TajimaD$, sont illustrés dans la figure 6.

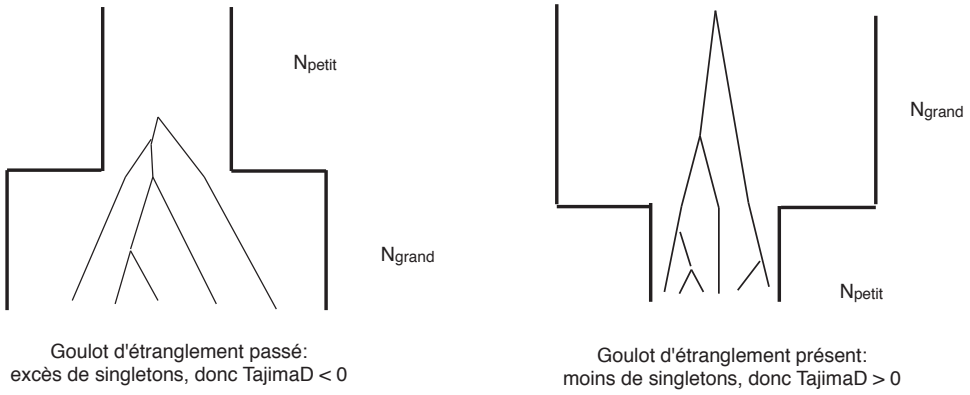

FIGURE 6 – TajimaD pour des goulots d'étranglements

5.2 Structuration

Si la population est divisée en plusieurs sous-populations, alors on dit que la population globale est structurée. Dans un modèle typique de population structurée, les individus appartenant à la même sous-populations sont panmictiques mais n'ont que des échanges limités avec les autres sous-population. On appelle migrants les individus bougeant d'une population à une autre.

Dans le cas le plus simple, on imagine que la population est sous-divisée deux sous-populations de taille égale ($N/2$) et qu'il existe une fraction m d'individus pouvant changer de sous-population. Dans ce cas, les taux de coalescence au sein de chaque population sont dictée par $N/2$ (donc plus rapide que la coalescence). Le taux de migration par génération est souvent défini comme m. Si le taux de migration est très petit, il ralenti considérablement le temps de coalescence pour deux séquences dans les deux sous-populations différentes. Selon la repartition du nombre de séquences de l'échantillon dans les deux sous-populations, les déformations de l'arbre sont différentes. Nous laisserons donc ce thème pour un futur développement.

5.3 Sélection

La coalescence est par essence défini dans des population neutre : les mutations observées et/ou liée à celle que l'on observe ne changent pas la généalogie des séquences.

Incorporé la sélection dans les modèles de coalescence est toujours délicat et requiert souvent des méthodes compliquées qui ne seront pas abordées ici. Nous étudierons plutôt deux cas "simplistes" de sélection que l'on peut comprendre en utilisant notre intuition.

5.3.1 Sélection négative

La sélection négative est celle qui tend à enlever les mutations délétères des populations. Il a été montré par une approche de diffusion (voir les travaux de Kimura et ses collaborateurs) la variation de fréquence d'une mutation délétère dans une population de taille finie est dictée par le rapport entre la taille de la population N et le facteur de sélection associé à cette mutation s. Grossièrement si $Ns \ll 1$, la mutation délétère est considérée (à cette taille de population) comme neutre. A l'inverse, si $Ns \gg 1$, la mutation tendra a être purgée de la population.

En ce qui concerne la coalescence, la présence de mutation délétères, qui ont $Ns \gg 1$, tend à réduire le nombre d'individus portant une séquence sans mutations délétère et pouvant ségreger assez longtemps pour être transmise sur de nombreuses générations. En ce sens, la sélection négative tend à donner $N_e < N$ et donc à accélérer le processus de coalescence.

5.3.2 Balayage sélectif

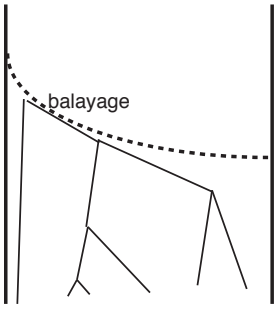

FIGURE 7 – Balayage sélectif

On parle de balayage sélectif (*selective sweep*) quand une mutation est tellement bénéfique que lorsque qu'elle apparaît, elle envahie "rapidement" toute la population et s'y fixe. Tous les nucléotides qui sont liées génétiquement à la mutation fortement sélectionnée profitent de cette fixation rapide pour augmenter en fréquence dans la population. C'est l'effet d'auto-stop (*hitchhicking*) des sites avoisinant la mutation sélectionnée. Lorsque la mutation arrive à fixation, tous les sites avoisinant doivent avoir un ancêtre commun au plus tard à la génération ou cette mutation est apparue. Le balayage est illustré sur la figure 7.

Ainsi, si l'on échantillonne pas longtemps après le balayage, on attend une situation équivalente à celle d'une population en croissance. Le balayage sélectif donc est un équivalent génétique d'une croissance démographique. Ainsi donc, on attend typiquement, après un balayage sélectif, une réduction drastique de la diversité et un $TajimaD$ négatif.

6 Recombinaison et Phylogénie

6.1 Recombinaison et coalescence

Jusqu'à présent, nous avons assumé qu'il n'existe qu'une généalogie pour tous les sites des séquences échantillonnées. Or si des événements de recombinaison ont eut lieu au cours de leur histoire généalogique, alors il n'existe plus qu'une seule généalogie pour tous les sites de la séquences.

n1 ...A...C...
n2 ...A...T...
n3 ...G...T...
n4 ...G...C...

TABLE 1 – Trace de la recombinaison

Imaginons, par exemple, que les séquences de quatre loci n1 à n4 sont telles que décrites dans la table 1 (les "." indiquent les sites non-polymorphes). Si on admet le modèle de sites infinis (toutes les mutations ont lieu sur un nouveau site), il n'existe pas un arbre unique capable de rendre compte des polymorphismes observés. En effet, le premier site (A/G) est compatible avec un arbre ((n1,n2)(n3,n4)) alors que le second site (C/T) est compatible avec un arbre ((n1,n3)(n2,n4)). Alors quel arbre est le vrai ?

En fait, les deux arbres sont vrais. Ici, nous avons un exemple ou un événement de recombinaison à découpler l'histoire du premier site de celui du second. Ainsi, lorsqu'il y a de la recombinaison entre les sites du locus étudié, il n'existe non plus un arbre unique pour la généalogie de la séquence mais plusieurs arbres plus ou moins indépendants attachés à une partie de cette séquence. Si la recombinaison est très forte, alors l'histoire de chaque site devient indépendante des autres sites.

Plus généralement, si l'on considère un échantillon de génomes où il y a de a recombinaison, il n'existe pas une unique généalogie valide pour tous les gènes de ces génomes. Au contraire, il en existe une par morceau de génome n'ayant pas recombiner depuis l'ancêtre commun. Il faut noter que les généalogies des sites les plus proches peuvent être partagée en partie : tant qu'il n'y a pas eut de recombinaison, l'histoire est la même pour les sites proches. Dès qu'un événement de recombinaison a lieu, leur histoire devient indépendante. Ainsi, comme les sites proches recombinent moins vite (en moyenne), ils partagent plus d'histoire généalogique.

6.2 Recombinaison et phylogénie

Si l'on effectue une reconstruction phylogénétique avec une séquence n'ayant pas recombinée (l'inverse ne ferait pas de sens), il existe plusieurs cause pour lesquels leur distance évolutive (nombre de mutations les séparant) peut varié. Tout d'abord pour un même temps de divergence, comme le nombre de mutations est dicté par un processus stochastique (loi de Poisson), on attend un nombre de mutations différent. Cependant, il existe un autre phénomène qui donne de la variance à la distance évolutive (nombre de mutations) entre deux gènes orthologues : les temps de coalescence de ces gènes dans la population de l'espèce ancêtre.

Nous venons de voir que, à cause de la recombinaison, l'arbre de coalescence de chaque gène d'un génome est plus ou moins indépendant. Cette observation a un impact direct sur la reconstruction phylogénétique faite à partir de plusieurs gènes d'un même génome. En effet, lorsque l'on tente de reconstruire la phylogénie d'une espèce, il faut considérer que deux gènes orthologues provenant d'espèces différentes ont évolué indépendamment tant que les deux espèces étaient séparées (comme deux sous-populations ayant un taux de migration nul). Mais dès lors que les deux gènes se retrouvent dans l'espèce ancêtre, leur temps de divergence est dicté par les processus stochastiques de coalescence. Ceci est illustré sur la figure 8

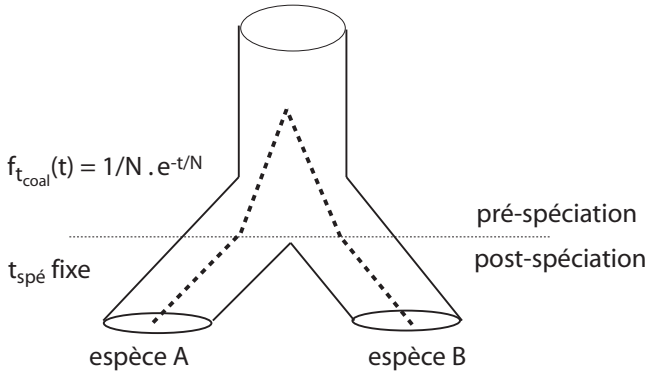

FIGURE 8 – Coalescence et Phylogénie

Ainsi, si le t_{spe} n'est pas trop grand par rapport aux valeurs de t_{coal}, la grande variance qui existe dans les arbres de coalescence aura pour effet de donner des temps de divergence très différents d'une paire de gène à une autre. Ainsi au cours de la reconstruction phylogénétique, la distance évolutive entre des paires de gènes orthologues peut être très différente d'une paire à une autre, d'une part à cause de la variance de la loi de Poisson (nombre de mutations pour un temps donné) et, d'autre part, à cause de variance de la loi exponentielle (temps de coalescence pré-spéciation).

7 Exercices

7.1 Mathématiques et Probabilité

Rappel sur les lois de probabilités importantes (Bernouilli, géométrique, binomiale et Poisson) et calcul de la moyenne et variance de la loi de Bernouilli ($E = p$, $Var = p(1-p)$), de la loi géométrique ($E = 1/p$, $Var = (1-p)/p^2$) et de la loi Poisson ($E = Var = \lambda$).

7.2 Diversité

7.2.1 *Homo sapiens*

On sait que si l'on prend deux séquences au hasard dans la population humaine moderne, on observe 1 nucléotide différent sur 1,300.

1. Combien vaut la diversité, θ ? Si on sait que $\mu \approx 10^{-8}$ mutations/base/génération, combien vaut N ? Quand vivait en moyenne l'ancêtre génétique de deux loci (en générations et en années) ?
2. Calculer t_{2min} et t_{2max}, les valeur bornes contenant les 95% de la distribution. Il faut pour cela calculer la fonction cumulative de la fonction de densité de probabilité de t_2 comme $F_{t_2}(t) = \int_0^t f_{t_2}(x)$ (c'est à dire la probabilité que t_2 soit plus petit ou égal qu'une valeur t). En utilisant cette fonction cumulative, on peut calculer les valeurs correspondant à correspondant aux 0.25 et 0.975.
3. Quel est le rapport entre le N estimé par la diversité et la taille réelle de la population humaine ($> 10^9$) ? Discussion autour de N_e (effectif efficace). Comment expliquer une telle différence entre N et N_e ?

7.2.2 *HIV*

On calcule que dans les séquences *gag-pol* d'une population de HIV-1 qui infecte un patient, on observe en moyenne 1% de nucléotide différent.

1. Calculer θ et N_e, sachant que l'on a mesuré expérimentalement le taux de mutation du virus à $\mu \approx 3.10^{-5}$ mutations/base/génération.
2. Si on estime N_e indépendamment de μ, on obtient $N_e \approx 10^3 - 10^4$. Qu'en conclure sur le taux de mutation ?

Discussion autour de μ_e (taux de mutation neutre). Sachant que le patient est infecté depuis plus de 10 ans et la population de HIV-1 réelle est de $N \approx 10^{10}$, imaginer des hypothèses pour expliquer la différence entre N et N_e.

7.3 Variance des arbres de coalescence

Créer des arbres de coalescence par simulation en utilisant le programme *tree_coal*. Visualiser par un logiciel les arbres obtenus (à l'aide du programme *njplot*) et contempler la variété des formes obtenues.

7.4 Génération et reconstruction d'arbres phylogénétiques

1. Créer des séquences issues d'un arbre de coalescence en utilisant le programme *generate_coalseq*. Ce ce programme crée un arbre de coalescence et distribue des mutations sur cet arbre. Puis, il génère les sites variables de ces séquences en créant des mutations $A \rightarrow G$. Faire varier θ entre 0.1, 1 ou 10 et n entre 4 et 10. Le programme génére deux fichier un fichier *.tree (l'arbre de coalescence) et *.phy (les séquences alignées au format phylip).
2. Reconstruire la phylogénie de ces séquences (par méthode parsimonie *dnapars*) puis comparé les arbres reconstruis aux arbres vrais (à l'aide de njplot). Discussion autour de la pertinence de la reconstruction.

7.5 Tajima's D pour le gène CCR5

Des chercheurs ont séquencer les séquences régulatrice du gène CCR5 d'individus de plusieurs localisations. Ce gène est impliqué dans la réponse au HIV-1 car une mutation, $\Delta 32$, donne à l'état homozygote une résistance complète au virus. Ce gène en général impliqué dans l'entrée des virus dans les cellules immunitaires (et p-e d'autres cellules). Voici les données qui ont été obtenus (Bamshad etal., PNAS, 2002) :

SNP	**(62) africain**	**(54) asiatique**	**(48) européen**	**(60) indien amér.**
a	0.258	0.074	0.104	0.150
b	0.016			
c	0.177	0.130	0.125	0.133
d	0.016			
e	0.016			
f	0.032	0.019	0.042	0.017
i	0.161	0.500	0.313	0.367
j	0.113			
k	0.161	0.280	0.292	0.217
l	0.016			
m	0.032			
n			0.104	
o			0.021	0.017

TABLE 2 – Données sur les fréquences alléliques du gène CCR5

On cherche à savoir si la répartition des mutations suit un modèle neutre.

1. Calculer le nombre de sites polymorphes (S) et le nombre moyen de différences (K) au sein de chaque populations.
2. Calculer le D de de Tajima et sa probabilité (en utilisant le programme tajimaD).

Astuce. On peut calculer K en utilisant les fréquences de chaque SNP grâce à :

$$K = 2 \times \frac{n}{n-1} \times (\sum_{i}^{SNP} p_i - p_i^2)$$

7.6 Le gène de la Lactase

Chez l'homme, le gène principal impliqué dans la disgestion du lait est le gènes LCT (fonction enzymatique : lactase-phlorizin hydrolase). Dans la plupart des espèces de grands singes, ce gène est exprimé seulement chez l'enfant. Chez certaines populations humaine, les adultes se nourrissent également de lait. On va chercher à étudier l'évolution "récente" de ce gène avec une approche coalescence.

1. Aller sur HapMap et récupérer les SNP du locus LCT.
2. En utilisant le programme *snp2fst.awk* fabriquer des sequences au format fasta, puis les "aligner" avec le programme *fst2aln*.
3. En utilisant, le programme *diversity*, calculer S, π et θ_S et θ_π.

Quel est la diversité des population du Japon (jpt), d'Europe (ceu), du Nigéria (yri) et de Chine (chb). Pourquoi ? Imaginer un scénario pouvant expliquer les observations.

7.7 Recombinaison sur les arbres phylogénétiques

A l'aide du simulateur (*generate_coalseq*), générer des séquences en autorisant une recombinaison totale entre chaque site. Reconstruire la phylogénie de ces séquences.

7.8 Transferts horizontaux ?

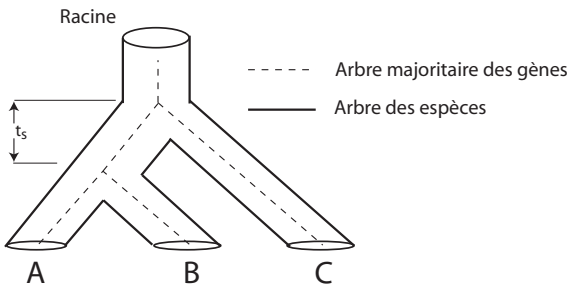

FIGURE 9 – Coalescence, Phylogénie et Spéciation

On dispose de trois espèces de bactéries proches et d'un groupe extérieur (outgroup) qui permet d'enraciner l'arbre. Dans chacune de ces espèces, on séquence plusieurs gènes

orthologues. La plupart des gènes suivent une phylogénie de l'espèce (((A,B),C),outgroup), décrite sur la figure 8, mais quelques gènes présentent une phylogénie (((A,C),B), outgroup) ou (((B,C),A), outgroup).

1. Doit-on faire appel à des transferts horizontaux pour expliquer les arbres "anormaux" ?
2. Comment-calculer en fonction du nombre de générations entre les deux événements de spéciation (t_s sur le dessin), la fraction de gènes ne suivant pas l'arbre des espèces ?