UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ESTIMATION DE L'HISTORIQUE DÉMOGRAPHIQUE D'UNE POPULATION DE VIRUS À PARTIR DE SÉQUENCES D'ADN PAR LA THÉORIE DE COALESCENCE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN MATHÉMATIQUES

PAR

SADOUNE AIT KACI AZZOU

JUIN 2016

REMERCIEMENTS

Mes remerciements vont en premier lieu à mes deux directeurs de thèse, Fabrice Larribe et Sorana Froda.

Merci Fabrice. Merci d'avoir accepté de prendre la direction de cette thèse. Merci pour ta grande disponibilité, ta confiance et tes encouragements qui permettent de transformer les difficultés rencontrées en une expérience enrichissante.

Merci Sorana. Merci pour ta générosité, pour ta rigueur, pour le temps consacré à corriger cette thèse, ainsi que les articles. Merci (Sorana et Fabrice) pour les longues heures passées, à débattre de mes idées, à discuter des moindres détails des formules mathématiques, souvent sur l'heure du midi.

Merci aux examinateurs Jinko Graham, Simon Gravel, et François Watier, qui ont accepté de siéger sur le jury de cette thèse.

Un petit clin d'œil à Marie Forest. Les discussions qu'on a eues à ton passage à Montréal m'ont beaucoup inspirées pour la suite du processus.

Merci à mes parents qui m'ont toujours soutenue et encouragé afin d'aller le plus loin possible dans mes études.

Merci à Lydia et à Yanis. Vous êtes le fruit de l'arbre que je suis.

Finalement, un merci tout particulier à mon amour Ghania. Merci pour ton soutien, ta compréhension, et ta patience.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

Symbole	Signification
μ	Taux de mutation par séquence par génération
N	Taille de la population au moment de l'échantillonnage des séquences
n	Nombre de séquences échantillonnées
T_k	Temps d'attente jusqu'au prochain événement de coalescence en présence de k séquences , avec $k = 2, ..., n$
T_{MRCA}	Temps de l'ancêtre commun
L	Longueur des séquences (nombre de nucléotides)
θ	Taux de mutation rééchelonné avec $\theta = 2N\mu$
$\mathcal{T}$	Topologie reliant les séquences
$\mathcal{W}$	Ensemble des longueurs des branches
$\mathcal{G}^{(j)}$	Généalogie $\mathcal{G}^{(j)} = (\mathcal{T}^{(j)}, \mathcal{W}^{(j)})$, $j = 1, 2, ..., J$
$\mathcal{D}$	Représente l'ensemble des séquences échantillonnées, avec $\lvert\mathcal{D}\rvert = n$
$L(\theta) = L(\mathcal{D}\lvert\theta)$	Vraisemblance complète de l'échantillon des séquences
$N_e(t)$	Taille de la population effective au temps t, $N_e(0) = N_e = N$
$\nu(t)$	Taille de population relative de $N_e(t)$ par rapport à N, $t > 0$
$V_{k+1} = \sum_{k=n}^{k+1} T_k$	$V_{k+1} = v_{k+1}$ est l'instant de départ du temps T_k
E	Ensemble des types de séquences échantillonnées
$\boldsymbol{P}$	Matrice de transition entre les 4 types de nucléotides ; {A,C,G,T}, de dimension 4×4

$P_{\alpha\beta}$	Probabilité de passage d'une séquence de type α vers une séquence de type β
$\boldsymbol{P}_{\alpha\beta}$	Matrice de probabilités de transition, de dimension $\|E\| \times \|E\|$ dont les éléments sont $P_{\alpha\beta}$, avec $(\alpha, \beta) \in E^2$
Q	Distribution d'échantillonnage pondéré ou distribution proposée
$P(\mathcal{G}^{(j)})$	Probabilité *forward* de la généalogie $\mathcal{G}^{(j)}$
$Q(\mathcal{G}^{(j)})$	Probabilité *backward* de la généalogie $\mathcal{G}^{(j)}$
$W_j = P(\mathcal{G}^{(j)})/Q(\mathcal{G}^{(j)})$	Poids d'échantillonnage pondéré de la généalogie $\mathcal{G}^{(j)}$
$w_j = W_j / \sum_{j=1}^{J} W_j$	Poids d'échantillonnage pondéré normalisé de la généalogie $\mathcal{G}^{(j)}$
H_i	Ensemble des séquences restantes à l'étape i ; $H_0 = \mathcal{D}$
$\mathcal{H}$	Ensemble des états $(H_{-m}, ..., H_{-1}, H_0)$ visités par le processus de Markov qui commence par le type génétique du $MRCA$ (H_{-m}), et se termine avec les types génétiques qui constituent l'échantillon H_0
n_i	Nombre de séquences restantes à l'état H_i
$n_i^{(\alpha)}$	Nombre de séquences de type α à l'état H_i
$p_\theta(H_i \| H_{i-1})$	Probabilité *forward* de passage de l'état H_{i-1} vers l'état H_i
$q_\theta(H_{i-1} \| H_i)$	Probabilité *backward* de passage de l'état H_i vers l'état H_{i-1}
$H_i = H_{i-1} - \alpha + \beta$	H_i est obtenu à partir de H_{i-1} par une mutation d'une séquence de type α vers le type β
$H_i = H_{i-1} + \alpha$	H_i est obtenu à partir de H_{i-1} par une division de la lignée de type α

RÉSUMÉ

L'évolution de la taille d'une population peut être retracée à partir d'un échantillon de séquences d'ADN. Dans cette thèse, nous proposons une nouvelle méthodologie non paramétrique basée sur une stratégie d'échantillonnage pondéré (*Importance Sampling*) qui permet d'explorer de tels historiques démographiques. L'essence de la méthode est de simuler un grand nombre de généalogies en utilisant le processus de coalescence, où l'information fournie par ces généalogies est combinée en utilisant les poids de cet échantillonnage pondéré.

En premier, nous proposons la méthode *skywis plot* qui débute par l'estimation de la taille de la population effective pour chaque généalogie, pour chaque intervalle de temps prédéfini, appelé *époque* ; ensuite, une moyenne pondérée de ces tailles de population estimées est calculée. Ainsi, les généalogies qui sont le plus en accord avec les données ont un poids plus élevé. Nous avons aussi généralisé notre méthodologie au cas d'un échantillonnage en série. Cela a nécessité la mise en œuvre d'une stratégie d'échantillonnage efficace qui permet de tenir compte de cette réalité qui est très utilisée, notamment dans le cas de virus qui évoluent rapidement comme le VIH.

Ensuite, nous proposons d'améliorer la performance de la méthode *skywis plot* à travers une procédure itérative appelée *iterative calibrated skywis plot* ; la taille de la population effective est approximée par une fonction en escalier, qu'on ré-estime après chaque itération en utilisant la méthode *calibrated skywis plot*. Ces fonctions en escalier sont utilisées pour générer les temps d'attente d'un processus de Poisson non homogène (coalescence avec mutation) sous un modèle avec une taille de population variable. Cela nous a aussi amené à adapter la distribution proposée de Stephens et Donnelly (2000).

Mots clés : historique démographique, échantillonnage pondéré, processus de coalescence, skywis plot, échantillonnage hétérochrone, processus non homogène.

INTRODUCTION

Les séquences d'ADN contiennent de l'information sur l'historique démographique de la population où les séquences ont été échantillonnées. Ainsi, avec la disponibilité de séquences complètes de plusieurs génomes de virus, le problème de l'estimation de la taille de la population est devenu un sujet important en statistique génétique, avec des applications pratiques qui permettent, par exemple, de prédire l'évolution des virus, d'étudier le lien entre les événements démographiques et climatiques, ou encore retracer la transmission et la propagation des virus.

Dans cette thèse, nous proposons une nouvelle méthodologie non paramétrique flexible où la connaissance de la fonction analytique qui régit la taille de la population effective n'est pas nécessaire.

Dans le premier chapitre, nous présentons les concepts de base de la biologie moléculaire, dont le but est de familiariser le lecteur avec les termes nécessaires à la compréhension des mécanismes d'évolution virale, comme la mutation qui est au cœur du sujet traité. Nous décrivons ensuite au chapitre 2 la théorie de coalescence, en commençant par le processus de coalescence classique qui suppose, entre autres, que la taille de la population reste constante à travers le temps. Cette hypothèse est ensuite levée en présentant le processus de coalescence dans le cas où la taille de population effective est variable, ainsi qu'en présence de la recombinaison. Au chapitre 3, on s'intéresse aux méthodes qui permettent d'approximer la vraisemblance $L(\theta)$ en utilisant l'échantillonnage pondéré. En particulier, on décrit en détails la distribution proposée par Stephens et Donnelly (2000), qui permet de simuler des généalogies de manière efficace. Le chapitre 4, quant à lui,

décrit les méthodes non paramétriques appelées *skyline plot* qui permettent d'estimer la taille de la population effective. En effet, notre méthodologie peut être considérée comme une amélioration de la méthode *skyline plot classique.*

Les chapitres 5 et 6 sont constitués de deux articles scientifiques en langue anglaise, qui présentent nos nouvelles méthodes. Le chapitre 5 décrit la méthode *skywis plot* qui s'appuie sur la simulation d'un grand nombre de généalogies en utilisant un échantillonnage pondéré. Ainsi, la taille de la population effective est d'abord estimée pour chacune des généalogies sur un nombre donné d'*époques* ; ces époques sont obtenues par cumul de temps de coalescence. Ensuite, une moyenne pondérée des estimés de la taille de la population effective est calculée pour chacune des *époques*, où les poids utilisés sont issus de l'échantillonnage pondéré. Notre méthode permet notamment d'affecter un plus grand poids aux généalogies les plus vraisemblables avec les séquences échantillonnées. De plus, notre méthodologie est généralisée au cas d'un échantillonnage hétérochrone. À cet effet, une nouvelle fonction d'importance est proposée afin de simuler des généalogies dans un cadre où les séquences étudiées sont échantillonnées à des intervalles de temps assez importants. Ainsi, nous montrons par simulation qu'un échantillonnage hétérochrone permet d'améliorer la qualité de l'estimation de l'historique démographique quand on se rapproche de l'ancêtre commun. Notons que cet article a été publié dans la revue *Frontiers in Genetics*, section *Evolutionary and Population Genetics.*

Le chapitre 6 décrit, à travers un deuxième article scientifique à paraître dans la revue *Theoretical Population Biology*, la méthode appelée *iterative calibrated skywis plot* ; cette méthode permet d'améliorer la performance du *skywis plot* dans le cas d'une évolution très rapide de la taille de la population effective. Cela est réalisé en approximant au préalable la taille de la population par un modèle où la taille de population est variable, mais constante par intervalle. Dans ce cas, le taux de coalescence est différent d'un intervalle à un autre, ce qui a nécessité d'adapter

la fonction d'importance de Stephens et Donnelly (2000) à cette problématique où le processus d'arrivée des événements (coalescence ou mutation) devient non-homogène.

CHAPITRE I

MÉCANISMES D'ÉVOLUTION DES VIRUS

Ce chapitre vise à présenter les concepts de base de la biologie moléculaire, ainsi que les mécanismes d'évolution virale comme la mutation, la recombinaison et le réassortiment. Le lecteur peut trouver plus de détails dans Holmes (2009).

1.1 Rappels de biologie moléculaire

1.1.1 L'ADN

L'information régissant les caractéristiques de tous les organismes est stockée dans une longue molécule d'acide désoxyribonucléique (ADN) qui se présente sous forme de deux chaînes de nucléotides, les brins. Chaque nucléotide est identifié par la base azotée qu'il contient ; il existe quatre principales bases azotées présentes dans l'ADN : l'adénine (A), cytosine (C), guanine (G) et thymine (T). On note que chaque gène correspond à une séquence précise d'ADN. Sur chaque brin d'ADN, les nucléotides sont liés entre eux par des liaisons 5'-3' phosphodiester entre le groupe phosphate d'un nucléotide et le sucre d'un autre nucléotide. La lecture de ces brins s'effectue dans le sens 5' vers 3' (voir figure 1.1).

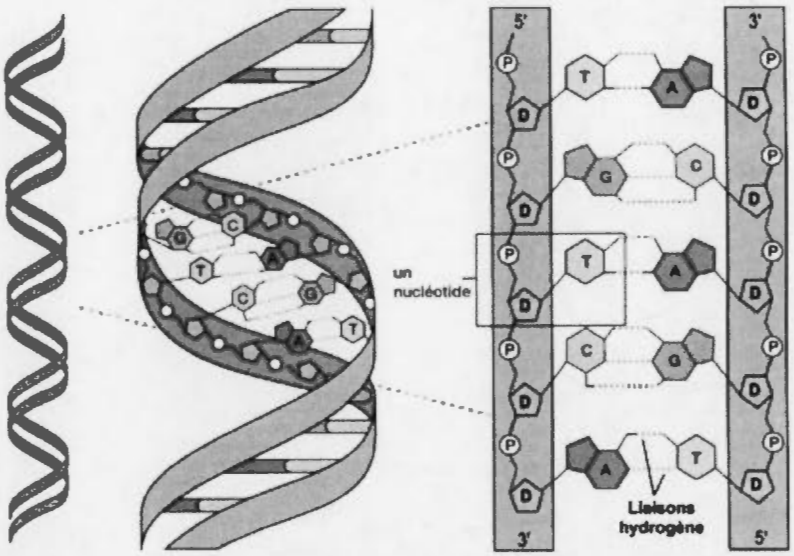


Figure 1.1: Appariement des bases azotées de l'ADN. Tiré à partir de :
`http://pta.nbed.nb.ca/bio/Bio\%2053411/Module\%204/notes4.1.html`

1.1.2 L'ARN

L'acide ribonucléique (ARN) est une molécule simple brin qui possède une structure similaire à l'ADN. Le processus de transfert d'information entre l'ADN, l'ARN et les protéines est appelé transcription. Au cours de la transcription, la séquence d'ARN messager (ARNm) obtenue est une copie simple brin linéaire de l'ADN dit codant, en remplaçant la Thymine (T) par l'Uracile (U). Les enzymes qui effectuent cette copie $ADN \rightarrow ARN$ s'appellent des ARN polymérases.

D'autres types d'ARN dits de structure (non codants) peuvent également être synthétisés. Par exemple, les ARN de transfert (ARNt) portent des acides aminés et permettent leur incorporation dans les protéines. Les acides aminés sont des molécules qui entrent dans la composition des protéines. Chaque acide aminé est représenté par au moins un codon, qui à son tour est sous forme d'une séquence de trois nucléotides.

1.2 Introduction aux virus

Un virus est une entité biologique qui nécessite une cellule hôte dont il utilise les constituants pour se multiplier. Le génome viral est constitué soit d'ADN ou d'ARN qui peuvent être simple brin ou double brin. On peut aussi distinguer les virus selon les caractéristiques physiques des virions[1] :

- virus nu ou enveloppé ;
- virus à symétrie hélicoïdale ou icosaédrique ;
- virus à génome linéaire ou circulaire, segmenté ou d'un seul tenant.

Le génome est entouré d'une coque de protéines appelée la capside dont la forme est à la base des différentes morphologies des virus. La taille des virus se situe entre 10 et 400 nanomètres[2] (nm) et le nombre de gènes sur les génomes des virus peut varier de 1 à 1 200. Le plus petit virus connu est le virus delta, qui parasite lui-même celui de l'hépatite B, et ne comporte qu'un seul gène ; le plus gros virus connu est le mimivirus, avec un diamètre qui atteint 400 nanomètres et un génome qui comporte 1 200 gènes.

Selon Lwoff (1957), un virus possède les caractéristiques suivantes :

- il ne contient qu'un seul type d'acide nucléique (ADN ou ARN) ;
- il y a multiplication de son matériel génétique ; par contre, il n'y a ni crois-

---

[1] Le virion constitue la forme infectieuse d'un virus, c'est-à-dire la forme sous laquelle il pénètre dans la cellule.

[2] Un nanomètre = un milliardième de mètre = 10^{-9} m.

sance ni fission[3] des virus ;

- il ne possède aucune des enzymes nécessaires pour produire de l'énergie ou pour se multiplier ;

- il est un parasite intracellulaire ; pour se reproduire, un virus doit impérativement pénétrer une cellule, détourner sa machinerie enzymatique afin qu'elle produise les protéines du virus puis quitter la cellule et en infecter une autre.

La figure 1.2 permet de visualiser et de comprendre la terminologie des différents constituants d'un virus.

Dans le cas des virus comme le VIH, l'ARN constitue le génome, alors que chez la grande majorité des organismes, c'est l'ADN qui remplit cette fonction. Ainsi, les virus peuvent justement être différenciés par type de génome (tableau 1.1).

Tableau 1.1: Classification des virus par type de génome

ADN	ARN
double brin (ADNdb)	double brin (ARNdb)
simple brin (ADNsb)	simple brin à polarité négative(ARNsb -)
	simple brin à polarité positive (ARNsb +)

On note que le génome des virus à ARN peut être codé dans deux directions différentes :

[3]La fission est un type de division d'une cellule parentale pour former deux cellules.

- soit les gènes sont stockés dans la direction $5' \rightarrow 3'$ (polarité positive), comme celle dans laquelle les gènes sont codés dans l'ARN messager des cellules ;

- soit ils sont stockés dans la direction opposée (polarité négative).

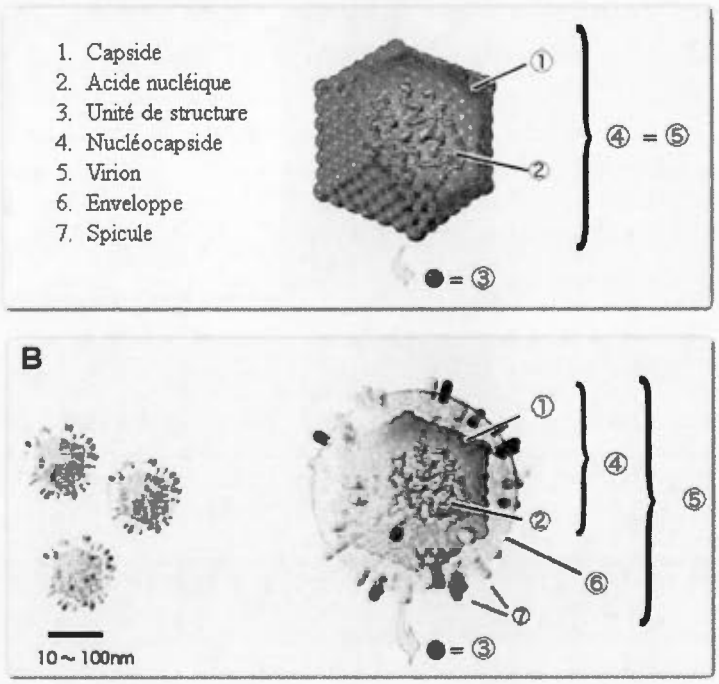


Figure 1.2: Schéma d'un virus. A) Virus nu . B) Virus enveloppé. Adapté à partir de : `http://commons.wikimedia.org/wiki/File:Virion.png`.

1.3 Les mécanismes de variabilité génétique

La variation est la base de l'évolution de toutes les formes de vie. Les virus utilisent toutes les stratégies possibles de la variation afin de rester en avance sur leur hôte. La variation crée de nouveaux génotypes qui aident le virus à s'adapter rapidement à des conditions environnementales qui changent. La mutation, la recombinaison

et le réassortiment sont quelques mécanismes par lesquels les virus créent de la variation.

1.3.1 Mutation

N'importe quel changement dans le matériel génétique (ADN, ARN) qui pourrait être transmis à la génération suivante est considéré comme étant une mutation. Les mutations se produisent principalement grâce à des erreurs lors de la réplication de l'acide nucléique. Le taux de mutation dépend du type d'acide nucléique ; contrairement à l'ADN polymérase où les erreurs de réplication sont rares, les erreurs de réplication sont beaucoup plus fréquentes pour l'ARN polymérase. La substitution est un type de mutation ponctuelle qui se traduit par le remplacement d'un nucléotide par un autre. Une substitution qui s'effectue entre une base purique vers une base purique (A, G) ou d'une base pyrimidique (C, T, U) à une autre base pyrimidique est appelée transition ($A \leftrightarrow G$ ou $C \leftrightarrow T$), par contre le remplacement d'une base purique par une base pyrimidique ou l'inverse est appelé transversion ($A \leftrightarrow C$, $C \leftrightarrow G$, $G \leftrightarrow T$ ou $A \leftrightarrow T$).
Il existe deux autres types de mutation nucléotidique (figure 1.3) :

- insertion : revient à l'ajout d'au moins une base dans la séquence ;
- délétion : un autre type de mutation qui correspond à la perte d'un ou de plusieurs nucléotides.

La division entre les virus ARN et ADN en termes de taux de mutation est reflété par leurs taux de substitution nucléotidique, qui peut différer jusqu'à 6 fois. Une preuve tangible de la rapidité de la substitution nucléotidique dans les virus ARN est que ce processus peut être souvent observé en temps réel, simplement en analysant les longueurs des branches des virus échantillonnés à différents temps (Drummond *et al.*, 2003b,a). Par conséquent, plusieurs virus ARN évoluent sur

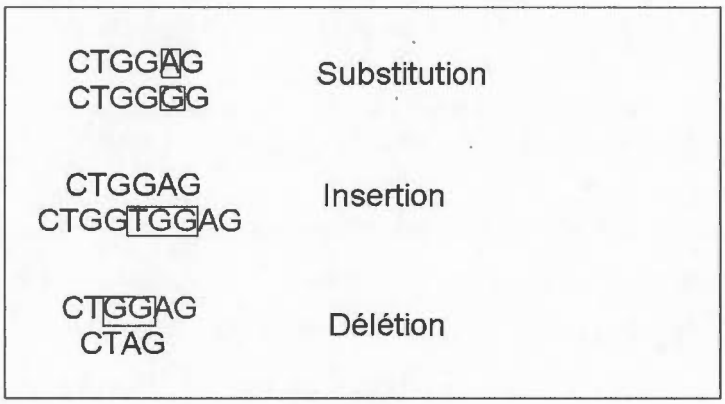


Figure 1.3: Types de mutations ponctuelles

une échelle de temps qui peut être enregistré par une observation humaine. Les taux de substitution nucléotidique des virus ARN qui ont été étudiés jusqu'à présent varient entre 10^{-2} et 10^{-5} (substitution/site/an) (Hanada *et al.*, 2004; Jenkins *et al.*, 2002).

1.3.2 Recombinaison et réassortiment dans l'évolution des virus ARN

Bien que la mutation soit la source principale de la variabilité génétique, il y a de plus en plus de travaux qui suggèrent que la recombinaison et le processus de réassortiment peuvent, dans certains cas, jouer un rôle significatif dans la mise en forme des « patterns » de la diversité génétique dans les virus ARN (Holmes , 2009).

Le processus de recombinaison permet d'échanger un fragment d'acide nucléique entre deux séquences d'ADN / ARN (figure 1.4). Ce type de recombinaison peut être homologue s'il se produit entre des régions de séquences homologues ou non homologue lorsque le matériel génétique se déplace entre des régions génomiques disjointes.

La recombinaison permet notamment aux virus d'éliminer les erreurs de réplication et à faire converger les souches virales. La recombinaison est observée principalement entre les virus à ARN qui leur permet de changer rapidement la virulence. Le taux de recombinaison est exprimé en nombre d'enjambements (crossover) par génome viral par réplication (par exemple, pour le VIH-1, on a 2.8 "crossovers" par génome, par cycle).

Il est important de faire la distinction entre la recombinaison, qui théoriquement peut se produire dans tous les virus ARN, et le réassortiment, qui se produit sur un sous ensemble de virus ARN qui possèdent des génomes segmentés - constitués de plusieurs fragments- (voir les figures 1.4 et 1.5). Bien que les deux processus puissent être perçus comme une forme de reproduction sexuelle au sens large, et exigent deux virus pour co-infecter la même cellule, ils sont mécaniquement très différents.La recombinaison dans les virus ARN se produit lorsque deux virus co-infectent la même cellule et une molécule hybride est produite à travers un processus appelé réplication par choix de copie (Lai, 1992).

Le processus de réassortiment se produit seulement sur des virus ARN segmentés. Dans ce cas, deux virus co-infectent une même cellule et les réassortis sont formés lorsque les segments du virus descendant proviennent de ceux des ancêtres différents.

Bien que les virus ARN et ADNsb soient soumis aux mêmes processus d'évolution de base que les autres organismes, on pourrait dire qu'ils occupent une région d'espace paramétrique d'évolution très différente des virus ADNdb. Le tableau suivant résume ces différences.

Tableau 1.2: Différence entre les virus ARN et ADNsb

ARN et ADNsb	**ADNdb**
Taux de mutation élevé (par nt)	Faible taux de mutation
Taille de génome réduite (<32 000 nt)	Grande taille du génome
Grandes tailles de population (toujours)	La taille de la population est faible
Duplication des gènes non fréquents	Duplication des gènes fréquents
En général, faible taux de recombinaison	recombinaisons fréquentes

Dans cette thèse, on s'intéresse à l'estimation de la taille de population de virus ARN ; or les virus ARN sont caractérisés par un taux de mutation élevé et une longueur de séquences réduite. Ces propriétés coïncident avec les hypothèses d'un modèle de mutation à sites finis décrit plus loin à la section 2.3.

Notre méthode qui permet d'estimer la taille de population effective dans le cas de modèles de mutation à sites finis est basée essentiellement sur la théorie de coalescence que l'on décrit à la section suivante.

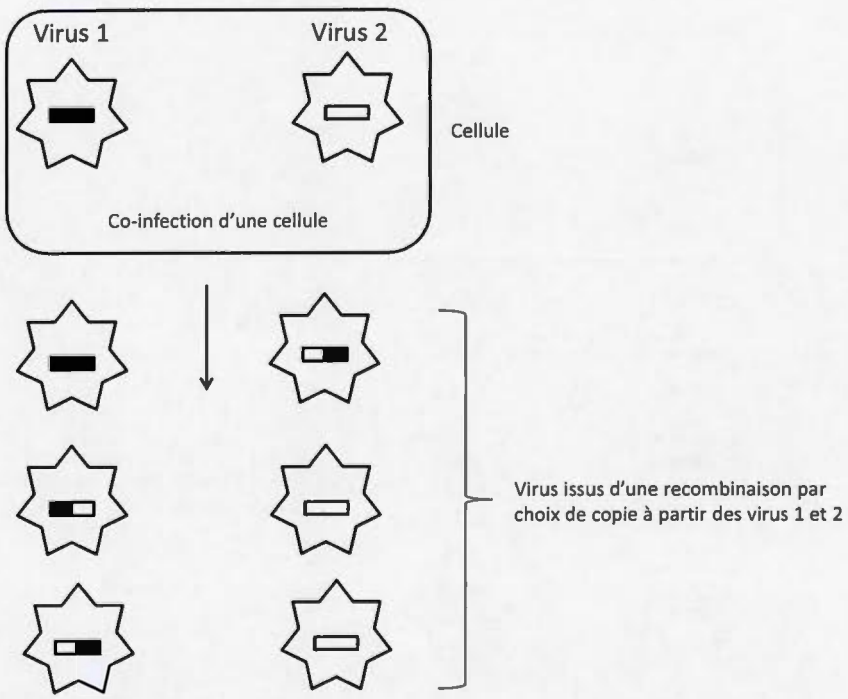


Figure 1.4: Illustration d'une recombinaison de virus ARN par "choix de copie". À partir de deux virus qui co-infectent une cellule, les virus répliqués sont soit du même type que les virus originaux, soit des combinaisons des virus originaux. Adapté à partir de Holmes (2009).

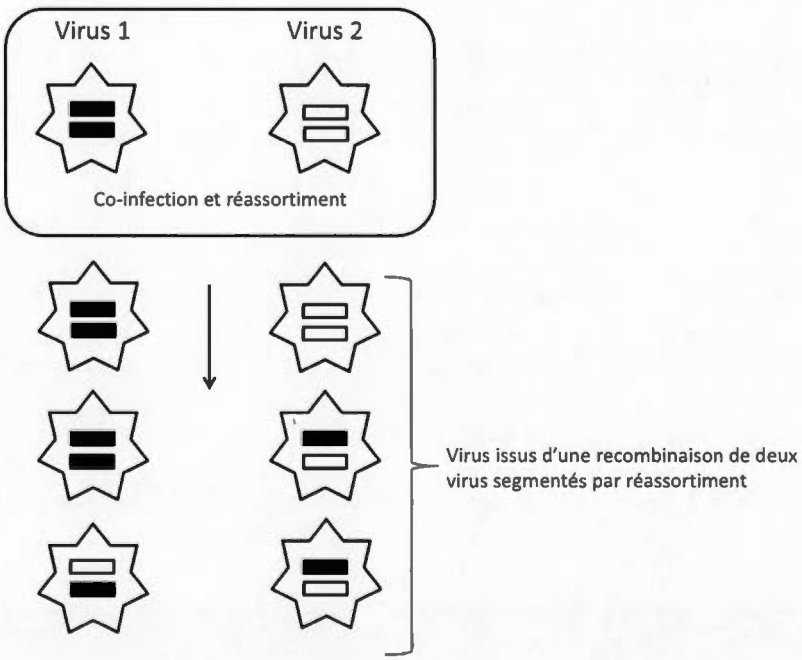


Figure 1.5: Illustration d'une recombinaison de virus ARN à génomes segmentés. À partir de deux virus qui co-infectent une même cellule. Le réassortiment crée de nouvelles configurations génétiques des virus ARN en échangeant des segments provenant des deux virus originaux. Adapté à partir de Holmes (2009).

CHAPITRE II

UN APERÇU SUR LA THÉORIE DE LA COALESCENCE

2.1 Introduction à la théorie de la coalescence

Le processus de coalescence permet de modéliser la généalogie de séquences d'ADN de manière rétrospective ; c'est-à-dire, en disposant d'un échantillon de séquences, la généalogie est recréée en remontant dans le temps, du présent vers le passé. Chaque séquence de l'échantillon de taille n, choisit ses parents de manière aléatoire parmi les séquences de la génération précédente. Lorsque deux séquences choisissent le même parent, on dit que les lignées coalescent. Ainsi, on arrive toujours à trouver le plus récent ancêtre commun pour l'échantillon en entier, le *MRCA* (*Most Recent Common Ancestor*), au bout de quelques générations. La figure 2.1 illustre ce que nous venons de décrire. Dans ce qui suit, on supposera que l'on dispose d'une population haploïde[1] de taille $2N$.

Dans sa forme la plus simple, la théorie de coalescence est basée sur les propriétés du modèle de Wright-Fisher sur une population haploïde. Le modèle de Wright-Fisher introduit par Wright (1931) et Fisher (1930), est un modèle de reproduction génétique de base qui décrit l'évolution d'une population idéale. Le modèle suppose une taille de population constante, pas de recombinaison, pas de sélection et pas

[1] Chaque individu de la population que nous étudions descend d'un seul parent.

de mutation. On suppose en plus que les générations ne se chevauchent pas ; ce qui veut dire que tous les individus (séquences) d'une même génération ont une espérance de vie égale.

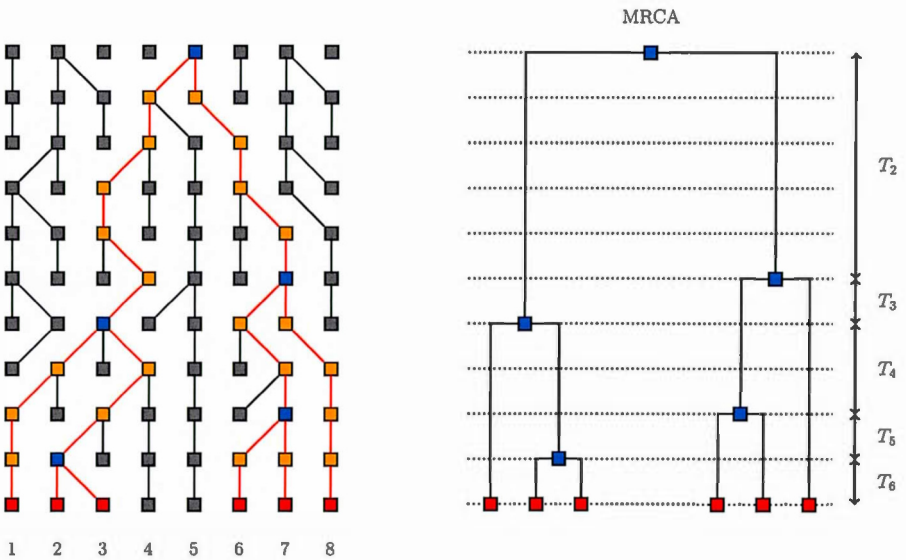


Figure 2.1: Généalogie d'une population de 8 séquences. Dans la partie gauche de la figure, les rectangles rouges représentent l'échantillon des séquences (dernière ligne) et leurs ancêtres pour une population de taille fixe $2N = 8$, sur 11 générations. La partie droite de la figure, résume l'information sur la connexion des séquences ainsi que les temps de coalescences, sous la forme d'un arbre.

Dans ce qui suit, nous présentons brièvement les résultats de la théorie de coalescence introduite par Kingman (1982a,b), qui sont directement liés à notre recherche (Chapitres 5 et 6).

2.1.1 Coalescence à temps discret

On commence par considérer le cas de deux séquences, et on décrit la loi de probabilité du temps d'attente pour que deux séquences trouvent un ancêtre commun, le *MRCA*.

La probabilité que deux séquences aient le même ancêtre dans la génération précédente est $1/(2N)$: la première séquence choisit son parent librement, et la deuxième séquence doit choisir le même parent. Puisque les générations sont indépendantes, la probabilité que deux séquences trouvent un ancêtre commun, j générations dans le passé est

$$\left(1-\frac{1}{2N}\right)^{j-1}\frac{1}{2N}.$$

En effet, dans les $(j-1)$ premières générations, les séquences choisissent des ancêtres différents, puis choisissent le même ancêtre à la génération j. Ainsi, le temps de coalescence T_2 pour que deux séquences trouvent un ancêtre commun suit une loi géométrique de paramètre $(1/2N)$:

$$P(T_2=j)=\left(1-\frac{1}{2N}\right)^{j-1}\frac{1}{2N}. \qquad j=1,2,... \tag{2.1}$$

On a, par conséquent,

$$E(T_2)=\frac{1}{(1/2N)}=2N \text{ générations,}$$

ce qui veut dire que le temps moyen pour que deux séquences trouvent le *MRCA* est égal au nombre de séquences dans la population.

Considérons maintenant qu'on dispose d'un échantillon de séquences de taille k et soit $P(k)$, la probabilité que les k séquences aient k différents ancêtres dans la génération précédente.

D'après la section précédente, la probabilité que deux séquences aient un ancêtre différent est égale à $(1-(1/2N))$. La probabilité que trois séquences ne descendent pas du même parent, est égale $(1-(1/2N))\times((2N-2)/2N))$. Le terme $((2N-2)/2N))$ représente la probabilité que la troisième séquence ait un ancêtre différent des deux premières. Cela donne

$$P(3) = \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right).$$

En utilisant le même raisonnement, on aura, pour $k \geq 2$,

$$P(k) = \frac{2N-1}{2N} \times \frac{2N-2}{2N} \times ... \times \frac{2N-k+1}{2N} = \prod_{i=1}^{k-1} \left(1 - \frac{i}{2N}\right)$$

$$= 1 - \sum_{i=1}^{k-1} \frac{i}{2N} + O\left(\frac{1}{N^2}\right) = 1 - \binom{k}{2}\frac{1}{2N} + O\left(\frac{1}{N^2}\right), \tag{2.2}$$

tel que $O\left(\frac{1}{N^2}\right)$ regroupe tous les termes divisés par N^2 ou par n'importe quelle puissance de N supérieure à deux. Cette approximation est équivalente à ignorer la possibilité d'avoir plus qu'une paire de séquences qui trouve un ancêtre commun à la même génération. Ainsi, en supposant que k est négligeable par rapport à N ($k \ll N$), la probabilité pour qu'il n'y ait pas d'événement de coalescence est

$$1 - \binom{k}{2}\frac{1}{2N};$$

par conséquent, la probabilité pour qu'un événement de coalescence se produise à une génération donnée, en présence de k séquences, est

$$\binom{k}{2}\frac{1}{2N}.$$

Soit T_k, le nombre de générations jusqu'à l'obtention d'un premier événement de coalescence en présence de k séquences. La probabilité que deux séquences parmi k trouvent un ancêtre commun $\{T_k = j\}$, j générations dans le passé est :

$$P(T_k = j) \approx \left(1 - \binom{k}{2}\frac{1}{2N}\right)^{j-1} \times \left(\binom{k}{2}\frac{1}{2N}\right), \tag{2.3}$$

et T_k suit approximativement une loi géométrique de paramètre $\binom{k}{2}\frac{1}{2N}$; de plus, $T_2, ..., T_n$ sont des variables indépendantes.

2.1.2 Coalescence en temps continu

Dans le modèle de Wright-Fisher, le temps est mesuré de façon discrète (générations). Il est par contre plus avantageux pour des raisons conceptuelles et de calcul, de considérer une approximation en temps continu.

L'échelle de temps choisie correspond au temps moyen pour que deux séquences trouvent un ancêtre commun ($2N$, d'après la section précédente). En effectuant cette transformation, le processus de coalescence devient indépendant de la taille de la population, car le temps moyen pour que 2 séquences trouvent un ancêtre commun est 1. Afin de déduire le processus de coalescence en temps continu, on pose $t = j/(2N)$ où j est mesuré en générations (Hein *et al.*, 2005). Ainsi, en présence de k séquences dans la généalogie, T_k suit une loi exponentielle de paramètre $\binom{k}{2}$, ce qui donne

$$P(T_k \leq t) = 1 - \exp\left(-\binom{k}{2} t\right). \tag{2.4}$$

2.1.3 Mesures de la taille d'une généalogie

Deux quantités d'intérêt qui mesurent la taille d'une généalogie sont : le temps total jusqu'à ce que l'on trouve l'ancêtre commun , T_{MRCA}, et la longueur globale de toutes les branches dans la généalogie, T_{total} (Wakeley, 2008). T_{MRCA} représente la somme des temps d'attente (voir figure 2.1)

$$T_{\text{MRCA}} = \sum_{k=2}^{n} T_k. \tag{2.5}$$

Le calcul de l'espérance et de la variance de T_{MRCA} développé ci-dessous est basé sur le fait que les T_k suivent des lois exponentielles indépendantes, de paramètres

$(k(k-1)/2)$. Ainsi, l'espérance du temps de l'ancêtre commun est donné par :

$$\begin{aligned} E(T_{\text{MRCA}}) =& E\left(\sum_{k=2}^{n} T_k\right) = \sum_{k=2}^{n} E(T_k) \\ =& \sum_{k=2}^{n} \left(\frac{2}{k(k-1)}\right) \\ =& \left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \ldots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n}\right) \\ =& 2\left(1 - \frac{1}{n}\right), \end{aligned} \tag{2.6}$$

et la variance du temps T_{MRCA} est donnée par :

$$\begin{aligned} Var(T_{\text{MRCA}}) = Var\left(\sum_{k=2}^{n} T_k\right) =& \sum_{k=2}^{n} Var(T_k) \\ =& 4\sum_{k=2}^{n} \frac{1}{k^2(k-1)^2} \\ =& 8\sum_{k=2}^{n} \frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)^2. \end{aligned} \tag{2.7}$$

On voit facilement qu'on a les deux propriétés asymptotiques suivantes (Wakeley, 2008) :

- l'espérance du temps jusqu'à l'ancêtre commun, $E(T_{\text{MRCA}})$, converge vers la valeur 2 ;
- $\lim_{n\to\infty} Var(T_{\text{MRCA}}) = 8\,(\pi^2/6 - 1) - 4 \approx 1.16$.

Rappelons que $E(T_2) = 1$, ce qui signifie que le temps moyen pour avoir un événement de coalescence, lorsque seulement deux séquences sont présentes dans la généalogie, représente environ la moitié de la hauteur de la généalogie.

La deuxième quantité d'intérêt est la longueur totale des branches de la généalogie, T_{Total}, qui est calculée en pondérant les temps de coalescence en présence d'un

certain nombre d'ancêtres par le nombre d'ancêtres présents au temps respectif.

$$T_{\text{Total}} = \sum_{k=2}^{n} kT_k. \tag{2.8}$$

L'espérance de la longueur totale des branches de la généalogie est donné par

$$E(T_{\text{Total}}) = E\left(\sum_{k=2}^{n} kT_k\right) = \sum_{k=2}^{n} \frac{2}{k-1}. \tag{2.9}$$

$E(T_{\text{Total}})$, se comporte comme $2(\log(n) + \gamma)$ (Wakeley, 2008), où $\gamma \approx 0.577216$ représente la constante d'Euler, qui se définit comme

$$\gamma = \lim_{n\to\infty} \left(\sum_{k=2}^{n} 1/k - \log(n)\right). \tag{2.10}$$

L'équation (2.9) montre que l'espérance de la longueur totale des branches d'une généalogie, $E(T_{\text{Total}})$, croît sans limite en fonction de n. De plus, la variance de T_{Total} peut être facilement calculée :

$$Var(T_{\text{Total}}) = 4\sum_{k=1}^{n-1} \frac{1}{k^2}, \tag{2.11}$$

dont la limite est $\frac{2}{3}\pi^2$ lorsque $n \to \infty$.

2.1.4 Modèle de Wright-Fisher avec mutation

Le modèle de Wright-Fisher de base est un modèle de reproduction sans aucune information sur le type génétique. On pourrait imposer un processus de mutation par dessus le modèle de reproduction, en supposant un modèle d'évolution neutre. On suppose qu'un événement de mutation se produit avec une probabilité μ, ce qui veut dire qu'une séquence peut être copiée sans modification (ou erreur) avec une probabilité $1 - \mu$.

Sous un modèle à sites finis, une position sur la séquence est choisie aléatoirement et une mutation se produit en cette position selon un modèle spécifié. Toutes

les mutations vont être supposées neutres dans le sens où le type d'une *séquence parent* dans une génération n'influence pas la probabilité de l'apparition d'une mutation pour la *séquence enfant* de la génération suivante.

La probabilité d'avoir une mutation pour la première fois sur une lignée donnée, j générations dans le passé est donnée par

$$P(T_M = j) = \mu(1-\mu)^{j-1},\ j = 1, 2, ... \tag{2.12}$$

où T_M représente le nombre de générations jusqu'au premier événement de mutation (T_M suit une loi géométrique de paramètre μ).

Comme pour obtenir la loi du temps de coalescence T_k en temps continue, l'approximation de T_M en temps continu est obtenue en posant $t = j/2N$ et $\theta = 4N\mu$ tel que j est mesuré en unités de $2N$ générations. Suite à cela, et en supposant que $2N$ est grand ($N \to \infty$), on a :

$$P(T_M \leq j) = 1 - (1-\mu)^j = 1 - (1 - \frac{\theta}{4N})^{2Nt} \approx 1 - e^{-\theta t/2}, \tag{2.13}$$

Le paramètre θ représente le taux de mutation de la population : il s'interprète comme l'espérance du nombre de mutations séparant un échantillon de deux séquences.

En présence de k lignées, le temps d'attente jusqu'au premier événement de mutation dans n'importe laquelle des k lignées va suivre une loi exponentielle de paramètre $k\theta/2$ en supposant que les k lignées mutent de façon indépendante.

De même, si on considère que l'on a des événements de coalescence et de mutation indépendants, alors le temps d'attente jusqu'au premier événement suit une loi exponentielle de paramètre :

$$\binom{k}{2} + \frac{k\theta}{2} = \frac{k(k-1+\theta)}{2}. \tag{2.14}$$

D'après ce qui précède, le prochain événement est une coalescence avec la probabilité

$$\frac{\binom{k}{2}}{\binom{k}{2}+\frac{k\theta}{2}} = \frac{k-1}{k-1+\theta}, \tag{2.15}$$

et un événement de mutation avec la probabilité

$$1-\frac{k-1}{k-1+\theta} = \frac{\theta}{k-1+\theta}. \tag{2.16}$$

On note que la paire de lignées qui fusionnent est choisie aléatoirement parmi toutes les paires et la mutation est choisie aléatoirement parmi les k lignées.

2.2 Algorithmes de simulation d'évolution de séquences

Dans notre recherche, nous avons recours à la simulation d'historiques. Ainsi, la simulation de tels historiques d'échantillons est très importante puisque :

- plusieurs quantités d'intérêt ne peuvent être calculées explicitement, ce qui nous oblige à recourir à la simulation ;
- la simulation donne une intuition sur la dynamique du processus de coalescence et du processus de mutation ;
- une bonne simulation d'historiques d'échantillons devient un enjeu majeur pour les procédures inférentielles.

On note que la simulation de l'historique d'un ensemble de séquences se fait toujours du présent vers le passé, ce qui a comme avantage de suivre seulement les ancêtres de l'échantillon considéré, plutôt que ceux de la population entière. Pour cela, on utilise en général deux algorithmes :

- un algorithme basé sur les équations (2.14), (2.15) et (2.16), noté algorithme 1 ;
- un algorithme, qui utilise le fait que les mutations pourront être ajoutées sur la généalogie une fois que celle-ci a été simulée (algorithme 2). Cette supposition est valable dans le cas où processus de mutation est neutre, et sera utilisée tout au long de la thèse.

Algorithme 1

1. poser $k = n$, où n est la taille de l'échantillon ;
2. simuler le temps du prochain événement selon une loi exponentielle de paramètre
$$k(k-1+\theta)/2;$$
3. on considère qu'on a un événement de coalescence avec une probabilité
$$(k-1)/(k-1+\theta),$$
et un événement de mutation avec une probabilité
$$\theta/(k-1+\theta);$$
4. selon le résultat à l'étape 3 :
   (a) on choisit aléatoirement une paire à coalescer et on pose $k \to k-1$, s'il s'agit d'un événement de coalescence ;
   (b) on choisit aléatoirement une lignée à muter et on laisse k inchangé, dans le cas où un événement de mutation se produit ;
5. on revient à l'étape 2 si $k > 1$, sinon fin de l'algorithme.

La deuxième méthode qui permet de simuler un historique d'un échantillon de séquences avec mutations est basée sur le fait que :

- le temps d'attente jusqu'à l'apparition d'une mutation sur une lignée suit une loi exponentielle de paramètre $\theta/2$.

- sur une branche de longueur t, le nombre de mutations, M_t sur la branche suit une loi de Poisson de paramètre $t\theta/2$, et donc :

$$P(M_t = j) = \frac{(t\theta)^j}{j!2^j} e^{-t\theta/2}. \tag{2.17}$$

Le nombre et les temps d'arrivée des mutations sur des branches différentes sont supposés indépendants. De ce fait, chaque branche pourrait être traitée indépendamment.

D'après ce qui précède, l'algorithme 2, pour la simulation de l'historique d'un ensemble de séquences avec mutations est énoncé ci-après :

Algorithme 2

1. simuler la généalogie de n séquences suivant un processus de coalescence avec un taux $\binom{k}{2}$ pour k lignées ;

2. pour chaque branche générer un nombre aléatoire, M_t à partir d'une loi de Poisson de paramètre $t\theta/2$ tel que t représente la longueur de la branche ;

3. les temps des événements de mutations, M_t, sont ensuite choisis aléatoirement sur la branche ;

Notons que les algorithmes 1 et 2 produisent des généalogies équivalentes.

2.3 Modèles de mutation

On a discuté jusque-là de la manière avec laquelle les séquences sont reliées entre elles dans la population jusqu'à leur ancêtre commun. Cependant, pour traiter des données réelles, il est nécessaire de considérer un modèle qui décrit comment les mutations causent des changements dans l'ADN/ARN. À cet effet, plusieurs modèles ont été proposés. Historiquement, le modèle à allèles infinis a été le premier à être proposé par Kimura et Crow (1964), suivi du modèle à sites infinis (Kimura, 1969), et le modèle à sites finis (Jukes et Cantor, 1969).

Dans ce chapitre, on introduit ces trois différents modèles de mutation en s'attardant sur le modèle à sites finis qui nous intéresse le plus dans notre cas, comme les virus ARN sont caractérisés par des taux de mutation élevés.

- *Modèles à sites infinis*

  Dans ce cas, on suppose que chaque mutation se produit à un nouveau site et le nombre de sites potentiels est infini. Le modèle à sites infinis décrit l'évolution des longues séquences d'ADN/ARN avec un faible taux de mutation à chaque site. Ainsi, les mutations sont non récurrentes dans ce modèle.

- *Modèles à allèles infinis*

  Dans ce cas, on fait l'hypothèse qu'une mutation provoque l'apparition d'un nouvel allèle qui n'était encore jamais apparu dans la population.

- *Modèles à sites finis*

  Un modèle d'évolution tout à fait réaliste dans le cas des virus, par exemple, implique une spécification complète des séquences d'ADN/ARN supposées de longueur fixe[2] L. En principe, un modèle de mutation devrait inclure les

[2]La longueur L d'une séquence représente le nombre de nucléotides dans cette séquence.

insertions et les délétions, mais ceux-là se produisent assez rarement dans une population de séquences qu'elles peuvent être identifiées facilement par alignement. C'est la raison pour laquelle il est usuel d'ignorer les insertions et les délétions par la suite.

Dans un modèle à sites finis, chaque position d'un nucléotide peut subir plus d'une mutation. De plus, deux séquences peuvent être identiques, même si plusieurs mutations ont eu lieu dans leur historique : une mutation en « avant » et « en arrière » pourrait effacer l'effet des deux mutations (par exemple, $A \to T$ et $T \to A$).

On note que, sous un modèle à sites finis, on s'intéresse aux nucléotides plutôt qu'aux séquences en entier. Le taux de mutation par nucléotide est $\theta_0 = \theta/L$ où :

- L représente le nombre de nucléotides d'une séquence ;
- $\theta = 2N\mu$, où μ est le taux de mutation par séquence par génération.

La figure (2.2) montre un exemple où, parmi la série d'événements ayant eu lieu, deux mutations se sont produites dans la même position (position 7, de la séquence 2), ce qui n'est pas possible sous un modèle à sites infinis.

Dans la section suivante, on s'intéresse exclusivement aux modèles de mutation à sites finis, une supposition adaptée à l'évolution des virus, sur laquelle on a travaillé tout au long de cette thèse.

2.3.1 Théorie des modèles d'évolution moléculaire

On propose dans cette section de décrire l'évolution des séquences ADN de longueur L par un L-processus de Markov au long d'un arbre, où chaque processus prend ses valeurs dans l'espace d'états $E = \{A, G, C, T\}$. Ces processus vont être considérés comme des chaînes de Markov à temps continu (Galtier *et al.*, 2005).

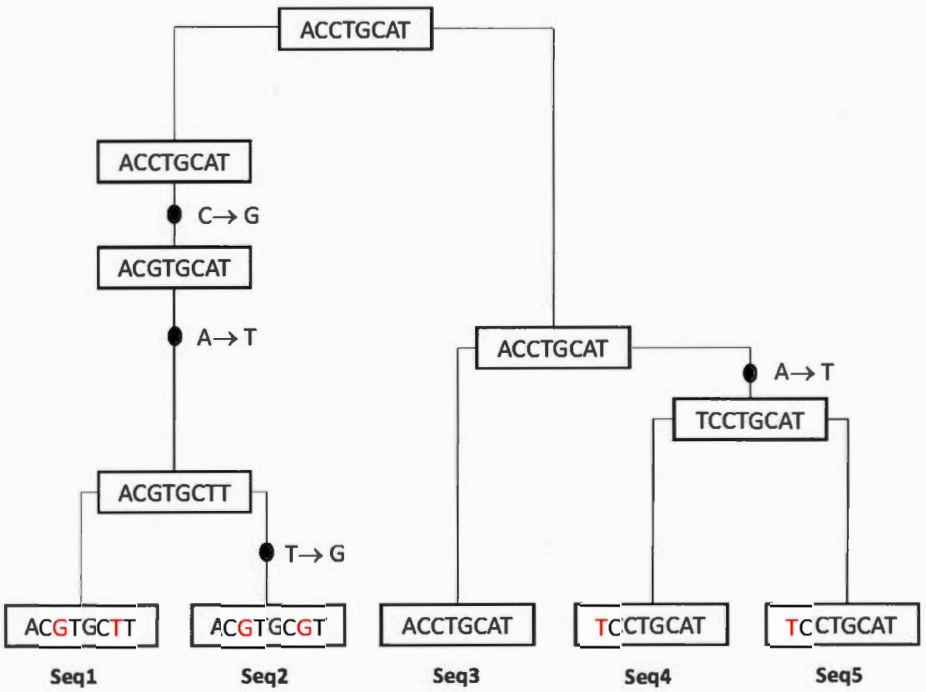


Figure 2.2: Exemple d'une généalogie sous un modèle à sites finis ; adapté à partir de Hein *et al.* (2005).

Dans ce qui suit, on suppose que les processus de substitution relatifs aux L sites vérifient les conditions suivantes :

(a) les sites évoluent indépendamment les uns des autres, puisqu'une mutation qui se produit en une position donnée n'influence pas la chance qu'on ait une mutation dans une autre position ;

(b) les sites sont identiquement distribués : l'ensemble des sites d'une séquence donnée est régit par le même processus de substitution E ;

(c) le processus de substitution E est markovien. En effet, on a, pour un site i,

$i = 1, 2, ..., L$, et tout $0 < s_1 < s_2 < ... < s_k < s$

$$\begin{aligned} P_{vw,s}(t) =& \Pr\left(E_i(s+t) = w \mid E_i(s_\ell) = v_\ell,\ 0 \leq s_\ell \leq s,\ 1 \leq \ell \leq k\right) \\ =& P\left(E_i(s+t) = w \mid E_i(s) = v\right), \end{aligned} \tag{2.18}$$

où $P_{vw,s}(t)$ représente la probabilité de changement du site i, de l'état v à l'instant s vers l'état w à l'instant $(s+t)$;

(d) le processus de substitution est homogène dans le temps, donc $P_{vw,s}(t)$ ne dépend pas du temps de départ ; c'est-à-dire, pour tout s, s', $\forall s \geq 0, s' \geq 0$,

$$\begin{aligned} P_{vw}(t) =& P\left(E_i(s+t) = w \mid E_i(s) = v\right) \\ =& P\left(E_i(s'+t) = w \mid E_i(s') = v\right). \end{aligned} \tag{2.19}$$

(e) le processus de substitution est supposé stationnaire, donc, il existe $\pi_v \geq 0$, $\sum \pi_v = 1$ telle que la probabilité d'observer un état donné ne dépend pas du temps t lorsque $t \to +\infty$, et on a

$$\lim_{t \to +\infty} P(E_i(t) = v) = \pi_v,$$

où π_v est la fréquence de l'état v lorsque le processus de substitution est rendu à l'état stationnaire.

(f) le processus de substitution est réversible :

$$\pi_v P_{vw}(t) = \pi_w P_{wv}(t). \tag{2.20}$$

2.3.2 Matrice des taux de substitution instantanés

Dans la définition des différents processus de mutation, on fait appel à la matrice des taux instantanés $Q = (q_{vw})$, où, l'on a lorsque $h \to 0$,

$$\Pr\left(E_i(s+h) = w \mid E_i(s) = v\right) \approx q_{vw} h.$$

Quand on travaille avec des séquences d'ADN, l'espace d'états E est donné par $\{A, G, C, T\}$ et Q s'écrit comme suit :

$$Q = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix},$$

où $q_{\ell k} \geq 0$ pour $\ell \neq k$, et $q_{\ell\ell} = -\sum_{k \neq \ell} q_{\ell k}$, avec $\ell, k \in \{A, G, C, T\}$.

À partir de la matrice des taux de substitution instantanés, la matrice des probabilités de transition est donnée par $P(t) = e^{Qt}$.

Dans ce qui suit, on introduit quelques modèles de substitutions caractérisés chacun par sa propre matrice de taux de substitution instantanés. Le modèle introduit par Jukes et Cantor (1969), noté JC69, est le modèle à sites finis le plus simple. Dans ce cas, toutes les positions ont la même probabilité de subir une mutation, où à partir d'une position donnée, le nucléotide mute vers les trois autres nucléotides possibles avec la même probabilité. Ce qui se traduit dans la matrice Q par

$$q_{\ell k} = \mu, (\ell, k) \in E^2, \ell \neq k,$$

et la matrice des taux de substitution instantanés est

$$Q = \begin{pmatrix} -\lambda & \mu & \mu & \mu \\ \mu & -\lambda & \mu & \mu \\ \mu & \mu & -\lambda & \mu \\ \mu & \mu & \mu & -\lambda \end{pmatrix}.$$

Ainsi, on a un taux de substitution égal à μ pour les quatre nucléotides, $\lambda_\ell = \lambda = 3\mu$, $\pi_k = 1/4,\ k \in \{A, G, C, T\}$, et avec la contrainte $p_{\ell\ell}(t) + 3p_{\ell k}(t) = 1$, $\ell \neq k$.

Le modèle de Kimura (1980), noté modèle K2 (deux paramètres), introduit le fait que les événements de transition ($A \leftrightarrow G$ et $C \leftrightarrow T$) se produisent plus fréquemment que les événements de transversion (tous les autres événements). Tout comme le modèle JC69, K2 n'est pas réaliste dans le sens où les deux modèles supposent qu'à l'équilibre, tous les nucléotides apparaissent avec la même fréquence, (1/4).

Dans le cas du modèle K2, le terme général de la matrice Q est donné par

$$q_{ij} = \begin{cases} \mu_1 & \text{pour les transitions,} \\ \mu_2 & \text{pour les transversions,} \end{cases}$$

si $i \neq j$.

Felsenstein (1981) a modifié JC96 pour permettre d'avoir des fréquences de base inégales. Ce modèle est généralement noté par F81. Ensuite, Hasegawa *et al.* (1985) ont combiné F81 et K80 afin de permettre une fréquence des bases inégale ainsi qu'un biais transition/transversion. Le modèle est généralement noté par HKY.

Dans le cas du modèle HKY, l'équation (2.20) suggère d'exprimer les transitions en fonction des $\pi_i, i \in E$ qui contrôlent explicitement la distribution stationnaire (Galtier *et al.*, 2005). Ainsi, le terme général de la matrice Q dans le cas du modèle HKY (cinq paramètres) est donné par :

$$q_{ij} = \begin{cases} \mu_1 \pi_j & \text{pour les transitions,} \\ \mu_2 \pi_j & \text{pour les transversions.} \end{cases}$$

2.3.3 Calcul de la probabilité de passage d'une séquence α vers une séquence β

On suppose qu'on a deux séquences $\alpha = (\alpha_\ell)_{1 \leq \ell \leq L}$ et $\beta = (\beta_\ell)_{1 \leq \ell \leq L}$ de longueur L, et que les L nucléotides mutent de manière indépendante les uns des autres. Alors, la probabilité de passage d'une séquence α vers une séquence β sur une période

courte de temps t est obtenue par le produit des probabilités $p_{\alpha_\ell\beta_\ell}(t), \ell = 1, 2, ..., L$, ce qui pourrait être écrit comme suit

$$p_{\alpha\beta}(t) = \prod_{\ell=1}^{L} p_{\alpha_\ell\beta_\ell}(t). \tag{2.21}$$

Les probabilités $p_{\beta_i\alpha_i}(t)$ sont calculées selon le modèle d'évolution d'ADN/ARN retenu, tel que décrit à la section 2.3.2.

Par exemple, si on suppose un modèle de mutation (JC69), la probabilité qu'un nucléotide passe de l'état i vers l'état j en t unités de temps dans le passé est donnée par :

$$p_{ij}(t) = \begin{cases} 0.25\left(1 - e^{-2t\theta_0/3}\right) & \text{si } i \neq j, \\ 0.25\left(1 + 3e^{-2t\theta_0/3}\right) & \text{sinon,} \end{cases} \tag{2.22}$$

tel que $i, j \in \{A, G, T, C\}$, et $\theta_0 = \theta/L$ (Hein *et al.*, 2005).

2.4 Taille de la population effective

Dans cette section, nous introduisons la notion de taille de population effective, qui est un concept central dans notre approche. En effet, cette notion est très importante lorsqu'une des hypothèses du modèle de Wright-Fisher n'est pas vérifiée.

En général, l'évolution d'une vraie population ne se comporte pas selon le modèle de Wright-Fisher, puisque souvent les populations présentent des formes particulières de structure de reproduction due à la proximité géographique des individus (gènes) ou à des contraintes sociales ; par exemple, ce ne sont pas toutes les femmes qui pourraient être mariées à tous les hommes et vice versa.

Il existe plusieurs définitions de la taille de la population effective que l'on dénote, N_e ; voir, par exemple, Ewens (2004). On s'intéresse dans notre cas à la définition de N_e qui nous est utile dans le cas des modèles à taille de population variable. La mesure est appelée taille de la population effective de consanguinité, *inbreeding*

effective population size, et se calcule comme suit :

$$N_e = \frac{1}{2P(T_2 = 1)}, \tag{2.23}$$

où T_2 est le temps de coalescence en présence de deux séquences, qui mesuré en unités de générations (Hein *et al.*, 2005). La généralisation se définit par

$$N_e^{(t)} = \frac{E(T_2)}{2}, \tag{2.24}$$

tel que t représente le temps et T_2 est exprimé en générations.

La différence la plus importante entre les mesures exprimées par les équations (2.23) et (2.24), est que N_e dépend seulement de la génération précédente tandis que $N_e^{(t)}$ dépend du nombre de générations jusqu'à ce que le *MRCA* soit trouvé.

Dans le cas d'un modèle haploïde de Wright-Fisher les deux quantités (N_e et $N_e^{(t)}$) sont équivalentes puisque

$$P(T_2 = 1) = 1/2N \text{ et } E(T_2) = 2N.$$

2.5 Extension du processus de coalescence au cas où la taille de population effective est variable

Le processus de coalescence élémentaire décrit à la section 2, est basé sur le modèle de reproduction de Wright-Fisher qui suppose une taille de population constante, une absence de sélection et de structure de population. Cependant, le cadre de la coalescence peut être généralisé dans le cas où l'une des hypothèses citées ci-haut n'est pas vérifiée. Dans notre cas, on s'intéresse au scénario dans lequel on aurait une fluctuation de nature déterministe de la taille de la population effective. On note dans ce qui suit la taille de la population effective au temps t, par $N_e(t)$ et la taille de la population effective au temps $t = 0$ par $N = N(0)$.

Nous avons vu que la probabilité pour que deux séquences trouvent un ancêtre commun dans la génération précédente dans un processus de coalescence standard à temps discret est $1/2N$.

Dans le cas où la taille de la population effective varie avec le temps, la probabilité que deux séquences coalescent dans la génération précédente est $p(t) = 1/(2N_e(t))$ qui peut différer de $p(0) = 1/(2N_e(0)) = 1/(2N)$. Par exemple, lorsque $N_e(t)$ est inférieure à N de telle sorte que la taille de la population décroît en allant vers le passé, la probabilité d'avoir un événement de coalescence augmente, et l'ancêtre commun, le $MRCA$ pourrait être trouvé plus rapidement que si $N_e(t)$ était constant à travers le temps (Hein *et al.*, 2005).

On définit la quantité $\Lambda(t)$ comme suit :

$$\Lambda(t) = \int_0^t \frac{1}{\nu(u)} du, \tag{2.25}$$

tel que $\nu(t) = N_e(t)/N$, représente la taille relative de $N_e(t)$ par rapport à N, et $\Lambda(t)$ le taux de coalescence cumulé au cours du temps relativement au taux au temps $t = 0$, avec l'hypothèse $\Lambda(\infty) = \infty$ pour assurer l'existence d'un $MRCA$ pour l'échantillon de séquences. Soit $A_n(t)$, le nombre d'ancêtres distincts d'un échantillon de séquences de taille n, après t unités de temps, dans le cas d'un processus de coalescence avec une taille de population constante. $\{A_n(t), t \geq 0\}$ est un processus de mort pur, où le nombre de lignées ancestrales passe de k vers $(k-1)$ lignées avec un taux $k(k-1)/2$ (Griffiths et Tavaré, 1994a).

De manière équivalente, dans le cas d'une taille de population variable, le nombre d'ancêtres distincts d'un échantillon de séquences de taille n, après t unités de temps, est donné par $\tilde{A}_n(t)$. Dans ce cas, $\{\tilde{A}_n(t), t \geq 0\}$ est un processus de mort non homogène où $\tilde{A}_n(t)$ peut être écrit en fonction de $A_n(t)$ comme suit :

$$\tilde{A}_n(t) = A_n(\Lambda(t)), t \geq 0. \tag{2.26}$$

Ainsi, si la taille de la population diminue avec le temps (du présent vers le passé), alors $\nu(t) \leq 1$ et $\Lambda(t) \geq t$ et, par conséquent,

$$\tilde{A_n}(t) \leq A_n(t),$$

et

$$\Pr(\tilde{A_n}(t) > k) \leq \Pr(A_n(t) > k).$$

On pourrait alors déduire que :

- le temps nécessaire pour trouver l'ancêtre commun dans une petite population est inférieur au cas d'une grande population ;
- la topologie d'un arbre est la même que dans le cas d'une taille de population constante ; par contre l'échelle du temps doit être changée afin de prendre en considération la fluctuation dans la taille de la population effective.

2.5.1 Loi de probabilité des temps d'attente dans le cas où la taille de la population effective est variable

Soit T_k le temps d'attente jusqu'au prochain événement de coalescence en présence de k séquences tel que $k = 2, ..., n$, et soit $V_k = T_n + ... + T_k$ le temps d'attente cumulé à partir des séquences échantillonnées jusqu'au moment où on a $(k-1)$ ancêtres.

La loi de probabilité de T_k conditionnellement à $V_{k+1} = v_{k+1}$ est donnée par (Hein *et al.*, 2005)

$$P(T_k > t | V_{k+1} = v_{k+1}) = \exp\left\{-\binom{k}{2}(\Lambda(t + v_{k+1}) - \Lambda(v_{k+1}))\right\}, \qquad (2.27)$$

avec $v_{n+1} = 0$.

On note que si on remplace $\Lambda(t)$ par t dans l'équation (2.27), on retrouve la densité de probabilité d'une loi exponentielle. Afin de simuler T_k, on a besoin de faire la

distinction entre les temps du processus de coalescence de base, où l'on suppose que la taille de la population est constante, et les temps de coalescence dans le cas où la taille de la population est variable.

On notera dans ce qui suit les temps du processus de coalescence de base par T_k^*. L'algorithme suivant permet de simuler T_k.

2.5.2 Algorithme de simulation des temps de coalescence dans le cas où la taille de population est variable, et application au cas exponentiel

1. On simule $T_2^*, ..., T_n^*$ selon un processus de coalescence standard (taille de population constante), où T_k^* est distribué selon une loi exponentielle de paramètre $\binom{k}{2}$. Les valeurs simulées vont être notées t_k^* ;

2. Résoudre $\Lambda(t_k + v_{k+1}) - \Lambda(v_{k+1}) = t_k^*$ pour $v_k, k = 2, ..., n$ et $v_{n+1} = 0$;

3. Les valeurs $t_k = v_k - v_{k+1}$ sont les réalisations du processus $T_2, ..., T_n$ décrit par l'équation (2.27).

Parfois, l'équation en v_k peut être résolue explicitement ; sinon, il faut faire appel à des algorithmes de calcul numérique.

Pour illustrer, considèrons le cas d'un modèle de croissance exponentielle de la population. Ainsi, on suppose que le taux de croissance instantané est proportionnel à la taille de la population courante,

$$N_e(t) = Ne^{-\beta t}, \tag{2.28}$$

où t est en unités de $2N$ générations.

Le coefficient β apparaît avec un signe négatif car le temps est mesuré de manière rétrospective. Dans le cas où la population croît de manière exponentielle (ou

plutôt décroît de manière exponentielle vers la passé). On note que les quantités $\nu(t)$, $\Lambda(t)$, ainsi que la loi des temps d'attente T_k se définissent comme suit :

$$\nu(t) = e^{-\beta t},$$

$$\Lambda(t) = \frac{1}{\beta}(e^{\beta t} - 1),$$

et

$$P(T_k > t | V_{k+1} = v_{k+1}) = \exp\left\{-\binom{k}{2}\frac{1}{\beta}e^{\beta v_{k+1}}(e^{\beta t} - 1)\right\}. \tag{2.29}$$

Alors, dans l'algorithme défini à la section 2.5.2, t_k est donné par

$$t_k = \frac{1}{\beta}\log(1 + \beta t_k^* e^{-\beta v_{k+1}}), \qquad k = 2, 3, ..., n, \tag{2.30}$$

où t_k^* est de loi exponentielle de paramètre $\binom{k}{2}$. Ainsi, pour un processus de coalescence avec une croissance exponentielle de la taille de la population effective de paramètre β, il est possible de calculer explicitement t_k en utilisant l'équation (2.30) à partir de t_k^* et de $v_{k+1} = \sum_{i=k+1}^{n} t_i$.

Notons que les temps d'attente $T_2, ..., T_n$ ne sont plus indépendants comme dans le cas d'un processus de coalescence de base, mais sont négativement corrélés : si l'un d'entre eux est plus grand, les autres temps d'attente sont susceptibles d'être petits parce que la variation dans le temps jusqu'au $MRCA$ est beaucoup plus petite par rapport au processus de coalescence de base.

2.6 Phylogénétique et théorie de la coalescence : deux méthodologies différentes

Plusieurs méthodes non paramétriques d'estimation de la taille de la population effective font appel à la théorie phylogénétique, une approche différente de la coalescence. Ainsi, on introduit brièvement la phylogénétique afin de mettre en évidence les différences avec la théorie de la coalescence.

Il est parfois difficile de faire la distinction entre les méthodes basées sur la phylogénétique de celles qui utilisent la théorie de coalescence, puisque les deux méthodes utilisent des arbres. Les différences entre la phylogénétique et la théorie de coalescence, sont mises en évidence dans ce qui suit.

- Méthodes phylogénétiques

  Les méthodes phylogénétiques estiment des arbres. Elles ont été développées pour déterminer le lien existant entre des espèces apparentées en supposant qu'elles sont les descendantes d'un ancêtre commun, et que la relation est sous forme d'un arbre (Rosenberg et Nordborg, 2002).

  On utilise en général une seule séquence provenant de chaque espèce et l'arbre généalogique estimé (*phylogénie*) est utilisé pour conclure sur la relation qui existe entre les différentes espèces.

  On confond souvent l'arbre de gènes (*Gene tree*) avec l'arbre des espèces (*Species tree)*, ce qui pourrait être expliqué par la forte « corrélation » entre l'arbre de gènes et l'arbre des espèces. Cependant, ce rapprochement n'est pas valide dans le cas d'un scénario démographique complexe. Dans ce cas, les conclusions tirées à partir de l'arbre des espèces ne pourraient pas être basées sur l'estimation d'un arbre de gènes - différents gènes peuvent entrainer la production de différents arbres.

- Méthodes basées sur la théorie de la coalescence

  Les méthodes basée sur la théorie de coalescence n'estiment pas des arbres. Néanmoins, elles sont utilisées pour estimer des paramètres d'un processus généalogique aléatoire, en considérant l'arbre généalogique comme un

paramètre de nuisance qui n'a pas d'intérêt en soi. De plus, les méthodes généalogiques basées sur la théorie de coalescence n'ont pas les limitations des méthodes phylogénétiques. En effet, contrairement aux méthodes phylogénétiques, la théorie de coalescence représente un cadre théorique cohérent qui considère la recombinaison, la migration, la sélection, ainsi que d'autres processus.

- Différence entre les méthodes phylogénétiques et les méthodes basées sur la théorie de coalescence

On illustre la différence entre l'utilisation des méthodes phylogénétiques par rapport aux méthodes basées sur la théorie de coalescence dans le cadre du calcul de vraisemblance comme présenté dans Rosenberg et Nordborg (2002). L'équation fondamentale pour l'inférence de vraisemblance en phylogénétique est

$$L(\mathcal{G}, \mu) = P(\mathcal{D}|\mathcal{G}, \mu), \tag{2.31}$$

tel que $\mathcal{D}$ représente les données de séquences d'ADN, $\mathcal{G}$ l'arbre phylogénique, et μ englobe l'ensemble des paramètres du modèle de mutation. Dans ce cas, l'objectif est l'estimation du paramètre $\mathcal{G}$.

En coalescence, l'équation de vraisemblance des données $\mathcal{D}$ s'écrit comme suit :

$$L(\mu, \alpha) = \sum_{\mathcal{G}} P(\mathcal{D}|\mathcal{G}, \mu) P(\mathcal{G}|\mu, \alpha), \tag{2.32}$$

où α englobe l'ensemble des paramètres, comme la taille de la population, le taux de migration, etc. L'objectif dans ce cas est l'estimation de α où $\mathcal{G}$ est considéré comme un paramètre de nuisance sur lequel une espérance est estimée sur l'ensemble des généalogies possibles.

Finalement, autant dans leurs objectifs que dans leurs modèles mathématiques, la théorie de la coalescence et la phylogénétique sont deux approches fondamentalement différentes.

Le chapitre qui suit, décrit l'inférence qui fait appel à la fonction de vraisemblance dans la théorie de coalescence, ce qui est à la base de la présente recherche.

CHAPITRE III

VRAISEMBLANCE ET ÉCHANTILLONNAGE PONDÉRÉ

Dans ce chapitre, on présente les méthodes qui permettent d'approximer la vraisemblance $L(\theta)$ d'un ensemble de données $\mathcal{D}$ dont l'évolution dépend d'un paramètre θ, en se basant sur la théorie de la coalescence. En général, on ne connait pas l'expression explicite de la vraisemblance complète $L(\theta)$ d'un échantillon de séquences. En effet, la vraisemblance s'écrit en intégrant sur toutes les généalogies possibles :

$$L(\theta) = P(\mathcal{D}|\theta) = \int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G},\theta)P(\mathcal{G}|\theta)d\mathcal{G}. \tag{3.1}$$

On note que, dans (3.1), l'intégration est réalisée sur l'ensemble de toutes les généalogies possibles $\mathcal{G}$, où chacune des généalogies peut être s'écrire comme $\mathcal{G} = (\mathcal{T},\mathcal{W})$, tel que $\mathcal{T}$ représente la topologie de coalescence (ordre de branchement) et $\mathcal{W}$ est l'ensemble des intervalles de temps écoulés entre les événements de coalescence.

Il est facile de calculer $P(\mathcal{D}|\mathcal{G},\theta)$ pour une topologie et un ensemble de temps de coalescence donnés. Par contre, intégrer $P(\mathcal{D}|\mathcal{G},\theta)$ sur toutes les topologies possibles n'est réalisable que pour des tailles d'échantillons réduites puisqu'il y a $n!(n-1)!/2^{n-1}$ différentes topologies de coalescence sur lesquelles on devrait sommer.

La méthode de Monte Carlo dite naïve pour estimer (3.1) est basée sur l'idée que

si une variable aléatoire X est de densité de probabilité $f_X(x)$, alors l'espérance pour n'importe quelle fonction de X, $g(X)$, pourrait être approximée en simulant plusieurs valeurs $X^{(1)}, X^{(2)}, ..., X^{(J)}$ à partir de la densité de probabilité $f_X(\cdot)$, et calculer la moyenne sur ces valeurs, ce qui donne

$$E(g(X)) = \int g(x) f_X(x) dx \approx \frac{1}{J} \sum_{j=1}^{J} g(X^{(j)}). \tag{3.2}$$

Sous certaines conditions de régularité sur la fonction $g(\cdot)$, cette approximation devrait être bonne pour un J suffisamment grand ; théoriquement, l'erreur tend vers zéro lorsque J tend vers l'infini. Si on applique ce type d'approximation à l'intégrale (3.1) on a

$$L(\theta) = \int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G}, \theta) P(\mathcal{G}|\theta) d\mathcal{G} \approx \frac{1}{J} \sum_{j=1}^{J} P(\mathcal{D}|\mathcal{G}^{(j)}, \theta), \tag{3.3}$$

avec $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)} \sim P(\mathcal{G}|\theta)$.

L'approximation (3.3) est facile à implémenter, puisque $P(\mathcal{D}|\mathcal{G}, \theta)$ peut être calculée pour une généalogie $\mathcal{G}$ reliant les séquences échantillonnées connues en utilisant le *peeling algorithm* de Felsenstein (1981). Malheureusement, l'approximation n'est pas utilisable pour de grandes tailles d'échantillon de séquences. La raison est que seulement quelques généalogies vont contribuer de manière importante à la somme de l'équation (3.3), et le temps de calcul va être gaspillé en incluant des généalogies qui ont une contribution négligeable à la vraisemblance.

Afin de répondre à cette difficulté, on peut utiliser l'échantillonnage pondéré, une approche qui est une partie intégrante de notre méthodologie, et qui est décrite dans la section suivante.

3.1 Échantillonnage pondéré

L'échantillonnage pondéré est une méthode statistique standard qui permet de réduire la variance dans la méthode de Monte Carlo décrite par l'equation (3.2). Dans le cas de l'équation (3.3), cela revient à limiter le nombre de simulations de généalogies qui ne contribuent pas beaucoup à la vraisemblance. Ainsi, au lieu de choisir les généalogies à partir de la loi de probabilité $P(\mathcal{G}|\theta)$, on simule à partir d'une loi de probabilité $Q(\mathcal{G})$, ce qui permet d'avoir des généalogies plus compatibles avec les données $\mathcal{D}$. Pour ce faire, on réécrit l'équation (3.1) sous la forme

$$\int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G},\theta)\frac{P(\mathcal{G}|\theta)}{Q(\mathcal{G})}Q(\mathcal{G})d\mathcal{G}, \tag{3.4}$$

où $Q(\mathcal{G})$ s'appelle *distribution proposée* ou encore *distribution de l'échantillonnage pondéré*, qui peut être en principe n'importe quelle loi de probabilité sur $\mathcal{G}$. Une simple approximation de Monte Carlo de (3.4) donne

$$L(\theta) \approx \frac{1}{J}\sum_{j=1}^{J} P(\mathcal{D}|\mathcal{G}^{(j)},\theta)\frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}, \tag{3.5}$$

où $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)} \sim Q(\mathcal{G})$.

Un bon choix de $Q(\cdot)$ permet de faire en sorte que l'approximation (3.5) soit beaucoup plus efficace que celle donnée par l'équation (3.3). Idéalement, on voudrait échantillonner à partir de $Q(\mathcal{G}) = \widetilde{Q}_\theta(\mathcal{G}) = P(\mathcal{G}|D,\theta)$ qui satisfait

$$\widetilde{Q}_\theta(\mathcal{G}) = P(\mathcal{G}|\mathcal{D},\theta) = \frac{P(\mathcal{G}|\theta)P(\mathcal{D}|\mathcal{G},\theta)}{P(\mathcal{D}|\theta)} = \frac{P(\mathcal{G},\mathcal{D}|\theta)}{P(\mathcal{D}|\theta)}. \tag{3.6}$$

Dans le cas où $Q(\mathcal{G}) = \widetilde{Q}_\theta(\mathcal{G})$, l'approximation (3.5) devient exacte :

$$\sum_{j=1}^{J} P(D|\mathcal{G}^{(j)},\theta)\frac{P(\mathcal{G}^{(j)}|\theta)}{\widetilde{Q}_\theta(\mathcal{G}^{(j)})} = \frac{1}{J}\sum_{j=1}^{J} L(\theta) = L(\theta). \tag{3.7}$$

On remarque que la connaissance de $\widetilde{Q}_\theta$ pour chaque généalogie $\mathcal{G}$ est équivalente à la connaissance de $L(\theta)$, ce qui rend cette approche non réaliste.

3.2 Surfaces de vraisemblance

On peut utiliser l'équation (3.5) afin d'approximer la vraisemblance $L(\theta)$ pour un ensemble de valeurs $\theta \in \{\theta_1, \theta_2, ..., \theta_R\}$ en utilisant des généalogies générées à partir d'une distribution $Q(\mathcal{G})$. Ainsi, la valeur qui maximise la probabilité $P(\mathcal{D}|\theta) = L(\theta)$ représente l'estimateur du maximum de vraisemblance de θ. Cependant, la fonction de l'échantillonnage pondéré optimale (3.6) dépend de θ. Ainsi, il n'existe pas de choix optimal de $Q(\cdot)$ pour toutes les valeurs de θ simultanément, ce qui fait que l'efficacité d'un estimateur de $L(\theta)$ peut varier en fonction de θ.

Une stratégie adoptée par Griffiths et Tavaré (1994a) commence par construire une fonction d'échantillonnage pondéré $Q_{\theta_0}(\cdot)$ pour approximer $\widetilde{Q}_{\theta_0}(\cdot)$ pour une certaine valeur θ_0 en utilisant (3.5). Ensuite, la courbe de vraisemblance $L(\theta)$ est estimée pour toutes les valeurs de θ dans un voisinage de θ_0. On dit dans ce cas, qu'on utilise la valeur θ_0 comme une *valeur conductrice* pour θ. Stephens et Donnelly (2000) ont démontré que ce type de méthode *à valeur conductrice* sous-estime la vraisemblance pour des valeurs de θ éloignées de θ_0 et aurait tendance à donner une valeur du maximum de la surface de vraisemblance proche de θ_0. Afin d'éviter ce genre de problème, on pourrait utiliser une fonction d'échantillonnage pondérée différente pour chaque valeur de θ. Cela revient à utiliser une fonction d'échantillonnage pondéré Q_{θ_i} différente pour chaque valeur θ_i, afin d'estimer la vraisemblance $L(\theta_i)$ sur une grille de valeurs, $\{\theta_1, \theta_2, ..., \theta_R\}$. L'utilisation de bonnes approximations $Q_{\theta_i}(\cdot)$ des fonctions optimales $\widetilde{Q}_{\theta_i}(\cdot)$ permet d'avoir de bons estimés pour toutes les valeurs de θ. On appellera cette approche *approche ponctuelle*.

Finalement, Stephens (2001) note qu'une méthode plus efficace serait de combiner l'approche ponctuelle et l'approche basée sur une *valeur conductrice*, ce qui revient

à utiliser un échantillon à partir d'une seule fonction d'échantillonnage pondéré,

$$Q(\mathcal{G}) = \frac{1}{R}\sum_{i=1}^{R} Q_{\theta_i}(\mathcal{G}), \tag{3.8}$$

afin d'estimer la surface de vraisemblance pour une grille de valeurs de θ. Cette approche efficace doit être implémentée ; si ce n'est pas le cas, il est conseillé de comparer les estimés des surfaces de vraisemblance en utilisant plusieurs valeurs conductrices pour vérifier si cela ne cause pas de problème. S'il s'avère que les surfaces obtenues sont très différentes, l'approche ponctuelle est recommandée (Stephens, 2001).

3.3 Mesures de performance de l'échantillonnage pondéré

Dans cette section, on décrit une mesure de performance de l'échantillonnage pondéré appelée taille de l'échantillon effective, qu'on notera ESS, une notation qui provient de l'anglais *effective sample size.*

L'utilisation de l'échantillonnage pondéré permet d'utiliser ce genre de mesure de performance, et de tirer profit de l'indépendance de nos échantillons afin de juger de la précision des estimateurs, contrairement aux méthodes MCMC, comme le remarquent Fearnhead et Donnelly (2001). Dans notre contexte, l'échantillonnage pondéré utilise des observations (généalogies) pondérées, où les poids sont donnés par

$$W_j = \frac{P(\mathcal{G}^{(j)})}{Q(\mathcal{G}^{(j)})}, j = 1, 2, ..., J,$$

et

$$w_j = \frac{W_j}{\sum_{j=1}^{J} W_j},$$

tel que J représente le nombre de généalogies simulées.

Il y a deux cas extrêmes possibles :

- un premier cas où il y a seulement un poids w_{i_0} qui est beaucoup plus grand que les autres, ce qui revient à utiliser seulement une seule généalogie ;
- un deuxième cas extrême, où on a $w_1 = w_2 = w_3.... = w_J \approx 0$, et il est évident que notre échantillonnage pondéré n'a pas fonctionné, car on ne différencie plus les généalogies.

Afin de justifier ce concept, soit la combinaison linéaire hypothétique suivante :

$$S_w = \frac{\sum_{j=1}^{J} w_j Z_j}{\sum_{j=1}^{J} w_j}, \tag{3.9}$$

où $Z_j, j = 1, 2, ..., J$ sont des variables aléatoires identiquement distribuées de variance $\sigma^2 > 0$ et $w_j \in [0, \infty)$.

Dans ce qui suit, on cherche à déterminer le nombre n_e d'observations indépendantes nécessaires pour que la variance de la moyenne pondérée décrite dans (3.9) soit équivalente à celle d'une moyenne arithmétique de n_e observations. Soit $\overline{Z}$ la moyenne de n_e variables aléatoires Z_j. La variance de $\overline{Z}$ est donnée par

$$Var(\overline{Z}) = \frac{\sigma^2}{n_e}, \tag{3.10}$$

et celle de S_w est ,

$$Var(S_w) = \frac{1}{(\sum_j w_j)^2} \sum_j w_j^2 \sigma^2. \tag{3.11}$$

En utilisant (3.10) et (3.11), on aura (Owen, 2009) :

$$n_e = \frac{(\sum_{j=1}^{J} w_j)^2}{\sum_{j=1}^{J} w_j^2} = \frac{J\overline{w}^2}{\overline{w^2}}, \tag{3.12}$$

où

$$\overline{w} = \frac{1}{J} \sum_{j=1}^{J} w_j, \text{ et } \overline{w^2} = \frac{1}{J} \sum_{j=1}^{J} w_j^2.$$

On voit bien que $n_e \leq J$, avec égalité si $w_j = \frac{1}{j}, j = 1, 2, ..., J$. De plus, le cas où les poids w_i sont très dispersés, revient à faire une moyenne seulement sur $n_e \ll J$.

On note qu'il est aussi possible d'exprimer la taille d'échantillon effective n_e en fonction du coefficient de variation de w. En effet, l'équation (3.12) peut être réécrite comme suit

$$n_e = \frac{J}{1 + CV(w)^2}, \tag{3.13}$$

tel que

$$CV(w) = \frac{1}{\overline{w}} \sqrt{\frac{1}{n-1} \sum_{j=1}^{J} (w_j - \overline{w})^2}, \tag{3.14}$$

représente le coefficient de variation des poids.

Pour finir cette section, on note que l'utilisation de ESS comme mesure d'efficacité de l'échantillonnage pondéré n'est pas parfaite, puisque :

- lorsque la mesure ESS est trop petite, on peut dire que les poids de l'échantillonnage pondéré sont problématiques ;
- par contre, lorsque la mesure ESS est élevée, on ne peut pas conclure que l'échantillonnage pondéré fonctionne car il est possible que la région explorée par les généalogies $(\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)})$ ne soit pas suffisante pour couvrir l'espace en entier.

3.4 Distribution proposée par Stephens et Donnelly (2000)

On avait mentionné à la section 3.1 qu'une bonne idée serait de choisir $Q(\cdot)$ proche de $\widetilde{Q}_\theta = P(\mathcal{G}|\mathcal{D}, \theta)$. Cette stratégie a été utilisée par Stephens et Donnelly (2000), qui ont proposé un échantillonnage pondéré efficace dans plusieurs cas. Cette méthode est à la base de notre méthodologie et on la résume dans ce qui suit.

Ainsi, on suppose que chaque séquence possède un *type* génétique associé, et on note l'ensemble de tous les types possibles par E, que l'on suppose dénombrable dans le cas d'un modèle à sites finis.

Indépendamment de tous les autres événements, une séquence de type $\alpha \in E$ peut soit muter vers une séquence de type $\beta \in E$ avec une probabilité $\mu P_{\alpha\beta}$, ou ne subir aucune mutation avec une probabilité de $1 - \mu$, où

- μ : le taux de mutation par séquence par génération selon une chaîne de Markov avec la matrice de transition $\mathbf{P}$ pour un modèle de substitution donné (voir section 2.3.2) ;
- $P_{\alpha\beta}$: la probabilité de mutation d'une séquence de type α vers une séquence de type β tel que décrit à la section 2.3.3.

On suppose que l'évolution est neutre, ce qui veut dire que la démographie est indépendante des types génétiques des séquences.

On définit l'historique $\mathcal{H}$ comme étant l'ensemble de tous les états $(H_{-m}, ..., H_{-1}, H_0)$ visités par le processus de Markov qui commence par le type génétique H_{-m} du $MRCA$ et se termine avec les types génétiques $H_0 \in E^n$, au temps présent ; m est de fait aléatoire.
La transition de l'état H_{i-1} vers l'état H_i est obtenue :

- soit par une mutation d'une séquence de type α vers le type β, événement qu'on note M_α^β ;
- soit par une division de la lignée de type α, événement qu'on note $\mathcal{C}_\alpha^\alpha$.

Les probabilités de transition de la chaîne de Markov qui décrit l'évolution du passé vers le présent, sont données par

$$p_\theta(H_i|H_{i-1}) = \begin{cases} \dfrac{n_{i-1}^{(\alpha)}}{n_{i-1}} \dfrac{\theta}{(n_{i-1}-1+\theta)} P_{\alpha\beta} & \text{si } M_\alpha^\beta, \\ \dfrac{n_{i-1}^{(\alpha)}}{n_{i-1}} \dfrac{n_{i-1}-1}{n_{i-1}-1+\theta} & \text{si } \mathcal{C}_\alpha^\alpha, \\ 0 & \text{sinon,} \end{cases} \tag{3.15}$$

où $n_{i-1} \geq 2$ représente le nombre de séquences dans H_{i-1} et $n_{i-1}^{(\alpha)}$ est le nombre de séquences de type α dans H_{i-1}.

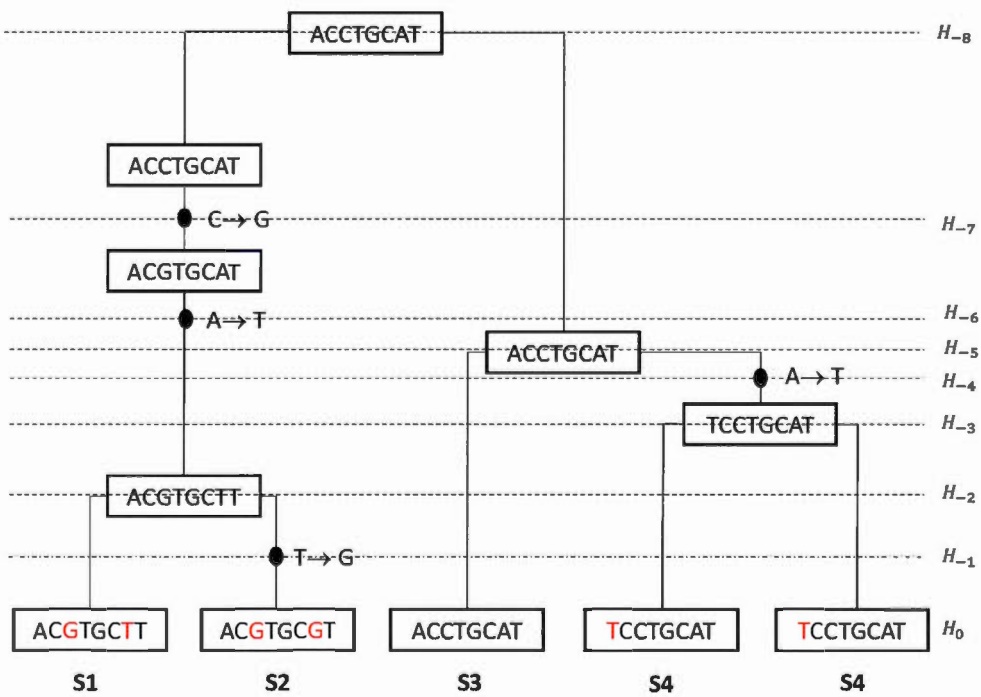


Figure 3.1: Illustration d'un arbre généalogique $\mathcal{G}$.

La figure 3.1 illustre un arbre généalogique $\mathcal{G}$ correspondant à un historique $\mathcal{H}$ pour cinq individus de types génétiques dans E={S1, S2, S3, S4, S5}, tel que S1= {ACGTGCTT}, S2={ACGTGCGT}, S3={ACCTGCAT}, S4= {TCCTG-CAT} et S5={ACGTGCAT}.

Si on décrit l'évolution des séquences de la figure 3.1 du présent vers le passé, on voit que l'on dispose au temps présent d'un ensemble H_0, constitué de cinq séquences de types génétiques $\{S1, S2, S3, S4, S4\}$. Le premier événement observé est une mutation, où la séquence de type S_2 mute vers une séquence de type S_1 ; ainsi, on obtient l'ensemble de séquences H_{-1} qui est constitué de 5 séquences de types {S1, S1, S3, S4, S4} ; ensuite, à partir de l'état H_{-1}, on observe que le deuxième événement observé est une coalescence du fait que les deux séquences de type S_1 coalescent, ce qui donne l'état H_{-2} qui est constitué de 4 séquences de types {S1, S3, S4, S4}. Le processus continue ainsi jusqu'au moment où on arrive à un état H_{-m} constitué d'une seule séquence. Finalement, on note que dans le cas de l'historique $\mathcal{H}$ de la figure 3.1, il y a eu quatre coalescences et quatre mutations, ce qui donne $m = 8$.

Ainsi, en énumérant du passé vers le présent, l'historique $\mathcal{H} = \{H_{-m}, H_{-m+1},$ ..., $H_{-1}, H_0\}$ peut être représenté par {{S3}, {S3, S3}, {S5, S3}, {S1, S3}, {S1, S3, S3},{S1, S3, S4}, {S1, S3, S4, S4}, {S1, S1, S3, S4, S4} , {S1, S2, S3, S4, S4}}.

L'ensemble de l'information disponible dans la figure 3.1 est notée $\mathcal{A}$, qu'on appelle *le type ancestral.* Par ailleurs, à l'état stationnaire et sous l'hypothèse de neutralité, la loi de probabilité du type du $MRCA$ est donnée par la loi stationnaire de la chaîne de Markov des mutations, $P_{\alpha\beta}$ (Stephens et Donnelly, 2000).

Dans la section suivante, on décrit la méthode de $S\&D$ (Stephens et Donnelly, 2000) qui s'intéresse à l'inférence du paramètre inconnu θ, en supposant que la matrice de transition $\mathbf{P}$ est connue.

3.4.1 Distribution proposée de Stephens et Donnelly (2000) : description de la méthode

Une classe naturelle de distributions proposées se fait en considérant une reconstruction aléatoire d'historiques de manière rétrospective, en utilisant le principe de Markov en partant d'un échantillon $\mathcal{D}$ constitué de n séquences, vers le $MRCA$. Ainsi, un historique aléatoire $\mathcal{H} = (H_{-m}, ..., H_{-1}, H_0)$ peut être reconstruit en commençant par $H_0 = \mathcal{D}$, et en choisissant H_{i-1}, $i-1 = -1, -2, ..., -m$ selon les probabilités de transition $q_\theta(H_{i-1}|H_i)$. Le processus s'arrête au premier instant pour lequel H_{-m} devient un singleton.

Pour utiliser l'équation (3.4), il est nécessaire de s'intéresser à une sous classe $\mathcal{M}$ de distributions proposées $q_\theta(.|H_i)$, telles que, pour chaque i, le support de $q_\theta(.|H_i)$ est l'ensemble

$$\{H_{i-1} : p_\theta(H_i|H_{i-1}) > 0\},$$

où p_θ représente la probabilité de transition du passé vers le présent.
En spécifiant les probabilités de transition vers le passé q_θ, on peut définir la loi de probabilité $Q_\theta(\cdot)$ ayant comme support l'ensemble de toutes les généalogies compatibles avec $\mathcal{D}$. De plus, il est simple de simuler des généalogies à partir de Q_θ et d'évaluer le rapport $P_\theta(\mathcal{H})/Q_\theta(\mathcal{H})$ afin d'appliquer ensuite l'approximation (3.5). Sous les conditions précédentes, Stephens et Donnelly (2000) ont prouvé le résultat suivant concernant le choix optimal $Q^*_\theta(\cdot)$ dans la classe $\mathcal{M}$.

Théorème 1 (Stephens et Donnelly, 2000)

Soit la loi conditionnelle du type de la $(n+1)^{ième}$ séquence échantillonnée α sachant $\mathcal{D}$,

$$\pi(\alpha|\mathcal{D}) = \frac{\pi_\theta(\mathcal{D}, \alpha)}{\pi_\theta(\mathcal{D})}, \tag{3.16}$$

tel que $\mathcal{D}$ représente les n séquences déjà échantillonnées auparavant. Alors, la

distribution proposée optimale Q_θ^* dans la classe $\mathcal{M}$ est donnée par

$$q_\theta^*(H_{i-1}|H_i) = \begin{cases} \dfrac{1}{C} \cdot \dfrac{\theta}{2} n_i^{(\alpha)} \times \dfrac{\pi(\beta|H_i - \alpha)}{\pi(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{si } M_\alpha^\beta, \\ \dfrac{1}{C} \cdot \binom{n_i^{(\alpha)}}{2} \times \dfrac{1}{\pi(\alpha|H_i - \alpha)} & \text{si } \mathcal{C}_\alpha^\alpha, \\ 0 & \text{sinon,} \end{cases} \tag{3.17}$$

tel que :

- n_i et $n_i^{(\alpha)}$ représentent, respectivement, le nombre global de séquences et le nombre de séquences de type α dans H_i ;
- C est une constante de proportionnalité donnée par

$$C = \frac{n_i(n_i - 1 + \theta)}{2};$$

- $H_i - \alpha$ représente l'ensemble des séquences dans H_i mis à part la séquence choisie de type α. ■

On a deux conséquences du théorème précédent :

1. pour n'importe quelle distribution proposée $Q_\theta \in \mathcal{M}$, comme celle proposée par Griffiths et Tavaré (1994c), on peut évaluer q_θ, en utilisant l'équation (3.17) afin de savoir si Q_θ se comporte bien ;
2. on peut espérer construire des distributions proposées qui se comportent bien, en développant de bonnes approximations de $\pi(\cdot|\cdot)$ et en les insérant dans l'équation (3.17).

Stephens et Donnelly (2000) proposent d'estimer les probabilités d'échantillonnage $\pi(\cdot|\mathcal{D})$ par $\hat{\pi}(\cdot|\mathcal{D})$ comme suit : on commence par choisir aléatoirement un individu

α dans E à partir de $\mathcal{D}$; ensuite, on fait muter α, m fois selon la matrice de mutation $\mathbf{P}_{\alpha\beta}$, où m est simulé selon une loi géométrique de paramètre $\theta/(n+\theta)$, ce qui donne

$$\hat{\pi}(\beta|\mathcal{D}) = \sum_{\alpha\in E}\sum_{m=0}^{\infty} \frac{n_\alpha}{n}\left(\frac{\theta}{n+\theta}\right)^m \frac{n}{n+\theta}(\mathbf{P}^m)_{\alpha\beta}. \tag{3.18}$$

De plus, Stephens et Donnelly (2000) prouvent les propriétés suivantes :

(a) À partir de l'équation (3.18), $\hat{\pi}(\cdot|\mathcal{D})$ peut être écrit de façon matricielle

$$\hat{\pi}(\beta|\mathcal{D}) = \sum_{\alpha\in E} \frac{n_\alpha}{n}\mathbf{M}^{(n)}_{\alpha\beta}, \tag{3.19}$$

pour une certaine matrice $\mathbf{M}^{(n)}$, telle que

$$\mathbf{M}^{(n)} = (1-\lambda_n)(I-\lambda_n\mathbf{P}_{\alpha\beta})^{-1}, \tag{3.20}$$

et $\lambda_n = \frac{\theta}{n+\theta}$. Par conséquent, une séquence β peut être échantillonnée à partir de $\hat{\pi}(\cdot|\mathcal{D})$ en commençant par choisir aléatoirement une séquence dans $\mathcal{D}$, puis choisir β à partir d'une distribution qui dépend seulement de la taille d'échantillon n, et du type de la séquence sélectionnée.

(b) La loi $\hat{\pi}(\cdot|\mathcal{D})$ est une loi stationnaire pour la chaîne de Markov sur E ayant comme matrice de transition

$$\mathbf{M}_{\beta\alpha} = \frac{\theta}{n+\theta}\mathbf{P}_{\beta\alpha} + \frac{n_\alpha}{n+\theta}, \tag{3.21}$$

car $\hat{\pi}(\alpha|\mathcal{D})$ satisfait

$$\hat{\pi}(\alpha|\mathcal{D}) = \sum_{\beta\in E}\hat{\pi}(\beta|\mathcal{D})\left(\frac{\theta}{n+\theta}\mathbf{P}_{\beta\alpha} + \frac{n_\alpha}{n+\theta}\right). \tag{3.22}$$

La distribution $Q_\theta^{SD} \in \mathcal{M}$ proposée par Stephens et Donnelly (2000), correspond aux probabilités de transition vers le passé $\hat{q}_\theta$ obtenues en insérant $\hat{\pi}(\cdot|\cdot)$ à la place de $\pi(\cdot|\cdot)$ dans l'équation (3.17) :

$$\hat{q}_\theta(H_{i-1}|H_i) = \begin{cases} \dfrac{1}{C} \cdot \dfrac{\theta}{2} \times n_i^{(\alpha)} \times \dfrac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} \times P_{\beta\alpha} & \text{si } M^{\beta\alpha}, \\ \dfrac{1}{C} \cdot \binom{n_i^{(\alpha)}}{2} \times \dfrac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{si } C_\alpha^\alpha, \\ 0 & \text{sinon,} \end{cases} \tag{3.23}$$

où $n_i^{(\alpha)}$ est le nombre de séquences de type α dans H_i, n_i est le nombre total de séquences dans H_i, et $C = n_i(n_i - 1 + \theta)/2$.

3.4.2 Application de la méthode de Stephens et Donnelly (2000)

Pour un H_i donné, les probabilités de transition vers le passé $\hat{q}_\theta(H_{i-1}|H_i)$ définies par l'équation (3.23), et dont la somme est égale à $(n_i^{(\alpha)}/n)$, peuvent être obtenues comme suit :

- on choisit une séquence aléatoirement à partir de H_i, et on note le type de la séquence choisie α ;
- pour chaque type $\beta \in E$ pour lequel $P_{\beta\alpha} > 0$, on calcule $\hat{\pi}(\beta|H_i - \alpha)$ à partir de l'équation (3.19) ;
- on échantillonne H_{i-1} comme suit :

$$H_{i-1} = \begin{cases} H_i - \alpha + \beta & \text{proportionnellement à } \theta\hat{\pi}(\beta|H_i - \alpha)\mathbf{P}_{\beta\alpha} \\ H_i - \alpha & \text{proportionnellement à } n_i^{(\alpha)} - 1. \end{cases} \tag{3.24}$$

Par conséquent, la génération des topologies en appliquant (3.24) s'appuie sur le calcul des quantités $P_{\beta\alpha}$ et $\hat{\pi}(\beta|H_i - \alpha)$.

On note que le bloc d'équations (3.24) est une version simplifiée du bloc d'équations (3.23). En effet, à partir de l'équation (3.23), une séquence de type α mute vers une séquence β avec une probabilité $p_m(\alpha)$, et coalesce avec une autre séquence de type α avec la probabilité $p_c(\alpha)$, de la manière suivante :

$$p_m(\alpha) = \frac{1}{C} \cdot \frac{\theta}{2} \cdot n_i^{(\alpha)} \cdot \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} \cdot P_{\beta\alpha}, \tag{3.25}$$

et

$$p_c(\alpha) = \frac{1}{C} \cdot \binom{n_i^{(\alpha)}}{2} \cdot \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)}. \tag{3.26}$$

Ainsi, en divisant les équations (3.25) et (3.26) par

$$\frac{1}{C} \cdot n_i^{(\alpha)} \cdot \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)},$$

on retrouve le bloc d'équations (3.24).

Dans ce chapitre, on a présenté la vraisemblance et l'échantillonnage pondéré qui sont au cœur de notre méthodologie qui permet d'estimer la taille de population effective dans le cas de virus. Cette nouvelle méthodologie est introduite brièvement au chapitre 4, et développé en détail aux chapitres 5 et 6.

CHAPITRE IV

MÉTHODES D'ESTIMATION DE L'HISTORIQUE DÉMOGRAPHIQUE À PARTIR DE SÉQUENCES DE NUCLÉOTIDES

4.1 Introduction

Dans la littérature, on retrouve un ensemble de méthodes qui permettent d'estimer l'historique démographique à partir des données de séquences d'ADN ; une classe de ces méthodes suppose que l'historique démographique pourrait être décrit par un simple modèle paramétrique comme les modèles à croissance exponentielle ou avec une taille de population constante (Hudson, 1990; Slatkin et Hudson, 1991) ou (Kuhner *et al.*, 1995, 1998). Dans ces cas, l'historique démographique pourrait être inféré à l'aide de modèles candidats en estimant les valeurs des paramètres de ces modèles.

Un bon exemple pratique de l'utilisation d'un modèle paramétrique « logistique par morceaux », pour estimer l'historique de la population infectée par l'hépatite C en Égypte, est donné par Pybus *et al.* (2003). Cette analyse a démontré une croissance rapide de l'hépatite C en Égypte entre 1930-1955, une période qui coïncide avec les campagnes de santé publique d'injection contre la schistosomiase, qui avait causé la propagation de l'épidémie de l'hépatite C en Égypte.

Dans la plupart des situations, la forme analytique de la taille de la population ne peut pas être supposée connue, et des fonctions simples de croissance de la taille

de population peuvent ne pas décrire adéquatement l'historique de la population d'intérêt. Cela a motivé le développement de modèles non paramétriques et semi paramétriques pour inférer l'historique démographique à partir de données de séquences ou d'une généalogie estimée ; citons, par exemple, Fu (1994) et Pybus *et al.* (2000).

Les méthodes non paramétriques fournissent une plus grande flexibilité pour l'estimation de la taille de population comme fonction de temps, en procédant directement à partir des données de séquences. Le résultat obtenu pourrait être utilisé dans le cadre des méthodes paramétriques pour une analyse ultérieure. Dans ce qui suit, on s'intéresse particulièrement à la littérature qui porte sur un ensemble de méthodes appelé *skyline plot*, un concept qui a été introduit par Pybus *et al.* (2000).

4.2 Famille des méthodes *skyline plot* : revue de littérature

On commence cette section par la description détaillée de la méthode *skyline plot classique*, pour ensuite introduire quelques généralisations méthodologiques qui ont été réalisées par la suite, donnant naissance à un ensemble de méthodes *skyline plot*.

4.2.1 Méthode skyline plot classique

La méthode *skyline plot classique* introduite par Pybus *et al.* (2000) nécessite l'estimation de la généalogie des séquences en supposant que l'erreur commise lors de l'estimation de la phylogénie est négligeable. Ainsi, on commence par estimer une phylogénie (unique) et, ensuite, la taille de population est estimée de manière indépendante pour chaque intervalle de la généalogie (figure 4.1).

La procédure de l'estimation de l'historique d'une population à partir d'une généalogie dépend seulement des temps écoulés entre les événements de coalescence et non pas de la relation généalogique entre les séquences. Par exemple, si les événements de coalescence se produisent rapidement, cela est un indicateur que la taille de la population est faible.

Soit $N_e(t)$, la taille de la population effective à l'instant t (voir section 2.4), où l'unité de temps est la génération. On note par $N_e(0)$ la taille de la population effective au moment d'échantillonner les séquences.

On considère un ensemble de n séquences de gènes échantillonnés aléatoirement à partir d'une population. Ainsi, la généalogie des séquences échantillonnées va contenir $(n-1)$ intervalles inter-noeuds, notés $t_2, t_3, \ldots, t_n$. Les temps de coalescence sont distribués selon la densité de probabilité (Griffiths et Tavaré, 1994a)

$$f_{T_k|V_{k+1}}(t_k|v) = \frac{\binom{k}{2}}{N_e(t_k+v)} \exp\left[-\int_v^{t_k+v} \frac{\binom{k}{2}}{N_e(x)} dx\right], \tag{4.1}$$

où l'indice k correspond au nombre de lignées présentes, et v représente le temps accumulé avant le temps de coalescence t_k, et $V_{k+1} = \sum_{i=k+1}^{n} T_i$. Si T_k est le temps de coalescence, on a, d'après Donnelly et Tavaré (1995),

$$U_k = \exp\left[-\int_{v_{k+1}}^{T_k+v_{k+1}} \frac{\binom{k}{2}}{N_e(x)} dx\right], \tag{4.2}$$

où U_k est une variable aléatoire uniforme sur $[0,1]$.

Dans la méthode *skyline plot*, on construit une phylogénie reconstruite sous l'hypothèse de l'horloge moléculaire.[1] Dans ce contexte, les temps de coalescence t_k,

[1] L'horloge moléculaire est une hypothèse qui permet de dater la distance temporelle

sont en unités de substitution par site, et non pas en générations, ce qui nécessite un changement d'échelle du processus de coalescence en posant $\gamma_k = \mu t_k$, où μ est le taux de mutation qui s'exprime en unités de nombre de substitutions par site par génération. L'estimateur non paramétrique de l'historique démographique est

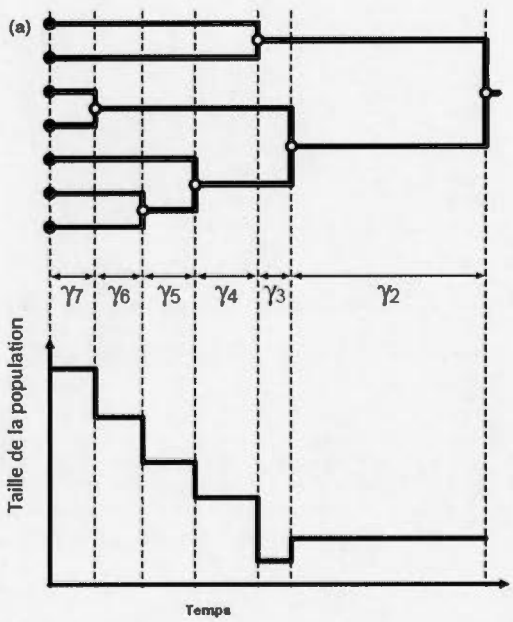


Figure 4.1: Estimation de l'historique démographique à partir d'une généalogie. (a) Une généalogie estimée à partir des longueurs de branches proportionnelles au temps. (b) Taille de la population estimée à partir de chaque intervalle de coalescence $\gamma_k = \mu t_k$ en unités de substitutions.

basé sur l'équation définie en (4.2), qu'on retrouve dans Griffiths et Tavaré (1994a) et Donnelly et Tavaré (1995).

---

entre plusieurs espèces de leur ancêtre commun, en supposant que le taux d'accumulation des mutations dans le génome d'organismes différents est du même ordre de grandeur dans des régions homologues du génome.

L'objectif est d'estimer la taille de population effective sur les intervalles $[v_{k+1}, t_k + v_{k+1}]$, en se basant sur les valeurs t_k et v_{k+1} données par la généalogie reconstruite.

Soit t_k une réalisation de T_k, et u_k la valeur correspondante selon l'équation (4.2). Ainsi, on obtient

$$t_k \binom{k}{2} = -\ln(u_k)\overline{Ne_k}, \tag{4.3}$$

où

$$\overline{Ne_k} = \left(\frac{1}{t_k} \int_{v_{k+1}}^{t_k + v_{k+1}} \frac{1}{N_e(x)} dx \right)^{-1},$$

s'interprète comme une moyenne harmonique de la taille de la population effective sur l'intervalle inter-nœud $[v_{k+1}, t_k + v_{k+1}]$. En multipliant les deux membres de (4.3) par μ, on obtient

$$\gamma_k \binom{k}{2} = -\ln(u_k)\overline{Ne_k}\mu. \tag{4.4}$$

En considérant $-\ln(u_k)$ comme une erreur aléatoire, Pybus *et al.* (2000) proposent d'estimer $\overline{Ne_k}\mu$ en posant

$$\mu\widehat{Ne}_k = \gamma_k \binom{k}{2}. \tag{4.5}$$

Cet estimateur peut être ainsi calculé à partir de la généalogie estimée.

Le graphique des $\widehat{Ne}_k$ en fonction du temps met en évidence une fonction constante par morceaux qui représente un estimateur non paramétrique de l'historique démographique.

Les auteurs affirment que $\overline{Ne_k}$ contient toute l'information concernant $N_e(x)$ qui peut être inférée sur l'intervalle observé $[v_{k+1}, t_k + v_{k+1}]$. De plus, en se basant sur l'équation (4.5), et en supposant que le taux de substitution μ est connu, $\overline{Ne_k}$ peut être estimé directement par $T_k\binom{k}{2}$.

La méthode skyline plot classique a plusieurs faiblesses qu'on énumère ci-dessous.

1. L'estimation d'une généalogie contient une erreur relative à la topologie et à la longueur des branches. La méthode ignore l'incertitude dans l'estimation de la généalogie, qu'on appelle aussi *erreur phylogénétique*. Or, cette erreur peut être importante si la généalogie contient de courtes branches, ce qui reflète, généralement, un faible taux de mutation.

2. La reconstruction de l'historique d'une population à partir d'une généalogie implique une incertitude qu'on appelle *erreur de coalescence* ; en effet, une généalogie représente une réalisation aléatoire du processus de coalescence. En particulier, l'estimation de la taille de la population dans chaque intervalle de coalescence est sujet à une grande erreur car la méthode s'apparente à l'estimation de la moyenne d'une loi exponentielle à partir d'un seul échantillon (Minin *et al.*, 2008).

   L'erreur de coalescence augmente en s'approchant de la racine de la généalogie (ancêtre commun) ; par exemple, la taille de la population sur le dernier intervalle (γ_2 de la figure 4.1) est estimée à partir de deux lignées seulement. Ceci pourrait avoir un très grand effet si la taille de la population demeure constante, puisque le plus ancien intervalle de coalescence représente en moyenne la moitié de la durée totale de la généalogie.

3. La méthode a tendance à produire des reconstructions de l'historique démographique très « bruités ». C'est particulièrement le cas lorsque la généalogie contient un nombre important de petits intervalles, avec de faibles longueurs de branches.

4.2.2 Méthode skyline plot généralisé

La méthode *skyline plot généralisé* qui a été introduite par Strimmer et Pybus (2001) permet d'éliminer les petits intervalles en les regroupant avec leurs voisins qui sont inférieurs à un certain seuil ϵ. Le choix de la valeur ϵ est réalisé en utilisant le critère AIC (*Akaike information criterion*) corrigé (Akaike, 1974), qui représente un compromis entre l'élimination du bruit et la conservation du signal démographique. Ainsi, le *skyline plot généralisé* permet de réduire le bruit existant dans le graphique *skyline plot* (Pybus *et al.*, 2000), ce qui produit des graphiques de tailles de population plus lisses. La méthode *skyline plot généralisé* exige la définition d'un ensemble de regroupements d'intervalles $A = \{a_1, a_2, ..., a_m\}$ où $(a_i > 0$, et $\sum_{i=1}^{m} a_i = n-1)$, où les valeurs a_i précisent le nombre d'événements de coalescence dans chaque intervalle, et m est le nombre d'intervalles groupés $(1 \leq m \leq n-1)$. De plus, on note le temps cumulé pour chaque intervalle groupé par Δw_j, $j = 1, 2, ..., m$, et $\widetilde{Ne} = \{Ne_1, ..., Ne_m\}$ le vecteur des tailles de populations effectives pour chaque intervalle groupé Δw_j.

Soit $h(\cdot)$ la fonction qui décrit le passage des intervalles t_i (avec un seul événement de coalescence) vers Δw_j (regroupement de plusieurs intervalles t_i). La fonction $h(\cdot)$ se définit comme

$$h(i) = \begin{cases} 1 \quad \text{si} & i \leq a_1, \\ j \quad \text{si} & \sum_{k=1}^{j-1} a_k < i \leq \sum_{k=1}^{j} a_k. \end{cases} \tag{4.6}$$

La log-vraisemblance du modèle démographique constant par morceaux est donnée par :

$$\log f_G(t_1, t_2, ..., t_{n-1} | A, \tilde{Ne}) = \sum_{i=1}^{n-1} \left(\log \left(\frac{k_i(k_i-1)}{2Ne_{h(i)}} \right) - \frac{k_i(k_i-1)}{2Ne_{h(i)}} t_i \right), \tag{4.7}$$

où $k_1, k_2, ..., k_{n-1}$, représentent le nombre de lignées dans chaque intervalle de coalescence.

La méthode *skyline plot généralisé* a été critiquée car, elle aussi, suppose que la généalogie est connue sans erreur (les temps de coalescence estimés sont exacts). Cependant, le *skyline plot généralisé* n'est plus influencé par la présence d'intervalles de coalescence courts et représente une nette amélioration par rapport au *skyline plot.*

4.2.3 Méthode skyline plot bayésien

Dans la plupart des cas, il n'est pas pratique d'ignorer l'erreur phylogénétique pour une généalogie inférée. L'incertitude des estimés des temps aux nœuds peut être peu importante dans le cas des virus ARN qui évoluent rapidement, mais devient très importante dans le cas de l'utilisation des données de séquences intraspécifique[2]. Pour palier au problème, le *skyline plot bayésien* a été introduit par Drummond *et al.* (2005).

La méthode *skyline plot bayésien* utilise une procédure d'échantillonnage MCMC pour estimer la loi *a posteriori* de la taille de la population effective à travers le temps directement à partir de séquences de gènes pour un modèle de substitution donné.

Contrairement aux autres méthodes, le *skyline plot bayésien* permet de proposer des intervalles de crédibilité pour la taille de la population effective estimée à n'importe quel point de temps. Les intervalles de crédibilité reflètent l'incertitude phylogénétique et l'incertitude liée au processus de coalescence en même temps. De plus, l'effet de moyenne des méthodes d'échantillonnage MCMC (Méthodes de Monte-Carlo par les chaînes de Markov) produit naturellement des estimateurs plus lisses que les autres méthodes de type *skyline plot.*

---

[2]Relation qui s'établit entre des individus appartenant à une seule et même espèce

Comme on dispose d'un vecteur $\widetilde{Ne} = \{Ne_1, ..., Ne_m\}$ et d'un ensemble $A = \{a_1, a_2, ..., a_m, \sum_{i=1}^{m} a_i = n-1\}$, l'historique démographique se caractérise par $2m-1$ paramètres démographiques et $n-1$ temps de coalescence.
Soit $\widetilde{g} = t_1, t_2, ..., t_{n-1}$; alors, la probabilité $f_G(\widetilde{g}|\widetilde{Ne}, A)$ de G en fonction des paramètres démographiques est donnée par l'équation (4.7), où m est choisi au préalable.

Drummond *et al.* (2005) ont utilisé un lissage simple sur $\widetilde{Ne}$ qui introduit l'hypothèse d'une taille de population autocorrélée dans le temps de telle manière que la taille de population Ne_j suit une loi de probabilité exponentielle avec une espérance égale à la taille de la population Ne_{j-1}

$$\mathcal{L}(Ne_j|Ne_{j-1}) \sim \exp(Ne_{j-1}). \tag{4.8}$$

Les auteurs supposent que $f_{Ne_1(Ne_1)} \propto 1/Ne_1$. On obtient ainsi la loi *a priori* :

$$f_{\widetilde{Ne}}(\widetilde{Ne}) \propto \frac{1}{Ne_1} \prod_{j=2}^{m} \frac{1}{Ne_{j-1}} \exp(-Ne_j/Ne_{j-1}). \tag{4.9}$$

La méthode *skyline plot bayésienne* échantillonne en même temps $\widetilde{Ne}$ et A. Ainsi, la loi *a posteriori*, dans le cas de séquences contemporaines, est donnée par :

$$f_1(\widetilde{Ne}, A, \Omega, \widetilde{g}|D, \mu) \propto Pr(D|\mu, \widetilde{g}) f_G(\widetilde{g}|\widetilde{Ne}, A) f_{\widetilde{Ne}}(\widetilde{Ne}) f_A(A) f_\Omega(\Omega), \tag{4.10}$$

où Ω représente l'ensemble des paramètres du modèle de substitution, et μ est le taux de mutation.

La méthode permet d'avoir une liste de J états, où chaque état est associé à une généalogie et à des paramètres démographiques $(Ne_j, A_j, \widetilde{g}_j)$, $j = 1, 2, ..., J$, ce qui permet de reconstituer l'historique démographique $Ne(t)$ comme une fonction constante par morceaux en fonction du temps pour chacun des J états. Cette fonction peut être décrite en calculant la distribution *a posteriori* de $Ne(t)$ pour une série de temps t_i, $i = 1, 2, ...$

Pour chaque t_i, on calcule les valeurs de $Ne(t_i)$ pour tous les J à partir de la distribution marginale *a posteriori* de la taille de la population au temps t_i.

4.2.4 Méthode skyride plot bayésien

Minin *et al.* (2008) ont développé la méthode *skyride plot bayésien* qui est une méthode alternative à la méthode *skyline plot bayésien* pour l'estimation de l'historique démographique à partir de séquences. Comme le skyline bayésien, le skyride bayésien suppose un certain degré d'autocorrélation dans les tailles de population à travers le temps. Dans cette méthode, on propose explicitement des trajectoires de taille de population ayant comme point de départ le modèle démographique constant par morceaux (Pybus *et al.*, 2000).

La construction est accomplie en imposant un lissage par un champ aléatoire Markovien Gaussien sur les paramètres de la trajectoire de la taille de la population constante par morceaux. Les auteurs ont utilisé un type lissage qui permet de pénaliser les changements de taille de population effective pour de petits intervalles de coalescence. Pour cela, Minin *et al.* (2008) ont muni les intervalles de coalescence avec des poids de lissage appropriés.

4.3 Notre méthode : le *skywis plot*

En prenant comme point de départ la méthode proposée par Pybus *et al.* (2000) afin d'estimer la taille de population effective, on propose une nouvelle méthode basée sur la théorie de coalescence. Cette méthode se différencie des autres méthodes *skyline plot* qui utilisent une seule phylogénie estimée à partir des séquences, ou à des approches de type MCMC. Dans notre cas, on a recours à un schéma d'échantillonnage pondéré efficace où notre estimé est obtenu comme une moyenne pondérée sur un grand nombre de généalogies simulées.

On rappelle que Pybus *et al.* (2000) ont proposé d'estimer la taille de la population effective sur chaque intervalle $[v_{k+1}, v_{k+1} + t_k]$ par $\widehat{Ne}_k = t_k\binom{k}{2}$, $k = 2, ..., n$ où :

- t_k est le temps d'attente jusqu'à la prochaine coalescence en présence de k séquences ; en abrégé temps de coalescence ;
- n est le nombre total de séquences échantillonnées ;
- v_{k+1} est le temps accumulé avant t_k, c'est-à-dire $v_{k+1} = \sum_{i=k+1}^{n} t_i$.

On note que pour la méthode *skyline plot*, les valeurs des t_k et v_{k+1} sont connues puisqu'elles peuvent être déduites à partir de la phylogénie estimée sous l'hypothèse de l'horloge moléculaire.

Dans notre cas, on propose une nouvelle stratégie basée essentiellement sur la théorie de la coalescence, puisque au lieu de se baser sur un estimé unique d'une phylogénie qui relie les séquences entre elles, on propose d'utiliser un grand nombre de généalogies simulées selon le processus de coalescence et ensuite calculer la moyenne pondérée des $\widehat{Ne}_k^{(j)}$, où $\widehat{Ne}_k^{(j)}$ représente l'estimé de la taille de la population effective pour le temps de coalescence t_k, et pour la généalogie $\mathcal{G}^{(j)}$, $j = 1, 2, ..., J$, $k = 2, ..., n$. Cette manière de faire permet d'explorer l'espace des généalogies possibles tout en affectant de plus grands poids aux généalogies les plus vraisemblables. L'effet de la moyenne permet ainsi d'avoir des graphiques en fonction du temps assez lisses. De plus, contrairement aux méthodes basées sur la phylogénétique, et compte tenu du fait que la théorie de coalescence s'applique dans un cadre plus général, il est possible de généraliser le travail présenté dans cette thèse à d'autres cas, comme la présence de recombinaison par exemple.

L'échantillonnage pondéré joue un rôle important dans notre méthode. Le schéma de l'échantillonnage pondéré fournit un moyen simple pour approximer la loi *a posteriori* d'une généalogie $P(\mathcal{G}|\mathcal{D},\theta)$ (voir le chapitre 3 pour les notations).

On rappelle que le poids $w^{(j)}$ associé à la généalogie $\mathcal{G}^{(j)}$, $j = 1, 2, ..., J$, est défini par

$$w^{(j)} = \frac{W^{(j)}}{\sum_{j=1}^{J} W^{(j)}}, \tag{4.11}$$

où

$$W^{(j)} = P(\mathcal{D}|\mathcal{G}^{(j)},\theta)\frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}. \tag{4.12}$$

La loi de probabilité donnée par les poids $w^{(j)}$, $j = 1, 2, ..., J$, est une approximation de la loi de probabilité *a posteriori* $P(\mathcal{G}|\mathcal{D},\theta)$. Par conséquent, il est possible d'estimer n'importe quel paramètre d'intérêt relié à une généalogie en faisant une moyenne pondérée des estimés de ces paramètres sur l'ensemble des généalogies simulées, tel que noté dans Stephens (2001).

Dans notre cas, la quantité d'intérêt est l'estimé de la taille de la population effective sur chacun des temps de coalescence t_k, notée $\widehat{Ne}_k$. D'après ce qui précède, il est possible d'obtenir $\widehat{Ne}_k$ en calculant une moyenne pondérée des $\widehat{Ne}_k^{(j)}$ sur toutes les généalogies générées. Ainsi, la valeur moyenne $\widehat{Ne}_k$ se calcule comme

$$\widehat{Ne}_k = \sum_j w^{(j)} \widehat{Ne}_k^{(j)}, \tag{4.13}$$

où

$$\widehat{Ne}_k^{(j)} = t_k^{(j)} \binom{k}{2}, \tag{4.14}$$

et $t_k^{(j)}$ représente le temps d'attente jusqu'au prochain événement de coalescence en présence de k séquences dans la généalogie $\mathcal{G}^{(j)}, j = 1, 2, ..., J$.

On suppose qu'en présence de n séquences de nucléotides, les temps de coalescence pour les généalogies $\mathcal{G}^{(1)}, ..., \mathcal{G}^{(J)}$ sont, respectivement

$$(t_n^{(1)}, t_{n-1}^{(1)}, ..., t_2^{(1)}), ..., (t_n^{(J)}, t_{n-1}^{(J)}, ..., t_2^{(J)}).$$

Ainsi, le calcul d'un estimateur non paramétrique de la taille de la population effective à travers le temps revient à :

1. l'application de l'échantillonnage pondéré en utilisant la méthode de Stephens et Donnelly (2000), ou une version modifiée de la fonction d'importance décrite au chapitre 3.1, en supposant que la matrice de transition $\boldsymbol{P}$ et le paramètre θ sont connus. Cela permet de simuler les généalogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ...\mathcal{G}^{(J)}$ et par conséquent les temps de coalescence $t_k^{(1)}, t_k^{(2)}, ..., t_k^{(J)}$ avec $k = 2, ..., n$. La procédure permet aussi de calculer les poids des généalogies $w^{(1)}, w^{(2)}, ..., w^{(J)}$ en utilisant l'équation (4.11) ;

2. calculer les valeurs de $\widehat{Ne}_k^{(1)}, \widehat{Ne}_k^{(2)}, ..., \widehat{Ne}_k^{(J)}$ pour $k = 2, ..., n$ en se basant sur l'équation (4.14) ;

3. calculer une moyenne pondérée des $\widehat{Ne}_k^{(j)}$, $j = 1, 2, ..., J$ obtenus à l'étape 2, en appliquant les poids des généalogies $\mathcal{G}^{(j)}$, $j = 1, 2, ..., J$ calculées à l'étape 1.

Cette section n'est en réalité qu'un bref aperçu de la nouvelle méthode *skywis plot*. Ainsi le prochain chapitre, sous forme d'un article scientifique en langue anglaise, contient le détail de la méthodologie ainsi que des simulations selon plusieurs scénarios démographiques où le *skywis plot* permet de reconstruire le vrai historique démographique. Vers la fin de l'article, on a aussi montré que l'utilisation de données hétérochrones (échantillonnées à divers moments dans le temps) permet d'améliorer la qualité des résultats.

On propose ensuite au chapitre 6, à travers un deuxième article scientifique en langue anglaise, d'améliorer la performance de la méthode *skywis plot*, en supposant dans un premier temps que la taille de la population relative a été préalablement estimée à quelques points dans le temps. Un cas pratique revient à utiliser la structure d'échantillonnage hétérochrone des séquences en introduisant de l'information disponible aux temps d'échantillonnage. Par exemple, Maretty et al. (2013) ont utilisé l'information sur la charge virale (Viral load) obtenue dans le cadre d'un échantillonnage en série à partir de patients infectés par un virus à évolution rapide qui est le VIH. En effet, Maretty et al. (2013) ont supposé une relation linéaire entre la charge virale et la taille de la population effective aux temps d'échantillonnage. Ainsi, entre deux échantillonnages successifs, il est possible d'estimer la taille de la population relative moyenne sur cette période en supposant une évolution neutre.

Dans ce nouveau cadre, il est possible d'utiliser une version modifiée du *skywis plot*, appelée *Calibrated skywis plot*, où la loi du temps du prochain événement, ainsi que la méthode de simulation des généalogies changent. De plus, la fonction d'importance doit être adaptée à ce nouveau contexte où le processus de coalescence et de mutation deviennent non-homogènes. Une fois la méthodologie qui permet d'estimer la taille de la population effective en partant de l'hypothèse que la taille de la population est constante par morceaux est établie, on définit une nouvelle méthode appelée *iterative calibrated skywis plot* qui est une méthode itérative où la première itération est une application de la méthode *skywis plot*, et dont l'estimation de la taille de la population effective constante par morceaux est déduite en appliquant la méthode *Calibrated skywis plot* à partir de l'itération précédente.

CHAPITRE V

PREMIER ARTICLE : A NEW METHOD FOR ESTIMATING THE DEMOGRAPHIC HISTORY FROM DNA SEQUENCES : AN IMPORTANCE SAMPLING APPROACH

Sadoune Ait Kaci Azzou, Fabrice Larribe, Sorana Froda

Abstract

The effective population size over time (demographic history) can be retraced from a sample of contemporary DNA sequences. In this paper, we propose a novel methodology based on importance sampling for exploring such demographic histories. Our starting point is the *generalized skyline plot* with the main difference being that our procedure, *skywis plot*, uses a large number of genealogies. The information provided by these genealogies is combined according to the importance sampling weights. Thus, we compute a weighted average of the effective population sizes on specific time intervals (epochs), where the genealogies that agree more with the data are given more weight. We illustrate by a simulation study that the *skywis plot* correctly reconstructs the recent demographic history under the scenarios most commonly considered in the literature. In particular, our method can capture a change point in the effective population size, and its overall performance is comparable with the one of the *bayesian skyline plot*. We also introduce the case of serially sampled sequences and illustrate that it is pos-

sible to improve the performance of the *skywis plot* in the case of an exponential expansion of the effective population size.

Keywords: importance sampling, effective population size, skywis plot, coalescent process, serially sampled

5.1 Introduction

The demographic history of a population leaves its signature in the genome, which means that DNA sequences contain information about the demographic history of the population from which they are sampled. Therefore, it is possible to use genetic data to infer demographic parameters, an issue with important implications in many fields such as public health, epidemiology and conservation biology (Minin *et al.*, 2008).

The first methods for estimating the demographic history from gene sequences were parametric and used coalescent theory. Such methods require a simple demographic model in order to describe the changes in the population size over time in terms of one or more parameters. They are based on importance sampling, e.g. (Stephens et Donnelly, 2000; Slatkin et Hudson, 1991), or Markov Chain Monte Carlo (MCMC) sampling, e.g. (Kuhner *et al.*, 1995, 1998). For example, in the case of exponential growth, the size of the population at any time t measured from the present to the past is given by $N(t) = N(0)\exp(-\beta t)$, and the unknown parameters are $N(0)$ and β.

Usually, in practice, it is not known in advance which demographic model fits the sampled gene sequences. Further, population histories are often more complex than those described by simple parametric models. This has motivated the development of nonparametric and semi-parametric methods for inferring the demographic history from sequence data or from an estimated genealogy (e.g. (Fu,

1994; Pybus *et al.*, 2000)) without resorting to some previous information about the demographic model.

Our method is nonparametric and is closely related to the family of *skyline plot* methods. The first method in this family was introduced by Pybus *et al.* (2000), and is referred to as the *classical skyline plot*. The *classical skyline plot* involves two separate steps, see (Ho et Shapiro, 2011): (1) estimating the genealogy from the sequence data and (2) estimating the population history from the estimated genealogy. Step 1 gives an estimated genealogy that includes the relationships among the individuals (tree topology) as well as their times of divergence. Genealogical estimation is done using standard phylogenetic methods under the so-called strict molecular clock. The strict molecular clock condition means that the branch lengths of the tree are proportional to time, with time being measured in mutations, and all lineages evolve at the same rate. It is also possible to estimate a genealogy in a relaxed-clock framework (Drummond *et al.*, 2006).

Further, in step 2 in order to estimate the population history from the estimated genealogy, Pybus *et al.* (2000) apply coalescent theory in a specific way by considering the times of divergence (node times) as coalescent times. When the true population size is constant, this assumption is equivalent to estimating the mean of an exponential distribution using a single realization from this distribution (Minin *et al.*, 2008). This uncertainty is referred to as 'coalescent error'. Further, the single phylogeny of the sequences is assumed to be known without error (i.e., phylogenetic error is assumed to be negligible).

Thus, Pybus *et al.* (2000) estimate the population size $\widehat{N}e_k\mu$, for each "coalescent interval" $\gamma_k = \mu t_k$, by the product of $\binom{k}{2}$ and γ_k, where μ is the mutation rate per site per generation and γ_k is measured in substitutions. Thus, the *classical skyline plot* produces a piecewise reconstruction of the demographic history that is quite

noisy, especially in the presence of small intervals when the sampled sequences are similar.

To improve the *classical skyline plot* estimation, several extensions have been proposed. Without being exhaustive, we discuss the extensions that are most relevant to our work.

Strimmer et Pybus (2001) developed a *generalized skyline plot* estimate based on the Akaike Information Criterion correction (AIC) in order to reduce the number of free parameters in the *classical skyline plot*. This method allows multiple coalescent events, i.e. for which little divergence time information is available, to be grouped together. Important developments were obtained in a Bayesian framework. Thus, Drummond *et al.* (2005) and R. Opgen-Rhein (2005) use multiple change-point (MCP) models to estimate population size dynamics.

In particular, Drummond *et al.* (2005) use a Markov chain Monte Carlo (MCMC) sampling procedure that efficiently samples a variant of the *generalized skyline plot*, given sequence data, and combines these plots in order to generate a posterior distribution of the effective population size through time. Due to the "averaging" effect of the MCMC sampling, the *Bayesian skyline plot* introduced by Drummond *et al.* (2005) produces smoother estimates than previous skyline plot methods. Also in the Bayesian framework, Minin *et al.* (2008) propose an alternative to change-point modeling that exploits Gaussian Markov random fields to achieve temporal smoothing of the effective population size. The advantage of the *skyride* method is that in contrast to estimates given by MCP models, explicit temporal smoothing does not require strong prior decisions like fixing the total number of change points *a priori*.

Finally, Heled et Drummond (2008) introduced the extended Bayesian skyline plot, which permits the analysis of multiple unlinked loci. Increasing the num-

ber of independent loci allows the uncertainty in the coalescent to be assessed, leading to an improvement in the reliability of the demographic inference and a substantial reduction in estimation error (Ho et Shapiro, 2011). Further, unlike previous *skyline plot* methods that use a piecewise-constant model, the extended Bayesian skyline plot permits the use of a piecewise-linear model to describe the demographic history, allowing the population size to change continuously along each interval.

In order to estimate the effective population size, we propose a new method in a likelihood-based perspective. Unlike some skyline methods that use a single estimated phylogeny of the sequences, or others that use MCMC approaches, we resort to an efficient importance sampling scheme and our estimate comes to an weighted average over a large number of simulated genealogies, each with a different set of coalescence times. The methodology is described in detail in Section 5.3.

5.2 Background

5.2.1 Coalescent theory

In this section, we present the basic ideas behind the standard coalescent, as well as its extension to the case of fluctuating population size. An introduction to coalescent theory can be found in Nordborg (2001). Coalescent theory allows one to produce genealogies relating the sampled sequences according to a large class of population genetic models. In particular, the classical coalescent process assumes a single, isolated and panmictic population (e.g. a Wright-Fisher model), which evolves with constant (haploid) size N over many generations. For sufficiently large N and a sample size n such that $n \ll N$, the ancestral relationships between the gene sequences can be approximated by Kingman's coalescent (Kingman, 1982b).

In short, the ancestry of a sample of sequences is modeled back in time, starting from the current sample and until the most recent common ancestor (MRCA) of the sample is found. At each step in the genealogical tree, one of the following events can occur: (1) two sequences coalesce if they share a common ancestor; (2) one sequence mutates. In the coalescent framework, time is measured in units of N generations, and N is large. The mutation rate μ per sequence per generation is rescaled so that $\theta = 2N\mu$. Further, one can consider that each pair of lineages coalesces independently as a Poisson process with rate 1, and so, when there are k ancestral lines, coalescent events occur as in a Poisson process with total rate $k(k-1)/2$ (Stephens, 2000).

In the classical coalescent process, and in the presence of k gene sequences, the waiting time T_k to the next coalescent event is exponentially distributed with rate $\binom{k}{2}$, while the distribution of the time until the first mutation event in any of the k lineages is exponential with parameter $k\theta/2$. Since mutations are assumed to occur independently of coalescence, the waiting time until a mutation or coalescent event is exponentially distributed with parameter

$$\binom{k}{2} + \frac{k\theta}{2} = \frac{k(k-1+\theta)}{2}. \tag{5.1}$$

The classical coalescent framework can be extended to include simple deviations from the idealized Wright-Fisher model, like recombination, fluctuating population size, population structure, and selection. In our paper, we focus on a single extension of the coalescent, namely variable population size.

In the case of non-constant population size, the number of descendants of a sequence in one generation does not follow the Poisson distribution with intensity one (Hein *et al.*, 2005). As a result, when the basic coalescent is used to model a real physical population, the size N of the population in the (haploid) Wright-Fisher model cannot be assumed to be equal to the size of the real population.

Let $N_e(t)$ denote the effective population size at time t with $N_e(0) = N$. The effective population size reflects the number of individuals that contribute offsprings to the descendant generation and is almost always smaller than the census population size.The variable population size coalescent model for contemporary gene sequences was introduced by Griffiths et Tavaré (1994b) and Donnelly et Tavaré (1995). In this case, the coalescence times $T_2, T_3, ..., T_n$ do not follow independent exponential distributions.

Let $V_k = T_n + ... + T_k$ be the accumulated waiting time so that the number of sequences pass from n to $k-1$ sequences, i.e.

$$V_k = \sum_{\ell=k}^{n} T_\ell, \tag{5.2}$$

and let $\Lambda(t)$ the cumulative coalescent rate over time measured relative to the rate at time $t = 0$:

$$\Lambda(t) = \int_0^t \frac{1}{\nu(u)} du, \tag{5.3}$$

where $\nu(t) = N_e(t)/N$, the relative size of $N_e(t)$ to N.

The waiting time until the next event depends only on the time of the previous event by the Markov property. The survival function of the time T_k conditional on $V_{k+1} = v$ is

$$P(T_k > t | V_{k+1} = v) = \exp\left\{-\binom{k}{2}(\Lambda(t+v) - \Lambda(v))\right\}, \tag{5.4}$$

where $v_{n+1} = 0$.

We note that when replacing $\Lambda(t)$ by t (i.e., in the case $N_e(t) = N, t > 0$) in equation (5.4), we get the survival function of the exponential distribution. From

(5.4), we obtain the density

$$f_{T_k|V_{k+1}}(t_k|v) = \frac{\binom{k}{2}}{N_e(t_k+v)} \exp\left[-\int_v^{t_k+v} \frac{\binom{k}{2}}{N_e(x)} dx\right]. \tag{5.5}$$

It is precisely from this equation that Pybus *et al.* (2000) derived the estimation of the effective population size $\widehat{N}e_k$ in the presence of k sequences.

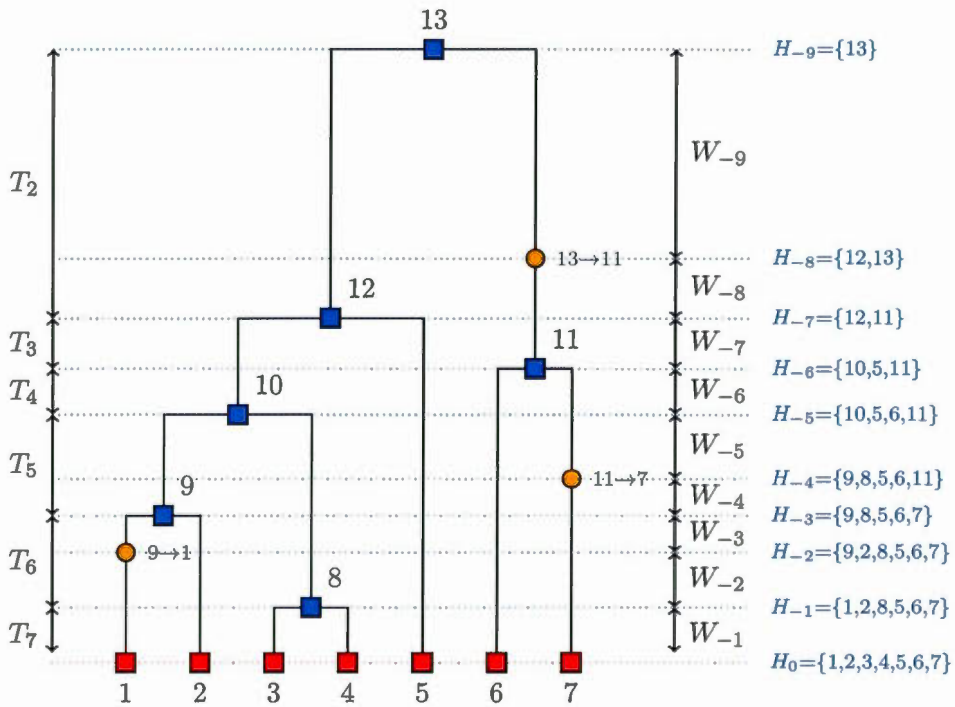


Figure 5.1: Example of a realization of the coalescent process viewed from past to the present with $n = 7$ sequences (red squares), with 6 coalescent events (blue squares) and 3 mutation events (orange circles).

5.2.2 Importance sampling

Parameter estimation in population genetic models require optimization of the likelihood of the data given the parameters, $P(\mathcal{D}|\theta)$. The likelihood is then evaluated by

$$L(\theta) = \int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G},\theta)P(\mathcal{G}|\theta)d\mathcal{G}, \tag{5.6}$$

where θ is the collection of parameters (such as population size and migration rates) for the population process. Typically, the objective of the analysis is to estimate these parameters by averaging the likelihood over all possible genealogies. A naive Monte Carlo method for the integral in (5.6) is given by

$$L(\theta) \approx \frac{1}{J}\sum_{j=1}^{J} P(\mathcal{D}|\mathcal{G}^{(j)},\theta), \tag{5.7}$$

where $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$ are an independent sample from $P(\mathcal{G}|\theta)$.

Importance Sampling (IS) allows us to improve the efficiency of the Monte Carlo integration. The main idea of the IS approach is to reduce the inefficiency of the approximation (5.7) by concentrating the simulation on the trees that are more likely with the observed data. Instead of choosing histories from the distribution $P_\theta(\mathcal{G})$, we want to sample genealogies from a proposal distribution $Q(\mathcal{G})$ that better supports the observed data, $\mathcal{D}$. The IS method is based on rewriting (5.6) as

$$\int_{\mathcal{G}} P(\mathcal{D}|\mathcal{G},\theta)\frac{P(\mathcal{G}|\theta)}{Q(\mathcal{G})}Q(\mathcal{G})d\mathcal{G}. \tag{5.8}$$

The Monte Carlo approximation of (5.8) gives

$$L(\theta) \approx \frac{1}{J}\sum_{j=1}^{J} P(\mathcal{D}|\mathcal{G}^{(j)},\theta)\frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}, \tag{5.9}$$

where $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)} \sim Q(\mathcal{G})$. Good choices of the distribution $Q(.)$ make this method of approximation much more efficient than (5.7). Ideally, we would like

to sample from $Q(\mathcal{G}) = P(\mathcal{G}|D)$. However, this is impossible because it supposes perfect knowledge of the likelihood which is not true in practice.

Importance sampling (IS) was first used in this context by Griffiths et Tavaré (1994a,b,c). Stephens et Donnelly (2000) proposed improvements to the method by suggesting an approximation to an optimal proposal distribution for IS, $P(\mathcal{G}|D)$.

5.3 The Skywis method

In this section, we describe our estimation method of the effective population size, when n gene sequences are available. The main idea behind our method is to simulate a large number of genealogies and create a weighted average of the effective population sizes, where the most probable genealogies are given larger weight. In short, reconstructing the demographic history from these sequences involves four distinct steps:

1. simulate J genealogies : $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$;
2. compute $\widehat{N}e_k^{(1)}, \widehat{N}e_k^{(2)}, ..., \widehat{N}e_k^{(J)}$ where $\widehat{N}e_k^{(j)}, k = 2, 3, ..., n$, represents the estimated effective population size for the genealogy $\mathcal{G}^{(j)}$ for each coalescent time $t_k^{(j)}$ (in the presence of k sequences);
3. compute the weights $w^{(1)}, w^{(2)}, ..., w^{(J)}$, where $w^{(j)}$ represents the weight of the genealogy $\mathcal{G}^{(j)}$ in the likelihood of the data;
4. estimate $\widehat{N}e_k$ based on genalogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$, by the weighted mean of $\widehat{N}e_k^{(j)}$, for $j = 1, 2, ..., J$, and $k = 2, 3, ..., n$, i.e.

$$\widehat{N}e_k = \sum_{j=1}^{J} w^{(j)} \widehat{N}e_k^{(j)}. \tag{5.10}$$

For example, with a variable population size that is expanding from the past to the present, as we progress towards the MRCA one can expect the population size

to be smaller, or coalescence times to be shorter, than in the case of a constant population size. This fact, of shorter coalescence times, should be reflected more faithfully by the most probable genealogies. Since such genealogies receive the largest weights, one can see that through the weighting system the estimator is adapting itself to the information contained in the data.

In what follows we describe our method in full detail, namely :

- how to simulate genealogies;
- how to set the weights;
- how to estimate the effective population size.

5.3.1 *Skywis plot* for homochronous sampling

Simulation of genealogies

In order to generate genealogies we use the proposal distribution $Q(.)$ introduced by Stephens et Donnelly (2000) assuming a constant population size and a finite sites model with known mutation parameters. Given the Stephens et Donnelly (2000) method is crucial to our approach, we describe it briefly.

Let:

- E: the set of possible types of gene sequences;
- H_{-i}: the set of all sequences when event i occurs (coalescence or mutation) where i increases from the present to the past in steps of 1 for each event (See Figure 5.1);
- $\mathcal{H} = \{H_0, H_{-1}, ..., H_{-m}\}$: a history of sequences where $H_0 = \mathcal{D}$, m is the total number of events in the history $\mathcal{H}$, and H_{-m} is a singleton (the MRCA);

- **P**: the mutation transition matrix;

In the Stephens et Donnelly (2000) method, the H_{-i} are viewed as states of a Markov process starting at genetic type $H_{-m} \in E$ and ending with $H_0 \in E$. Let **P** be the mutation transition matrix. Let $P_{\alpha\beta}$ be the probability of a DNA sequence of type α to mutate to a DNA sequence of type β, and let M_α^β denote a mutation of a gene sequence from type α to type β according to **P**; let C_α^α denote a coalescence of two gene sequences of type α. Then, the forward transition probabilities $p_\theta(H_i|H_{i-1})$, are defined by equation (5.11):

$$p_\theta(H_i|H_{i-1}) = \begin{cases} \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1}-1+\theta)} P_{\alpha\beta} & \text{if } M_\alpha^\beta, \\ \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1}-1}{n_{i-1}-1+\theta} & \text{if } C_\alpha^\alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{5.11}$$

where $n_{i-1}^{(\alpha)}$ is the number of sequences of type α in H_{i-1}, $n_{i-1} \geq 2$ is the number of sequences in H_{i-1}.

Stephens et Donnelly (2000) consider randomly constructing histories backward in time in a Markov way, from the sample H_0 to an MRCA (single type), according to some backward transition probabilities $q_\theta(H_{i-1}|H_i)$ in the class $\mathcal{M} = \{H_{i-1}|P_\theta(H_i|H_{i-1}) > 0\}$ with the constraint $q_\theta(H_{i-1}|H_i) \propto p_\theta(H_i|H_{i-1})$. Their proposed backward transition probabilities $\widetilde{q}_\theta(H_{i-1}|H_i)$ which define $Q(.)$ are given by equation (5.12), namely:

$$\widetilde{q}_\theta(H_{i-1}|H_i) = \begin{cases} C^{-1}\dfrac{\theta}{2}\cdot n_i^{(\alpha)}\cdot\dfrac{\hat{\pi}(\beta|H_i-\alpha)}{\hat{\pi}(\alpha|H_i-\alpha)}\cdot P_{\beta\alpha} & \text{if } M_\beta^\alpha, \\ C^{-1}\dbinom{n_i^{(\alpha)}}{2}\cdot\dfrac{1}{\hat{\pi}(\alpha|H_i-\alpha)} & \text{if } C_\alpha^\alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{5.12}$$

where $n_i^{(\alpha)}$ is the number of sequences of type α in H_i, n_i is the number of sequences in H_i, $\{H_i-\alpha\}$ is the set of all sequences in H_i without the chosen sequence α, and $C = n_i(n_i-1+\theta)/2$ is a constant of proportionality. The estimated conditional probability $\hat{\pi}(\alpha|H_i)$ is described below.

In the proposed reconstruction, when H_i contains n_i chromosomes, the new type α is obtained by choosing a chromosome from H_i at random and then mutating it a geometric number of times. If $n_i^{(\beta)}$ is the number of chromosomes of type β in H_i, then (Stephens et Donnelly, 2000)

$$\hat{\pi}(\alpha|H_i) = \sum_{\beta\in E}\sum_{m=0}^{\infty}\frac{n_i^{(\beta)}}{n_i}\left(\frac{\theta}{n_i+\theta}\right)^m\frac{n_i}{n_i+\theta}(\mathbf{P}^m)_{\alpha\beta}. \tag{5.13}$$

In our approach, the genealogies are simulated backwards in time by the following algorithm based on (5.12) :

1. initialize $n_i := n$, where n is the number of DNA sequences at time $t = 0$ (present), and $i = 0$;

2. simulate the time to the next event, W_{-i-1}, as an exponential distribution with parameter $\dbinom{n_i}{2} + \dfrac{n_i\theta}{2}$;

3. randomly choose a sequence from H_i; the chosen sequence type is denoted α;

4. for each type $\beta \in E$ for which $P_{\alpha\beta} > 0$, compute $\hat{\pi}(\beta|H_i - \alpha)$;

5. compute the quantities x_1 and x_2, where

$$x_1 = \theta \sum_{\beta \in E} \hat{\pi}(\beta|H_i - \alpha) P_{\beta\alpha} \quad \text{and} \quad x_2 = n_i^{(\alpha)} - 1.$$

   Then, choose:

   - a coalescence event with probability $\frac{x_2}{(x_1 + x_2)}$;
   - a mutation event (to β) with probability $\frac{x_1}{(x_1 + x_2)}$.

6. depending on the type of event chosen in step 5, we continue as follows:

   - if there is a coalescence event, choose another sequence of type α randomly, and let $n_{i-1} := n_i - 1$;
   - if there is a mutation event, mutate the sequence α into a sequence β, without changing n_i, i.e. let $n_{i-1} := n_i$;

7. let $i := i - 1$ and continue until $n_i = 1$.

After implementing the above algorithm, the coalescence times that are at the core of our method can be deduced. In the genealogy $\mathcal{G}$ given in Figure 5.1, we can deduce the coalescent times from the event times. For example, $T_7 = W_{-1}$ whereas $T_6 = W_{-2} + W_{-3}$ because we have a mutation event before a coalescence event.

Weights of genealogies

After generating genealogies using the Stephens et Donnelly (2000) proposal distribution, it is possible to compute the importance weight $w^{(j)}$ for each genealogy

$\mathcal{G}^{(j)}$, with $j = 1, 2, ..., J$. Then $w^{(j)}$ is given by :

$$w^{(j)} = \frac{W^{(j)}}{\sum_{j=1}^{J} W^{(j)}}, \tag{5.14}$$

where

$$W^{(j)} = P(\mathcal{D}|\mathcal{G}^{(j)}, \theta)\frac{P(\mathcal{G}^{(j)}|\theta)}{Q(\mathcal{G}^{(j)})}, \tag{5.15}$$

with

$$Q(\mathcal{G}^{(j)}) = \prod_{i=0}^{-m+1} \widetilde{q}_\theta(H_{i-1}|H_i), \tag{5.16}$$

and

$$P(\mathcal{G}^{(j)}|\theta) = \prod_{i=0}^{-m+1} p_\theta(H_i|H_{i-1}). \tag{5.17}$$

Estimation of the effective population size

When building genealogies backwards in time, as we move backwards in time, fewer coalescence events occur. As a result, coalescence times close to the present are very short and become larger gradually going back in time. These short coalescence times create an undesirable variability in the estimation of the effective population size. Therefore we propose to cumulate small coalescence times in order to improve the estimation of the effective population size. These cumulated time intervals are called epochs. To define epochs that get larger as we go backwards in time, we followed Durbin and Li (2011), and used a special time scale based on the TMRCA. Forest (2014) adopted the same method.

Finally, we note that the idea of cumulating small coalescence times in order to smooth the graph of the estimator of the effective population size was first proposed by Strimmer and Pybus (2001); it has since become quite standard in the related literature.

Once the genealogies have been simulated using the method described in Section 5.3.1, we cumulate the coalescence times as follows :

- we fix the total number of epochs, n_{cum}, i.e. the total number of time intervals where we estimate the effective population size;
- for each simulated genealogy $\mathcal{G}^{(j)}$, we compute the MRCA time, $T_{MRCA}^{(j)}$;
- we use formula (6.42) proposed by Durbin et Li (2011) in order to define epochs where estimates of the effective population size are computed. In other words, the following time cutting points in a genealogy $\mathcal{G}^{(j)}, j = 1, 2, .., J$ are used :

$$t_{\text{cut},b}^{(j)} = 0.1 \cdot \exp\left(\frac{b}{n_{\text{cum}}} \cdot \log(1 + 10 \cdot T_{MRCA}^{(j)})\right) - 0.1, \quad b = 1, 2, ..., n_{\text{cum}}, \quad (5.18)$$

where $t_{\text{cut},n_{\text{cum}}}^{(j)} = T_{MRCA}^{(j)}$.

For each genealogy, formula (5.18) gives the boundaries of the epochs, measured from the present to the past where $b = 0, 1, 2, ...n_{\text{cum}}$ (in units of N generations). The boundaries of epochs are different for each genealogy $\mathcal{G}^{(j)}$ and depend on the length of the tree. For example if $T_{MRCA}^{(1)} = 1$ in units of N generations and $n_{\text{cum}} = 5$, then according to (5.18), the boundaries of the intervals are 0.0615, 0.1609, 0.3215, 0.581 (backward in time). For example, for the first epoch, this means that we must cumulate coalescence times from T_n until reaching 0.0615 N generations.

The *skyline plot* can be viewed as a method of moments estimator based on the standard coalescence distributions (Strimmer et Pybus, 2001). For a genealogy $\mathcal{G}^{(j)}$, we have :

$$E\left(T_k^{(j)} \cdot \binom{k}{2}\right) = N \text{ (generations)}, \quad (5.19)$$

because $T_k^{(j)}$ is exponentially distributed as $\exp\left(\binom{k}{2}\right)$. Therefore, we use the estimate

$$\widehat{N}e_k^{(j)} \approx t_k^{(j)} \binom{k}{2}, j = 1, 2, ..., J. \tag{5.20}$$

The expectation of the accumulated waiting time in order to pass from n to ℓ lineages, $T_{n\to\ell}^{(j)} = \sum_{k=\ell}^{n} T_k^{(j)}$, is given by (see, for example, (Rodrigo *et al.*, 1999))

$$E\left(T_{n\to\ell}^{(j)}\right) = \frac{2c}{n(n-c)} N \text{ (generations)}, \tag{5.21}$$

where $c = n - \ell$ represents the number of coalesced sequences. From equation (5.21), we can see that we can estimate, using the method of moments, the effective population size for the cumulated time of c coalescence times by :

$$t_{n\to\ell}^{(j)} \cdot \frac{n(n-c)}{2c}, \tag{5.22}$$

where $t_{n\to\ell}^{(j)} = \sum_{k=\ell}^{n} t_k^{(j)}$, and $c = n-\ell$. In our case, the cumulated waiting times for each genealogy $\mathcal{G}^{(j)}$ are deduced from equation (5.18) : once the boundaries of the intervals of epochs are computed, the cumulated waiting times, $\Delta t_b^{(j)}$ numbered from present to the past, are derived as :

$$\Delta t_b^{(j)} = t_{\text{cut},b}^{(j)} - t_{\text{cut},b-1}^{(j)}, \tag{5.23}$$

where $b = 1, 2, ..., n_{\text{cum}}$, $j = 1, 2, ..., J$ and $t_{\text{cut},0}^{(j)} = 0$. It follows from equations (5.22) and (5.18) that the estimated effective population size for an epoch b, and genealogy $\mathcal{G}^{(j)}, j = 1, 2, ..., J$, is given by :

$$\widehat{N}e_b^{(j)} = \Delta t_b^{(j)} \cdot \frac{d_b^{(j)}\left(d_b^{(j)} - c_b^{(j)}\right)}{2c_b^{(j)}}, \tag{5.24}$$

where $d_b^{(j)}$ is the number of sequences at the beginning of the $\Delta t_b^{(j)}$ interval, and $c_b^{(j)}$ is the number of cumulated coalescence times in the epoch $\Delta t_b^{(j)}$, $b = 1, 2, ..., n_{\text{cum}}$, $j = 1, 2, ..., J$.

The distribution of importance weights of genealogies described by the equation (5.15) is an approximation to the posterior distribution $P(\mathcal{G}|\mathcal{D},\theta)$. As a result, one can approximate quantities of interest related to the tree by forming a weighted average of these quantities over the sampled trees as suggested in Stephens (2001).

In our case, we are interested in the estimation of $E(Ne_b)$, $b = 1, 2, ..., n_{\text{cum}}$ from the J estimates $\widehat{N}e_b^{(j)}$, $j = 1, 2, ..., J$ and we let

$$E(Ne_b) \approx \sum_{j=1}^{J} w^{(j)} \widehat{N}e_b^{(j)}. \tag{5.25}$$

In our algorithm, the weighted average of $\widehat{N}e_b^{(j)}$ is computed for the same time interval for all $j = 1, 2, ..., J$ that represent the intersections of epochs for the J simulated genealogies. This way of proceeding gives us weighted estimates of effective population sizes under the assumption that the effective population size is constant for an epoch. The reason for taking common intervals accross genealogies is that $\widehat{N}e_k^{(j)}$ estimates the integral (see Pybus *et al.* (2000))

$$\left(\frac{1}{t_k} \int_{v_{k+1}^{(j)}}^{t_k^{(j)} + v_{k+1}^{(j)}} \frac{dx}{N_e(x)} \right)^{-1}, j = 1, 2, ..., J. \tag{5.26}$$

Therefore, to estimate the integral (5.26) by a weighted average of estimates from J genealogies, we must use the same time intervals.

For illustration, in Figure 5.2, we assume that two genealogies $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are simulated using the method described in Section 5.3.1 with respective weights $w^{(1)}$ and $w^{(2)}$. Further, we assume that we cumulate coalescence times to obtain $n_{\text{cum}} = 3$ epochs. The limits of epochs for a genealogy $\mathcal{G}^{(j)}$ are denoted $t_{\text{cut},b}^{(j)}, b = 1, 2$, and the time to the MRCA by $\text{TMRCA}^{(j)}, j = 1, 2$. The detailed calculation of the weighted effective population size per epochs is summarized in the following table :

Time interval	$\widehat{N}e_\ell, \ell = 1, 2, ..., 2 \cdot n_{\text{cum}}$
$[0; t^{(2)}_{\text{cut},1})$	$w^{(1)}\widehat{N}e^{(1)}_1 + w^{(2)}\widehat{N}e^{(2)}_1$
$[t^{(2)}_{\text{cut},1}; t^{(1)}_{\text{cut},1})$	$w^{(1)}\widehat{N}e^{(1)}_1 + w^{(2)}\widehat{N}e^{(2)}_2$
$[t^{(1)}_{\text{cut},1}; t^{(2)}_{\text{cut},2})$	$w^{(1)}\widehat{N}e^{(1)}_2 + w^{(2)}\widehat{N}e^{(2)}_2$
$[t^{(2)}_{\text{cut},2}; t^{(1)}_{\text{cut},2})$	$w^{(1)}\widehat{N}e^{(1)}_2 + w^{(2)}\widehat{N}e^{(2)}_3$
$[t^{(1)}_{\text{cut},2}; TMRCA^{(2)})$	$w^{(1)}\widehat{N}e^{(1)}_3 + w^{(2)}\widehat{N}e^{(2)}_3$
$[TMRCA^{(2)}; TMRCA^{(1)}]$	$\widehat{N}e^{(1)}_3$

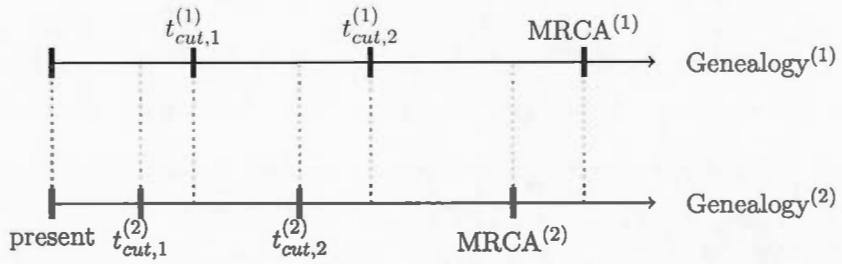


Figure 5.2: Division of time axis in the presence of two genealogies.

5.3.2 *Skywis plot* for heterochronous sampling

The algorithm described in Section 5.3.1 can be generalized to the case of serially sampled sequences i.e. sequences sampled at different moments in time. Such samples are also called heterochronous. Figure 5.3 illustrates a case where we sampled sequences at times $t_0 < t_1 < t_2$, and the time is measured from the present to the past. Let S be the number of instants where we sampled sequences ($S = 3$ in Figure 5.3). Rodrigo et Felsenstein (1999) extend the coalescent likelihood for such heterochronous sequences, a very important issue in the case of rapidly

evolving viruses such as HIV. For example, Rodrigo *et al.* (1999) have estimated, using heterochronous sequences, the viral generation time of HIV typel (HIV-1). Also, serially sampling rapidly evolving populations is used for dating evolutionary events and divergence times (see, e.g., Drummond *et al.* (2003b)).

In the presence of serially sampled sequences, we have to adapt the method of Stephens et Donnelly (2000) in order to simulate genealogies in this case. This necessarily involves developing new formulas for the probabilities $p_\theta(H_i|H_{i-1})$ and $\widetilde{q}_\theta(H_{i-1}|H_i)$, as given below.

Backward and forward probabilities, and weights of genealogies

Let $n^{(s)}$ be the number of additional sampled sequences at time t_s, with $s = 1, 2, \ldots, S-1$. The main difference between the algorithm for homochronous sequences presented in Section 5.3.1, and the new algorithm for heterochronous sequences is that the number of lineages increases at the (known) instants $t_s, s = 1, 2, \ldots, S-1$ where samples of sequences are added. Further, it is necessary to use event times, because the embedded chain differs according to the relative position of these event times with respect to $t_s, s = 0, 1, 2, \ldots, S-1$.

In other words, the probabilities $p_\theta(H_i|H_{i-1})$ and $\widetilde{q}_\theta(H_{i-1}|H_i)$ are calculated differently from the case of a single sample of sequences, which has an impact on how the weights of genealogies, $w^{(j)}$, $j = 1, 2, ..., J$, are computed. In order to present our results, we introduce these additional notations :

- $D_{i,v} = \{H_i, v\}$: represents the set of all sequences present in the population after the i^{th} event at time v; this is a generalization of H_i with the specification of the time of event i;

- $\mathcal{E}_s$: represents the set of all sequences added at time t_s;

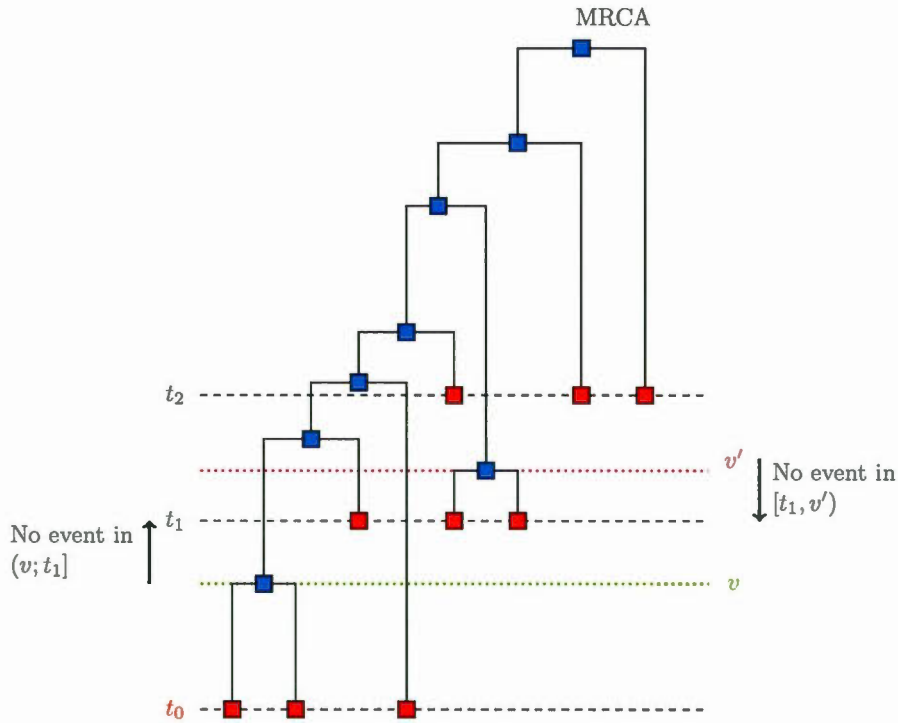


Figure 5.3: Example of serially sampled sequences with $S = 3$. The red squares are the sampled sequences and the blue squares are the sequences derived from coalescence.

Our proposal distribution is an adapted version of the Stephens et Donnelly (2000) method for simulating genealogies, to the case of heterochronous sequences. In this case, as mentioned above, we consider that there is a list of pre-specified sampling times $t_s, s = 0, 1, 2 \ldots, S-1$ which are dividing the time axis. In what follows, time is measured from the present to the past and by *event* we mean either a coalescence or a mutation. If an event time v is such that $t_{s-1} < v < t_s$ and the time v' of the next event is such that $v' > t_s$, v' is truncated at t_s, i.e. $v' \leq t_s$. Then, either there is a next event at time $v' \leq t_s$ or the time is truncated at t_s, new sequences are added, and the process starts anew. Thus, from $D_{i,v}$ one can

move to either $D_{i-1,v'} = \{H_{i-1}, v'\}, v < v' \leq t_s$, where H_{i-1} is obtained from H_i by a coalescence or a mutation, or to D_{i-1,t_s} where $H_{i-1} = H_i + \mathcal{E}_s$. In this last case we add $\mathcal{E}_s$ sequences at time t_s and the process starts anew, with a new set of sequences that includes those at v. The moves of the process (embedded chain) are given by the following formulas, and we consider separately the case $v' < t_s$ and the case $v' = t_s$.

Case 1: $t_{s-1} < v < v' < t_s$.

$$\widetilde{q}_\theta(D_{i-1,v'}|D_{i,v}) = \begin{cases} C^{-1}\frac{\theta}{2} n_i^{(\alpha)} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{if } M_\beta^\alpha \\ C^{-1}\binom{n_i^{(\alpha)}}{2} \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{if } C_\alpha^\alpha \\ 0 & \text{otherwise,} \end{cases} \tag{5.27}$$

Case 2: $t_{s-1} < v < t_s, v' = t_s$.

$$\widetilde{q}_\theta(D_{i-1,t_s}|D_{i,v}) = \begin{cases} \Pr(\ \exists \text{ an event in } (v; t_s]\ ) \cdot C^{-1}\frac{\theta}{2} n_i^{(\alpha)} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{if } M_\beta^\alpha, \\ \Pr(\ \exists \text{ an event in } (v; t_s]\ ) \cdot C^{-1}\binom{n_i^{(\alpha)}}{2} \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} & \text{if } C_\alpha^\alpha, \\ \Pr(\text{no event in } (v, t_s]) & \text{if } H_{i-1} = H_i + \mathcal{E}_s, \\ 0 & \text{otherwise.} \end{cases} \tag{5.28}$$

Normally (i.e. in homochronous sampling), the waiting time W_{-i-1} from the state $D_{i,v}$ with $t_{s-1} < v < t_s$ to the next event has an exponential distribution with

rate $\lambda_i = \binom{n_i}{2} + \frac{n_i\theta}{2}$, where n_i is the number of lineages at time v. Thus, the probability that there are no events in the interval $(v, v'] \equiv (v, ts]$ is given by the survival function

$$Pr(W_{-i-1} > t_s - v) = \exp\left(-\lambda_i(t_s - v)\right), \tag{5.29}$$

where W_{-i-1} is the waiting time from state H_i to state H_{i-1} in a process with homochronous sampling.

In the case of heterochronous sequences, the algorithm for simulating the genealogies backward in time is the following:

1. initialize $n_i = n$ and $s = 0$, where n is the number of sampled sequences at time $t_0 = 0$ (present), and s is the index of times where we perform the sampling. Further, initialize the cumulated time t_{cum} to 0;

2. simulate the time to the next event, W_{-i-1}, as an exponential distribution with parameter $\binom{n_i}{2} + \frac{n_i\theta}{2}$; let t_{evt} be the observed value;

3. compute $t^*_{\text{cum}} := t^{(i)}_{\text{cum}} + t_{\text{evt}}$;

4. if $t^{(i)}_{\text{cum}} < t_s$ and $t^*_{\text{cum}} > t_s$, then

   - let $t^{(i-1)}_{\text{cum}} = t_s$;
   - let $n_{i-1} := n_i + n^{(s)}$ (add a sample of sequences at time t_s);
   - let $s := s + 1$ and $i := i - 1$, and go to step 2;

   otherwise, go to step 5;

5. let $t^{(i-1)}_{\text{cum}} := t^*_{\text{cum}}$ and randomly choose a sequence from n_i; the chosen sequence type is denoted α;

6. compute the quantities x_1 and x_2, where

$$x_1 = \theta \sum_{\beta \in E} \hat{\pi}(\beta | H_i - \alpha) P_{\beta\alpha} \quad \text{and} \quad x_2 = n_i^{(\alpha)} - 1.$$

   Then, choose:

   - a coalescence event with probability $\frac{x_2}{(x_1 + x_2)}$;
   - a mutation event (to β) with probability $\frac{x_1}{(x_1 + x_2)}$.

7. depending on the result in step 6:

   - if there is a coalescence event, choose another sequence of type α randomly, and let $n_{i-1} := n_i - 1$;
   - if there is a mutation event, mutate the sequence α into a sequence β, without changing n_i, i.e. let $n_{i-1} := n_i$;

8. let $i := i - 1$ and continue until $n_i = 1$.

After the definition of how to build a genealogy in the case of serially sampled sequences, and the proposal distribution Q, we introduce the probability P of the genealogy by specifying the probability of passing from the state $D_{i-1,v'} = \{H_{i-1}, v'\}$ to the state $D_{i,v} = \{H_i, v\}$ when there are n_{i-1} sequences, and we suppose that an event time v' is such that $t_s < v' < t_{s+1}$ (a coalescence corresponds to a split when viewed from the past to the present). Therefore, as for the backward transition probabilities, we consider separately the case $v > t_s$ and the case $v = t_s$.

Case 1: $t_s < v < v' < t_{s+1}$.

$$p_\theta(D_{i,v}|D_{i-1,v'}) = \begin{cases} \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1}-1+\theta)} P_{\alpha\beta} & \text{if } M_\alpha^\beta \\ \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1}-1}{n_{i-1}-1+\theta} & \text{if } C_\alpha^\alpha \\ 0 & \text{otherwise.} \end{cases} \tag{5.30}$$

Case 2: $t_s < v' < t_{s+1}$ and $v = t_s$.

$$p_\theta(D_{i,v}|D_{i-1,v'}) = \begin{cases} \Pr(\exists \text{ an event in } [t_s; v')) \cdot \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\theta}{(n_{i-1}-1+\theta)} P_{\alpha\beta} & \text{if } M_\alpha^\beta \\ \Pr(\exists \text{ an event in } [t_s; v')) \cdot \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1}-1}{n_{i-1}-1+\theta} & \text{if } C_\alpha^\alpha \\ \Pr(\text{no event in } [t_s, v')) & \text{if } H_i = H_{i-1} - \mathcal{E}_s \\ 0 & \text{otherwise,} \end{cases} \tag{5.31}$$

where :

- the probability that there are no events in the interval $[t_s, v')$ is given by :

$$\Pr(Wt_{-i-1} > v' - t_s) = \exp\left(-\lambda_i(v' - t_s)\right). \tag{5.32}$$

- $n_{i-1}^{(\alpha)}$ represents the number of sequences of type α in $D_{i-1,v'} = \{H_{i-1}, v'\}$;

- $H_i = H_{i-1} - \mathcal{E}_s$: represents the event of adding the set of sequences $\mathcal{E}_s$ at time t_s.

As in the case of homochronous sequences, after computing the probabilities $p_\theta(D_{i,v}|D_{i-1,v'})$, and $\widetilde{q}_\theta(D_{i-1,v'}|D_{i,v})$ for a genealogy $G^{(j)}, j = 1, 2, \ldots, J$, the importance weights may be derived from equations (5.14), (5.15), (5.16), and (5.17).

Estimation of the effective population size for heterochronous sequences

For heterochronous sequences, the method for producing a *skywis plot* is similar to the one defined in Section 5.3.1. The main difference lies in the definition of epochs in this case. [1] In the presence of S serially sampled sequences, we cumulate the coalescence times as follows :

- for each simulated genealogy $\mathcal{G}^{(j)}$, we compute the MRCA time, $T^{(j)}_{MRCA}$, $j = 1, 2, ..., J$;
- we fix the number of epochs at $n^{(s)}_{\text{cum}}$ in each time interval $(t_s; t_{s-1})$ where no new sample is added, $s = 1, 2, ..., S$, $t_S = T^{(j)}_{MRCA}$, and $t_0 = 0$ (present);
- in order to define the epochs, the time cutting points in a genealogy $\mathcal{G}^{(j)}, j = 1, 2, .., J$ are computed as follows:

$$t^{(j,s)}_{\text{cut},b} = t_{s-1} + 0.1 \cdot \exp\left(\frac{b}{n_{\text{cum}}} \cdot \log(1 + 10 \cdot t_s)\right) - 0.1, \tag{5.33}$$

where $b = 1, 2, ..., n^{(s)}_{\text{cum}}$ and $s = 1, 2, ..., S - 1$.

For each genealogy and for each time interval $(t_s; t_{s-1})$, $s = 1, 2, ..., S$, formula (5.33) gives the limits of the epochs from the present to the past in units of N generations. In practice, we performed minor smoothing at times t_s, because the addition of new sequences creates an artificial discontinuity at t_s, $s = 1, 2, ..., S$.

---

[1] The reason we changed the way we define the epochs is that the number of sequences rises at the instants of the serial sampling, so the method used in Section 5.3.1 is not appropriate.

Therefore, the population size in the first epoch after t_s is set to be equal to the effective population size in the epoch preceding the addition of new sequences.

5.4 Results

To test the ability of our method to capture the demographic signal contained in the DNA sequences, we simulated several demographics scenarios. Further, we compared the results of the *skywis plot* with those of the *generalized skyline plot* that uses single tree, and the *Bayesian skyline plot* that uses MCMC approach. These methods are the closest to our approach.

The DNA sequences were simulated using the *fastsimcoal* program (Excoffier et Foll, 2011) which allows us to consider several demographic scenarios and different mutation models. The genealogies were simulated [2] using the method described in Section 5.3.1. In all our simulations, the coalescence times were cumulated into $n_{\text{cum}} = \sqrt{n} - 1$ epochs according to the method described in section 5.3.1, where n represents the number of simulated DNA sequences. After that, we derive the *skywis plot* using equations (5.24) and (5.25).

The *generalized skyline plot* was performed as follows:

1. From the DNA sequences generated by *fastsimcoal*, we estimated a phylogeny using the PHYLIP program (the PHYLogeny Inference Package (Felsenstein, 1989)) using the maximum likelihood method with a molecular clock constraint (we used the *dnamlk* program).

2. Based on the estimated tree produced by PHYLIP, we used the APE package

[2]The simulation of the genealogies was performed using MATLAB programming language (MATLAB, 2013) and the *Parallel Computing Toolbox* which allows parallelization of the simulation of genealogies.This is possible when using IS.

(Paradis *et al.*, 2004) to produce the *generalized skyline plot* according to the optimal strategy for grouping adjacent coalescent intervals introduced by Strimmer et Pybus (2001).

The *Bayesian skyline plot* was performed using the BEAST program, version 1.8.1. In order to reproduce a parametrization which is as close as possible to ours, we used (Hasegawa *et al.*, 1985) substitution model with equal base frequencies, and a strict clock with rate 1.

Below, we present our results according to the demographic models we considered.

5.4.1 Constant effective population size

In this case, we consider 50 simulated DNA sequences with parameters:

- number of nucleotides: 10000;
- constant effective population size: 2000;
- no recombination and no population structure;
- mutation rate equals to $2 \cdot 10^{-7}$: therefore ($\theta = 8$);
- JC69 (Jukes et Cantor, 1969) finite sites model.

The estimate of the effective population size (*skywis plot*) is shown in Figure 4(a). We observe that the *skywis plot* (orange line) gives a relatively smooth curve of the effective population size. Further, the estimation turns around the real value N, with a slight over-estimation close to the present, which can be explained by the fact that when the mutation rate θ is large, the sampled sequences are all different, and we have many mutations before one coalescence; thus, coalescence

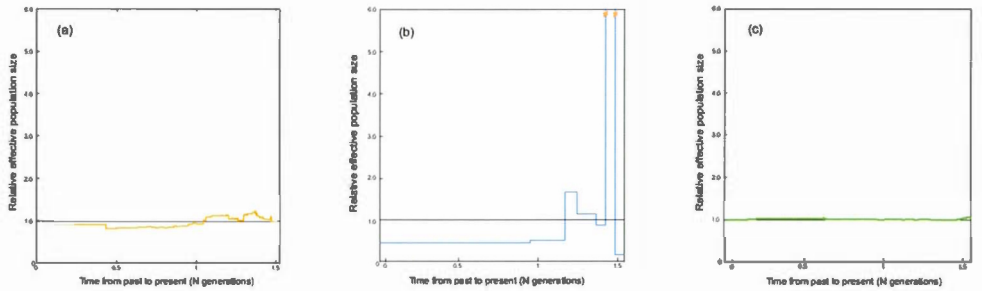


Figure 5.4: Constant effective population size: (a) *skywis plot*, (b) *generalized skyline plot*, (c) *Bayesian skyline plot*

times are longer, and the corresponding population sizes are larger (see section 5.3.1.)

In Figure 4(b) we present the *generalized skyline plot* . In this case, the form of the graph is not recognizable as a constant line.

The *Bayesian skyline plot* is given in Figure 4(c). In this case, the graph is very smooth and is easily recognizable as a constant line.

5.4.2 Piecewise constant function

In this section, we present results where 25 DNA sequences of length 10000 nucleotides and mutation rate $\mu = 5 \cdot 10^{-4}$ were simulated under the JC69 mutation model. We assume that the effective population size follows the piecewise constant model function :

$$N_e(t) = \begin{cases} N & \text{if } t < x \\ aN & \text{otherwise,} \end{cases} \tag{5.34}$$

where $N = N(0) = 10^4, x = 5000$ generations, $a = 0.25$ (see Figure 5.5), and the time t is measured from present to the past.

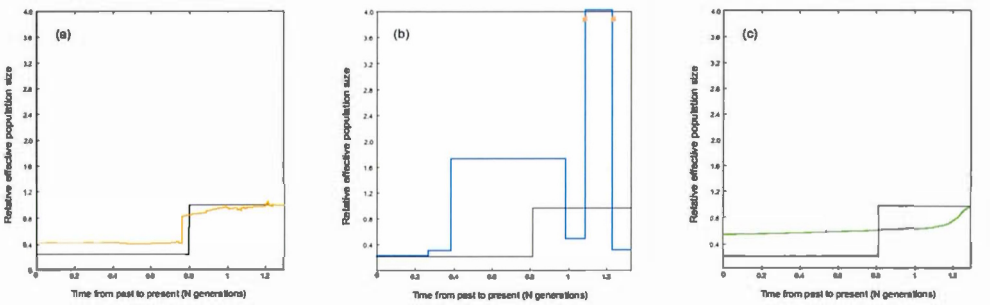


Figure 5.5: *Skywis plot* for data simulated from the population model where $N(t) = 10000$, if $t < 5000$ generations, and $N(t) = 2500$ otherwise (time from the past to the present) : (a) *skywis plot*, (b) *generalized skyline plot*, (c) *Bayesian skyline plot.*

Figure 5(a) represents the non-parametric estimate (*skywis plot*) of the effective population size for a number of epochs equal to $n_{\text{cum}} = 4$. We note that the *skywis plot* was able to detect well enough the change-point of the size of the actual population which dates back 5000 generations. However, the *skywis plot* seems to overestimate the effective population size for $t > 5000$ generations. In figure 5(b) we present the *generalized skyline plot.* The *skywis plot* gives a better result than the *generalized skyline plot* close to the present, while the estimate given by the *generalized skyline plot* is closer to the true value when we approach the MRCA.

The *Bayesian skyline plot* presented in Figure 5(c) is very smooth and generally reproduces the true history except closer to the present, where the Bayesian skyline plot over-smoothes the effective population size.

5.4.3 Exponential population growth

In this section, we suppose that the effective population growth is exponential assuming an instantaneous growth rate that is proportional to the current population size according to the equation $N_e(t) = N \exp(-\beta t)$ from present to the past where t is in units of generations.

Using the *fastsimcoal* program, we simulate 50 DNA sequences with the following parameters:

- Number of nucleotides : 1000;
- $N = N_e(0)$ at time $t = 0$: 10000;
- no recombination, and no population structure;
- mutation rate : $5 \cdot 10^{-7}$ ($\theta = 1$);
- JC69 finite sites model;
- $\beta = 1$.

The *skywis plot* for the simulated DNA sequences from the exponential model described above is given in Figure 6(a).

The result given in Figure 6(a) is quite good in the sense that the size of the effective population decreases steadily from the present to the past and follows the exponential curve quite closely most of the time. However, we note that the estimated effective population size is almost constant from some point in time when approaching the TMRCA. This is explained by the fact that for the last two sequences the theoretical average time to coalesce represents half the length of the tree, and from this point in time there is no much variability in the estimate of

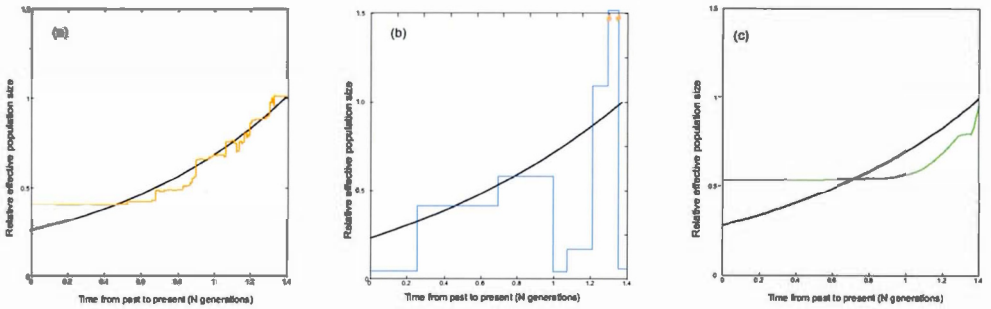


Figure 5.6: *Skywis plot* for DNA sequences simulated from an exponential model with $\beta = 1$: (a) *skywis plot*, (b) *generalized skyline plot*, (c) *Bayesian skyline plot.*

the population size. In particular, this remark led us to consider heterochronous sampling in order to improve the effective population size estimate.

In Figure 6(b) the time is measured in substitution units and we present the *generalized skyline plot.* As before, the *generalized skyline plot* has a fluctuating shape but it exhibits a certain tendency to decrease toward the past. In the end, when we approach the time of the MRCA, the *generalized skyline plot* gives an estimate that is close to the true value.

In Figure 6(c), we present the *Bayesian skyline plot.* As in the other scenarios, the *Bayesian skyline plot* produces a very smooth curve; in this case it suggests that the population had a mild exponential expansion. However, we note that the curve remains constant closer to the MRCA.

5.4.4 Exponential population growth and heterochronous sequences

In order to test the methodology proposed in Section 5.3.2, we use the same parameters as in Section 5.4.3, but by assume that the 50 sequences were collected at different moments in time such as:

- $n_0 = 25$ (present);
- $n_1 = 15$ at time $t_1 = 0.5$ in units of N generations (measured from present to the past);
- $n_2 = 10$ at time $t_1 = 1$ (N generations).

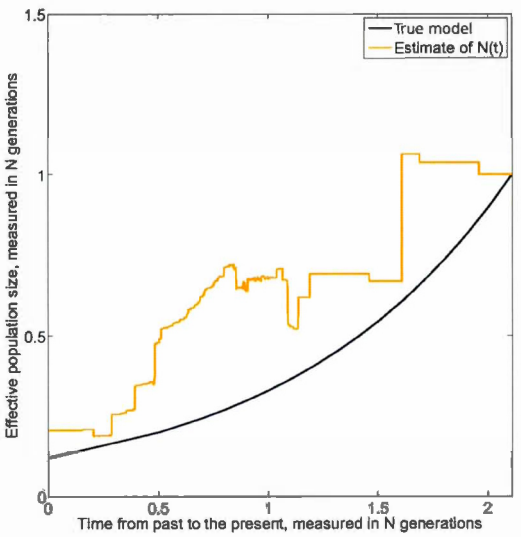


Figure 5.7: *Skywis plot* for DNA sequences simulated from an exponential model with 3 serial samples at times $t_0 = 0$, $t_1 = 0.5$, $t_1 = 1$ (in units of N generations) from the present to the past, and $\beta = 1$.

The result given in Figure 5.7 suggests that the effective population decreases exponentially from present to the past. Further, we note that the estimated effective population size continues to decrease when approaching the time of the MRCA, which is a net improvement over the homochronous case. This could be explained by the fact that as more sequences are added over time, more information is available as one approaches the MRCA.

5.5 Discussion

The *skywis plot* is a new flexible method for exploring the demographic history of a sample of DNA sequences based on coalescent theory. Our nonparametric method is likelihood-based and uses Importance Sampling. More precisely, we generate a large number of genealogies, both their times and their topology; further, we use the importance weights of these genealogies to compute a weighted average of the effective population size per epoch. This allows us to produce estimates that exhibit clear cut population growth tendencies over time, which is the main purpose of this approach, given that it is nonparametric. In practice, we expect our method to be used as a preliminary procedure that could be supplemented by a parametric analysis.

We present a framework of the new method and test by simulation its ability to capture the demographic signal contained in the DNA sequences under several demographic scenarios. Moreover, we consider both homochronous and heterochronous data using a simple substitution model, JC69 (Jukes et Cantor, 1969). We could also have considered more complicated substitution models, except those that allow variation in evolutionary rates among lineages.

For illustration we present the results given by the generalized skyline plot that uses a single genealogy, and those obtained by the *Bayesian skyline plot* that uses an MCMC approach. Although the *Bayesian skyline plot* is smoother than the *skywis plot*, our estimator is able to capture the shape of the effective population size Ne(t), as well as its main change points, but in some examples it had a (slight) tendency to overestimate the population size as we approached the MRCA. This is not surprising, given that the simulation our estimation method entails first setting a constant population size (where coalescence times are longer) and further operating an adjustment through a weighting system. Further, note that,

unlike the methods based on a single tree, it is possible to extend the *skywis plot* and include recombination. Indeed, recombination induces a graph structure rather than a tree, and IS methods in this context already exist (e.g. Fearnhead et Donnelly (2001)).

As a future development, we expect the method to be improved by considering an iterative procedure, in which the present approach would be the first estimation step. As a new approach, the *skywis plot* remains to be tested on more complex demographic models, and models of substitution that could be more realistic, especially for rapidly evolving RNA viruses. Also, the *skywis plot* can be easily extended to include multilocus data because, when there is no recombination, the same importance sampling scheme can be applied.

CHAPTER VI

DEUXIÈME ARTICLE: INFERRING THE DEMOGRAPHIC HISTORY FROM DNA SEQUENCES: AN IMPORTANCE SAMPLING APPROACH BASED ON NON-HOMOGENEOUS PROCESSES

Sadoune Ait Kaci Azzou, Fabrice Larribe, Sorana Frona

Abstract

In Ait Kaci Azzou *et al.* (2015) we introduced an Importance Sampling (IS) approach for estimating the demographic history of a sample of DNA sequences, the *skywis plot*. More precisely, we proposed a new nonparametric estimate of a population size that changes over time. We showed on simulated data that the *skywis plot* can work well in typical situations where the effective population size does not undergo very steep changes. In this paper, we introduce an iterative procedure which extends the previous method and gives good estimates under such rapid variations. In the *iterative calibrated skywis plot* we approximate the effective population size by a piecewise constant function, whose values are re-estimated at each step. These piecewise constant functions are used to generate the waiting times of non homogeneous Poisson processes related to a coalescent process with mutation under a variable population size model. Moreover, the present IS procedure is based on a modified version of the Stephens et Donnelly (2000) proposal distribution. Finally, we apply the *iterative calibrated skywis Plot* method to a sim-

ulated data set from a rapidly expanding exponential model, and we show that the method based on this new IS strategy correctly reconstructs the demographic history.

Keywords: non-homogeneous processes, importance sampling, effective population size, *calibrated skywis plot*, coalescent process.

6.1 Introduction

The methods for estimating the demographic history from gene sequences using coalescent theory can be classified into two categories: parametric and nonparametric. The parametric methods require a definite analytic demographic model which describes the changes in the population size. Such methods are typically based on importance sampling or Markov Chain Monte Carlo (MCMC) sampling and infer the demographic history by estimating the demographic parameters (Slatkin et Hudson, 1991; Kuhner *et al.*, 1998; Drummond *et al.*, 2002). Because it is not known in advance which demographic model fits the sampled gene sequences, nonparametric and semi-parametric methods for inferring the demographic history from sequence data or from an estimated genealogy have been developed (e.g. Pybus et al., 2000; Fu, 1994). In practice, the result of a nonparametric method can be used as a preliminary estimate that could be supplemented by a parametric analysis.

Among the most known nonparametric methods using coalescence theory to estimate the effective population size, we mention the family of *skyline plot* methods that was introduced by Pybus *et al.* (2000), referred in the literature as the *classical skyline plot*. This first method produces a piecewise reconstruction of the demographic history which is considered as quite noisy. Therefore, several extensions have been proposed in order to improve the performance of the estimator of the demographic history. In what follows, we mention the most popular among

them.

Strimmer et Pybus (2001) developed the *generalized skyline plot* estimate based on the Akaike Information Criterion correction (AIC) by allowing to cumulate multiple coalescent events. Later, important developments were obtained in a Bayesian framework. Thus, Drummond *et al.* (2005) and R. Opgen-Rhein (2005) use multiple change-point (MCP) models to estimate population size dynamics. Also, in order to achieve temporal smoothing of the effective population size, Minin *et al.* (2008) propose an alternative to change-point modeling that resorts to Gaussian Markov random fields. This last method does not require to set a prior total number of change points. Finally, Heled et Drummond (2008) introduce the *extended Bayesian skyline plot*, which permits the analysis of multiple unlinked loci, leading to an improvement in the reliability of the demographic inference and a substantial reduction in estimation error.

Ait Kaci Azzou *et al.* (2015) developed a new method (*skywis plot*) based on coalescent theory in order to estimate the effective population size. They are using an efficient importance sampling scheme where the estimate comes to an average over a large number of simulated genealogies. In this approach, one computes a weighted average of the effective population sizes on specific time intervals (*epochs*), where the genealogies that better agree with the data are given more weight. Moreover, Ait Kaci Azzou *et al.* (2015) illustrated by simulation that the *skywis plot* correctly reconstructs the recent demographic history under scenarios where the slope of the population size is not too steep.

In order to improve the performance of the *skywis plot*, it is possible to use the structure of heterochronous sequences (serial sampling) by introducing some information at the sampling times. A good example is given by Maretty *et al.*

(2013) who used the information about the Viral Load [1] (V) obtained at the time of sequencing in the case of serial sampling from patients infected with rapidly evolving viruses like HIV. Maretty *et al.* (2013) assumed a linear relationship[2] between the Viral Load V, and the effective population size N_e at the sampling times $t_0, t_1, ..., t_S$ ($N_e = \lambda V$). Thus, between two successive sampling times, it is possible to estimate N_e so that $N_{t_i} = (V_{t_{i-1}} + V_{t_i})/2$ assuming neutral evolution. The *skywis plot* can be refined by using this type of additional information in the case of serial sampling.

In the case of homochronous sequences and rapidly evolving population size, one can improve the skywis plot by considering an iterative procedure, in which the skywis plot would be the first estimation step and it would provide the starting values for this iterative method.

6.2 Preliminaries

6.2.1 Coalescent theory

Case of constant population size

Coalescent theory allows one to produce genealogies relating the sampled sequences according to a large class of population genetic models. In its simplest form, the coalescent process (Kingman, 1982b) provides a model for the genealogy assuming a single, isolated and panmictic population (e.g. a Wright-Fisher model), and a constant population size; for more details see, for example, Nordborg (2007), Hein *et al.* (2005), and Wakeley (2008). This classical coalescent framework can

---

[1] For example, in the case of the HIV virus, the viral load is the amount of HIV in a sample of blood. When the viral load is high, a patient has more HIV in in his body.

[2] For more information about the relationship between the viral load and the effective population size, see for example Gutierrez *et al.* (2012)

be extended to include simple deviations from the idealized Wright-Fisher model, like recombination, fluctuating population size, population structure, and selection. In what follows, we focus on a single extension of the coalescent, namely variable population size.

Case of variable population size

The case of non-constant (variable) population size requires to introduce the concept of effective population size, $N_e(t)$. If the population is constant in time, N, then $N_e(t) \equiv N$.

The effective population size reflects the number of individuals that contribute offsprings to the descendant generation and is almost always smaller than the census population size.

The variable population size coalescent model for contemporary gene sequences has been introduced by Griffiths et Tavaré (1994a) and Donnelly et Tavaré (1995). In this case, the coalescence times $T_2, T_3, ..., T_n$ do not follow independent exponential distributions.

In what follows, we let $N_e(0) = N$ and we assume that, relative to the population size N at time 0, the size of the population time t units ago is $\nu(t)$.
Let $V_k = T_n + ... + T_k$ be the cumulated waiting time so that the number of sequences pass from n to $k-1$ sequences, i.e.

$$V_k = \sum_{\ell=k}^{n} T_\ell, \tag{6.1}$$

and let $\Lambda(t)$ be the cumulative coalescent rate over time measured relative to the rate at time $t = 0$:

$$\Lambda(t) = \int_0^t \frac{1}{\nu(u)} du. \tag{6.2}$$

The survival function of the time T_k conditional on $V_{k+1} = v$ is

$$P(T_k > t \mid V_{k+1} = v) = \exp\left\{-\binom{k}{2}(\Lambda(t+v) - \Lambda(v))\right\}, \tag{6.3}$$

where $V_{n+1} = 0$.

In what follows, we present some theoretical aspects which led, for instance, to inferring the coalescence times distribution in the case of variable population size as in (6.3).

Let $A_n(t)$ be the process that counts the number of ancestors at time t of a sample of size n in the case of constant population size. It is known that $\{A_n(t), t \geq 0\}$ is a pure death process that moves from state k to state $k-1$ at rate $k(k-1)/2$ (Griffiths et Tavaré, 1994a). Further, let $\tilde{A}_n(t)$ be the process that counts the number of ancestors at time t of a sample of size n in the case of variable population size; then the process $\tilde{A}_n(t)$ jumps from $\tilde{A}_n(t) = k$ to $k-1$ at rate $\frac{k(k-1)}{2\nu(t)}$, and is a non-homogeneous death process. The non-homogeneous process $\{\tilde{A}_n(t), t \geq 0\}$ can be written as a function of $A_n(t)$ as follows (Tavaré et Zeitouni, 2004):

$$\tilde{A}_n(t) = A_n(\Lambda(t)),\ t \geq 0,$$

where

$$\Lambda(t) = \int_0^t \frac{1}{\nu(u)} du.$$

For example, if the effective population size decreases from the present to the past, then $\nu(t) \leq 1$ and $\Lambda(t) \geq t$. We have in this case that $\tilde{A}_n(t) \leq A_n(t)$. As a result

- the total time required to find the most recent common ancestor in a small population is shorter than in a large one;

- the topology of the tree in the case of variable population size is the same than in the constant population size, but its time scale has to be changed to account for the fluctuations in the population.

Let $p(t)$ be the probability that two genes coalesce in the previous generation; then $p(t) = 1/N(t)$ where $N(t)$ in considered as the natural local scale of the coalescent process. Thus, one way to take into account the variability of the effective population size is to generate genealogies under the constant size coalescent process and then to stretch or compress time, according to whether $p(t)$ is smaller or larger than $p(0) = 1/N$ as argued in Hein *et al.* (2005). This principle is illustrated in Figure 6.1, and the time is measured in units of N generations.

Let $\nu(\Delta t)$ be the relative population size in an interval of length Δt , in the presence of k genealogies. The local coalescence rate in Δt is given by :

$$\binom{k}{2} \cdot \frac{1}{\nu(\Delta t)} = \binom{k}{2} \cdot R(\Delta t), \tag{6.4}$$

where

$$R(\Delta t) = \frac{\binom{k}{2} \times \frac{N}{N_e(\Delta t)}}{\binom{k}{2}}$$

In Figure 6.1 the bottom time axis represents the time corresponding to a constant effective population size, while the top time axis represents the rescaled time when one takes into account the relative population size on each interval Δt. Further, the relative coalescence rate is represented on each interval Δt.

For example, in the second time interval $[1, 2)$ (in N generations), on the constant population size time axis, we have $\nu(\Delta t) = 0.5$, and this means that the effective population size is half the one at time $t = 0$. It follows that

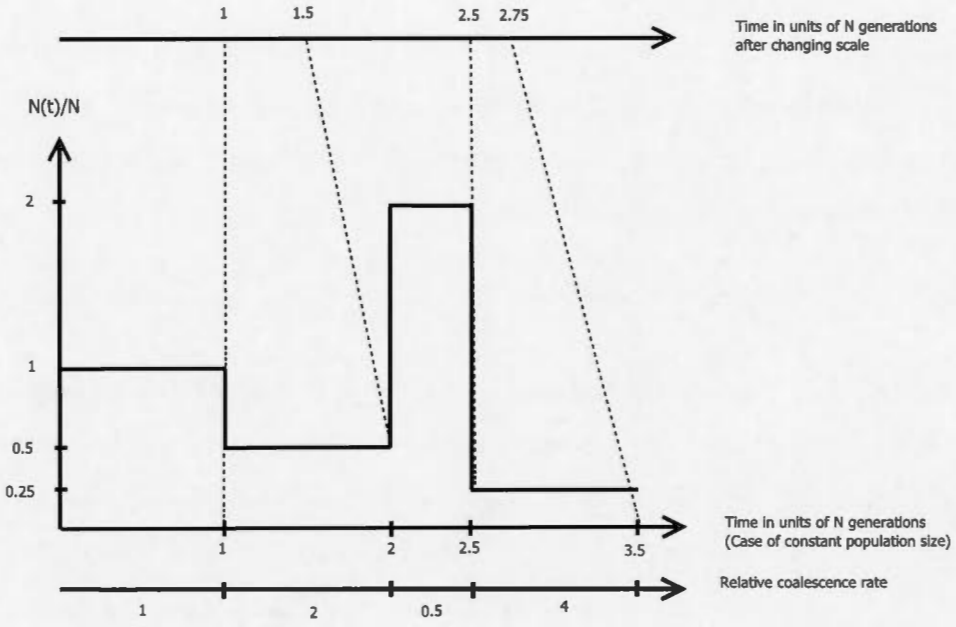


Figure 6.1: Stretching and compressing time in the coalescence process with variable population size.

- the sequences will coalesce faster, with a relative rate equal to 2: it is twice as high as in the case of a coalescent process with constant effective population size;

- therefore, the time must be divided by 2, to consider the value of $\nu(\Delta t)$ for the interval $[1, 2)$.

From this illustration, one can see that a variable population size has an impact on the local coalescence rate, and the time to the Most Recent Common Ancestor, MRCA (T_{MRCA}).

Indeed, in Figure 6.1 the T_{MRCA} is equal to 3.5 in the case of constant population size, while the T_{MRCA} is equal to 2.5 in the proposed demographic model of variable population size. In other words, all sequences coalesce faster under this model than in the constant population size case.

6.2.2 The *skywis plot* method

In this section, we remind the principle of the *skywis plot* estimate introduced in Ait Kaci Azzou *et al.* (2015), when n gene sequences are available at time $t = 0$. The main idea behind this approach is to simulate a large number of genealogies using the proposal distribution $Q(\cdot)$ introduced by Stephens et Donnelly (2000) and create a weighted average of the effective population sizes; the most probable genealogies receive a larger weight, where the weights are those given by the importance sampling method.

In short, reconstructing the demographic history from these sequences involves four distinct steps:

1. simulate J genealogies : $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$;
2. for each genealogy $\mathcal{G}^{(j)}$, cumulate the small coalescence times in order to obtain n_{cum} *epochs* as described in Ait Kaci Azzou *et al.* (2015); further, compute an estimate $\widehat{N}e_a^{(j)}$ for an epoch a, $a = 1, ..., n_{\text{cum}}$, and for each genealogy $\mathcal{G}^{(j)}, j = 1, 2, ..., J$.
3. compute the weights $w^{(1)}, w^{(2)}, ..., w^{(J)}$, $j = 1, 2, ..., J$ where $w^{(j)}$ represents the weight of the genealogy $\mathcal{G}^{(j)}$ in the likelihood of the data;

4. For each *epoch* a, $a = 1, ..., n_{\text{cum}}$, compute the estimate $\widehat{N}e_a$, based on genealogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$, as the weighted mean of $\widehat{N}e_a^{(j)}$, for $j = 1, 2, ..., J$.

The estimate $\widehat{N}e_a^{(j)}$ is given by:

$$\widehat{N}e_a^{(j)} = \Delta t_a^{(j)} \cdot \frac{d_a^{(j)} \left(d_a^{(j)} - c_a^{(j)} \right)}{2c_a^{(j)}}, \tag{6.5}$$

where $d_a^{(j)}$ is the number of sequences at the beginning of *epoch* a and $c_a^{(j)}$ is the number of cumulated coalescence times in *epoch* a, $a = 1, 2, ..., n_{\text{cum}}$, $j = 1, 2, ..., J$.

From (6.5), one can see that the estimate of the effective population size for an *epoch* a and for a genealogy $\mathcal{G}^{(j)}$, is proportional to the interval of length $\Delta t_a^{(j)}$ of cumulated coalescence times. Therefore, it is important to simulate correctly the time to the next event in order to have a good estimate of the demographic history from gene sequences. Further, when the population size is constant, the estimator given by equation (6.5) can be viewed as a method of moments estimator for the effective population size in the cumulated time interval of length $\Delta t_a^{(j)}$. For more details, see, for example, Rodrigo *et al.* (1999), Ait Kaci Azzou *et al.* (2015).

In the *skywis plot* method and in the presence of k sequences, we simulated the time to the next event (coalescence or mutation) as an exponential distribution with rate $\left(\binom{k}{2} + \frac{k\theta}{2} \right)$. As described in Section 6.2.1, this assumption is not true when the effective population size is variable. In the specific case where $N_e(t)/N$ is firstly approximated by a step function, as the behavior of the estimate is similar to the constant size case on each interval, it is possible to calibrate the time to adapt the shape of the genealogy in this context. The new method is called *calibrated skywis plot*, and is described in what follows: the new way of simulating the time to the next event is given in Section 6.3.2 and the adaptation of the proposal distribution $Q(\cdot)$ Stephens et Donnelly (2000) is given in Section 6.3.3.

6.3 The calibrated skywis plot method

6.3.1 Distribution of the next event time based on a Poisson process

In the case of a constant population size, the coalescent process can be viewed as a **sequence** of Poisson processes, where the rate changes from $\lambda_{\text{coa},k} = \binom{k}{2}$ to $\lambda_{\text{coa},k-1} = \binom{k-1}{2}$ when the number of sequences passes from k to $k-1$ (Ewens, 2004). Actually, the time to the next coalescence corresponds to the first waiting time of a Poisson process. This remark has been used in devising simulation techniques for the coalescence process.

Let $k \in \{n, n-1, ..., 2\}$ be the number of sequences. In the case of a constant population size, the number of events in the coalescent and mutation processes have been modeled as two independent homogeneous Poisson processes which we denote $M_{\text{coa},k}(t)$ and $M_{\text{mut},k}(t)$, with respective rates $\lambda_{\text{coa},k}$, and $\lambda_{\text{mut},k}$. Thus, in the presence of k sequences, the sum $M_{\text{Tot},k}(t) = (M_{\text{coa},k}(t) + M_{\text{mut},k}(t))$ is a Poisson process with parameter $(\lambda_{\text{coa},k}+\lambda_{\text{mut},k})$. Further, the waiting time until the next event in the process $M_{\text{Tot},k}(t)$ is the minimum of the waiting times until the next coalescence or mutation event, i.e. the minimum of the waiting times of the processes $M_{\text{coa},k}(t)$, and $M_{\text{mut},k}(t)$. We note also that in these processes, all waiting times are independent and exponentially distributed; further, for each event time, the probability that it is a coalescence or mutation event is, respectively,

$$\frac{\lambda_{\text{coa},k}}{(\lambda_{\text{coa},k} + \lambda_{\text{mut},k})} \quad \text{and} \quad \frac{\lambda_{\text{mut},k}}{(\lambda_{\text{coa},k} + \lambda_{\text{mut},k})}.$$

In the case of a variable population size, the coalescent process is non-homogeneous in time. Therefore, in the presence of k sequences, $k = n, n-1, ..., 2$, we note that:

1. the arrival of coalescence events can be viewed as a **sequence** of arrivals in

successive Poisson processes with rates depending on k

$$\lambda_{\text{coa},k}(t) = \binom{k}{2}/\nu(t), \tag{6.6}$$

where $\nu(t)$ is the relative population size $N_e(t)/N$;

2. the arrival of mutation events can be viewed as a **sequence** of arrivals in successive Poisson processes with rates depending on k

$$\lambda_{\text{mut},k}(t) = \frac{k\theta}{2}; \tag{6.7}$$

3. the total number of events

$$M_{\text{Tot},k}(t) = M_{\text{coa},k}(t) + M_{\text{mut},k}(t),$$

is a non-homogeneous Poisson process with rate $(\lambda_{\text{coa},k}(t) + \lambda_{\text{mut},k})$;

4. the inter-arrival times of the coalescent process, are neither independent, nor exponentially distributed; the arrival time of the next coalescence is distributed as the first waiting time of a non-homogenous Poisson process of rate $\lambda_{\text{coa},k}(t)$;

5. once the arrival time of the next event is known, the event is either a coalescence or a mutation, with respective probabilities (see, e.g. Ross (1995), ch.5)

$$\frac{\lambda_{\text{coa},k}(t)}{(\lambda_{\text{coa},k}(t) + \lambda_{\text{mut},k})}, \quad \frac{\lambda_{\text{mut},k}}{(\lambda_{\text{coa},k}(t) + \lambda_{\text{mut},k})}. \tag{6.8}$$

6.3.2 Simulation of the next event times according to non-homogeneous Poisson processes

We base our simulation on the theory described in section 6.3.1.

A Poisson process is specified through its intensity function $\lambda(t)$, or by the function $m(t) = \int_0^t \lambda(u)du$. In order to simulate the next event time, we settled for the

inversion method because in our case it is easy to invert $m(t)$. (The method is described in Ross (2013), for instance).

Let Wt_{-i} be the waiting (inter-arrival) time until the next event (time measured from the present to the past), and let $Tc_{-i} = \sum_{\ell=-i+1}^{-1} Wt_\ell$ be the cumulated waiting time.

The survival function of the next event time Wt_{-i} conditional on the cumulated time Tc_{-i} is given by

$$\Pr(Wt_{-i} > t \mid Tc_{-i} = v) = \exp\left[-\left(m(v+t) - m(v)\right)\right]. \tag{6.9}$$

In view of (6.9), (6.6) and (6.7), and in the presence of k sequences,

$$m(t) \equiv m_k(t) = \int_0^t \lambda_k(u)du = \int_0^t \left(\lambda_{\text{coa},k}(u) + \lambda_{\text{mut},k}(u)\right) du. \tag{6.10}$$

Thus, in the presence of k sequences

$$\begin{aligned} m_k(t) &= \int_0^t \left(\binom{k}{2}\frac{1}{\nu(u)} + \frac{k\theta}{2}\right) du \\ &= \int_0^t \left(\binom{k}{2}\frac{1}{\nu(u)}\right) du + \frac{k\theta}{2}t, \end{aligned} \tag{6.11}$$

which gives

$$\begin{aligned} 1 - F(t)_{\{Wt_{-i}|Tc_{-i}=v\}} &= \Pr(Wt_{-i} > t \mid Tc_{-i} = v) \\ &= \exp\left[-\left(\int_v^{v+t} \left(\binom{k}{2}\frac{1}{\nu(u)}\right) du + \frac{k\theta}{2}t\right)\right]. \end{aligned} \tag{6.12}$$

The simulation of the next event time Wt_{-i} by inversion comes to solving the following equation in t

$$F(t)_{\{Wt_{-i}|Tc_{-i}=v\}} = u \Leftrightarrow 1 - F(t)_{\{Wt_{-i}|Tc_{-i}=v\}} = 1 - u, \tag{6.13}$$

where u is an observed value of a uniform variable $U[0,1]$.

Equation (6.13) can be rewritten as

$$\begin{aligned} \exp\left[-\left(\int_v^{v+t}\left(\binom{k}{2}\frac{1}{\nu(u)}\right)du+\frac{k\theta}{2}t\right)\right] &= 1-u \\ \Leftrightarrow \left(\int_v^{v+t}\left(\binom{k}{2}\frac{1}{\nu(u)}\right)du+\frac{k\theta}{2}t\right) &= -\ln(1-u). \end{aligned} \tag{6.14}$$

Solving equation (6.14) in t depends on the intervals to which v and $v+t$ belong.

In what follows, we assume that $N(t)$ is approximated by a step function, where the time is split into $(C+1)$ intervals using the cutting points $b_0 = 0$ (present), $b_1, b_2, ..., b_C$; $b_{C+1} = T_{MRCA}$. In a practical context, this comes to assuming that the relative effective population size denoted $\nu(I_c) = \delta_c$, has been previously estimated on each interval I_c, $c = 1, 2, ..., C+1$, where $I_c = [b_{c-1}; b_c)$.

In order to illustrate how the times are simulated, let $C = 3$, and let, for example, $v < b_1$; then, $(v+t)$ can belong to $[0; b_1)$, $[b_1; b_2)$, $[b_2; b_3)$, or $[b_3; T_{\text{MRCA}}]$. Assume, for illustration, that $v + t \in [b_1; b_2)$. Then, equation (6.14) becomes

$$\begin{aligned} &\int_v^{v+t}\left[\binom{k}{2}\frac{1}{\nu(u)}+\frac{k\theta}{2}\right]du \\ &\quad=\int_v^{b_1}\left(\binom{k}{2}\frac{1}{\delta_1}+\frac{k\theta}{2}\right)du+\int_{b_1}^{v+t}\left(\binom{k}{2}\frac{1}{\delta_2}+\frac{k\theta}{2}\right)du \\ &\quad=\binom{k}{2}\frac{1}{\delta_1}(b_1-v)+(t+v-b_1)\binom{k}{2}\frac{1}{\delta_2}+\frac{k\theta}{2}t, \end{aligned} \tag{6.15}$$

and solving equation (6.14) is equivalent to solving in t :

$$\binom{k}{2}\frac{1}{\delta_1}(b_1-v)+(t+v-b_1)\binom{k}{2}\frac{1}{\delta_2}+\frac{k\theta}{2}t=-\ln(1-u). \tag{6.16}$$

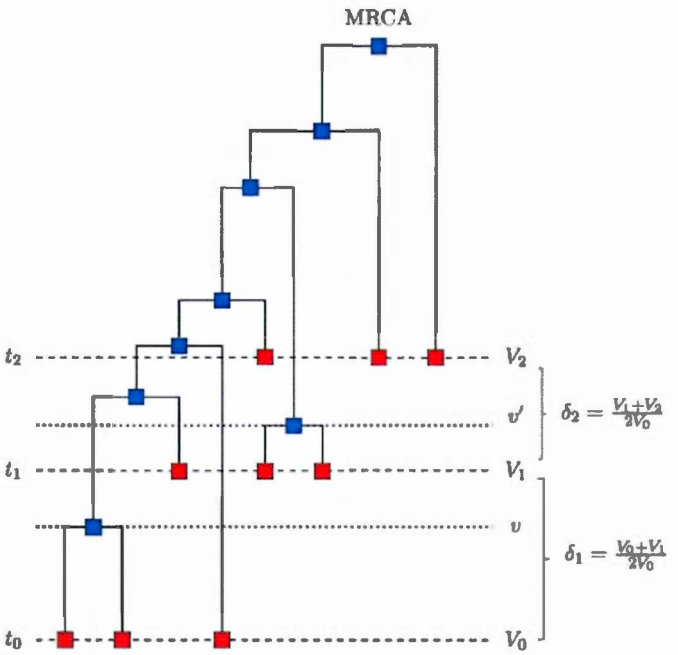


Figure 6.2: Example of serially sampled sequences with $S = 3$. The red squares are the sampled sequences and the blue squares are the sequences derived from coalescence; δ_j represents the estimate of the relative population size on the inter-sampling time intervals $[t_{j-1}; t_j)$, $j = 1, 2$.

Developing equation (6.16) gives :

$$\begin{aligned} t &= \frac{(b_1 - v)\binom{k}{2}\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right)}{\binom{k}{2}\frac{1}{\delta_2} + \frac{k\theta}{2}} - \frac{\ln(1-u)}{\binom{k}{2}\frac{1}{\delta_2} + \frac{k\theta}{2}} \\ &= \frac{(b_1 - v)\binom{k}{2}\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right)}{C_{\delta_2}} - \frac{\ln(1-u)}{C_{\delta_2}}, \end{aligned} \tag{6.17}$$

where

$$C_{\delta_c} = \binom{k}{2}\frac{1}{\delta_c} + \frac{k\theta}{2}, c = 1, 2, \ldots, C+1. \tag{6.18}$$

When simulating the next event time, we know v, u, and $-\ln(1-u)$, but we do not know the location of $v+t$. Therefore, we have to identify the domains in u which correspond to domains in $(v+t)$. In our example, where $v < b_1$ and $b_1 \leq v+t \leq b_2$, one can see that $b_1 \leq v+t < b_2$ is equivalent to :

$$(b_1 - v) \leq \frac{(b_1 - v)\binom{k}{2}\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right)}{C_{\delta_2}} - \frac{\ln(1-u)}{C_{\delta_2}} < (b_2 - v), \tag{6.19}$$

after t is replaced by its value in (6.17).

Thus, (6.19) comes to :

$$(b_1 - v) - \frac{(b_1 - v)\binom{k}{2}\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right)}{C_{\delta_2}} \leq -\frac{\ln(1-u)}{C_{\delta_2}} < (b_2 - v) - \frac{(b_1 - v)\binom{k}{2}\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right)}{C_{\delta_2}}. \tag{6.20}$$

Similar calculations can be done for each interval to which v and $v+t$ belong. Thus, Table 6.3 represents the formulas to be used in order to simulate the next event time when approximating the effective population size by a model with a constant population size per interval I_c, $c = 1, 2, 3, 4$, while the Table 6.4 summarizes the corresponding domains. Further, the details of the simulation with these values are given in Appendix A.

The previous computing can be easily extended to any number of intervals. This generalization is based on the following Proposition.

Proposition 1. *Let v be the observed cumulated time in the $M_{Tot,k}(t)$ process defined in Section 6.3.1, and let t be a realization of the next event time.*

Let

$$\frac{\widetilde{N}(t)}{N} = \delta_c I_{\{b_{c-1} \le t < b_c\}},\ c = 1, 2, ..., C+1,$$

($b_{C+1} = T_{MRCA}$) be an approximation to the effective population size, and $M_{Tot,k}$ its associated process and its waiting and total times (Wt_i, Tc_i), defined in Section 3.1, $k = n, n-1, \ldots, 2$. Further, fix to k the number of sequences,

let

$$C_{\delta_j} = \binom{k}{2}\frac{1}{\delta_j} + \frac{k\theta}{2}, j = 1, 2, \ldots, C+1,$$

and consider the conditional probability $\Pr(Wt_i > t | Tc_i = v)$ and the integrals given in (6.14), with $b_{\ell-1} \le v < b_\ell, b_{j-1} \le v + t < b_j, 1 \le \ell \le j \le C+1$. Then, for $j \ge \ell + 2$, the following results hold:

1. *The integral*

$$\int_v^{v+t} \left\{ \binom{k}{2}\frac{1}{\nu(u)} + \frac{k\theta}{2} \right\} du \tag{6.21}$$

can be written as

$$\binom{k}{2}\left\{ \frac{b_\ell - v}{\delta_\ell} - \frac{b_{j-1} - v}{\delta_j} + \sum_{i=\ell+1}^{j-1} \frac{b_i - b_{i-1}}{\delta_i} \right\} + t \cdot C_{\delta_j}.$$

2. *The value E^* of the integral in (6.21) satisfies the following inequalities:*

$$E^* \ge (b_{j-1} - v)C_{\delta_j} + \binom{k}{2}\left\{ \frac{(b_\ell - v)}{\delta_\ell} - \frac{(b_{j-1} - v)}{\delta_j} + \sum_{i=\ell+1}^{j-1} \frac{(b_i - b_{i-1})}{\delta_i} \right\}$$

and

$$E^* < (b_j - v)C_{\delta_j} + \binom{k}{2}\left\{ \frac{(b_\ell - v)}{\delta_\ell} - \frac{(b_{j-1} - v)}{\delta_j} + \sum_{i=\ell+1}^{j-1} \frac{(b_i - b_{i-1})}{\delta_i} \right\}.$$

Proof.

1. Given that the population size is piecewise constant, the integral $\int_v^{v+t} \frac{1}{\nu(u)} du$ can be decomposed as follows:

$$\begin{aligned} &\int_v^{b_\ell} \frac{1}{\nu(u)} du + \int_{b_\ell}^{b_{j-1}} \frac{1}{\nu(u)} du + \int_{b_{j-1}}^{v+t} \frac{1}{\nu(u)} du = \\ &\frac{b_\ell - v}{\delta_\ell} + \sum_{i=\ell+1}^{j-1} \frac{b_i - b_{i-1}}{\delta_i} + \frac{v + t - b_{j-1}}{\delta_j} = \\ &\frac{b_\ell - v}{\delta_\ell} - \frac{b_{j-1} - v}{\delta_j} + \sum_{i=\ell+1}^{j-1} \frac{b_i - b_{i-1}}{\delta_i} + \frac{t}{\delta_j}, \end{aligned} \tag{6.22}$$

for all $j-1 \geq \ell+1$. The final result is proven by noting that $\int_v^{v+t} (k\theta/2)\, du = (k\theta/2)t$ and grouping the terms in t.

Special cases are:

- $j = \ell$, where the integral (6.21) reduces to

$$\int_v^{v+t} \frac{1}{\nu(u)} du = \frac{b_\ell - v}{\delta_\ell} + \frac{t}{\delta_j};$$

in particular, if $\ell = C + 1$ then $j = C + 1$;

- $j = \ell + 1$ where the integral (6.21) reduces to

$$\int_v^{v+t} \frac{1}{\nu(u)} du = \frac{b_\ell - v}{\delta_\ell} - \frac{b_\ell - v}{\delta_{\ell+1}} + \frac{t}{\delta_j}.$$

2. Under our assumptions, t satisfies the inequalities

$$b_{j-1} - v \leq t < b_j - v$$

and the result follows obviously from 1. ■

The previous proposition allows to solve (6.14) in the general case (see Corollary 1 in Appendix B).

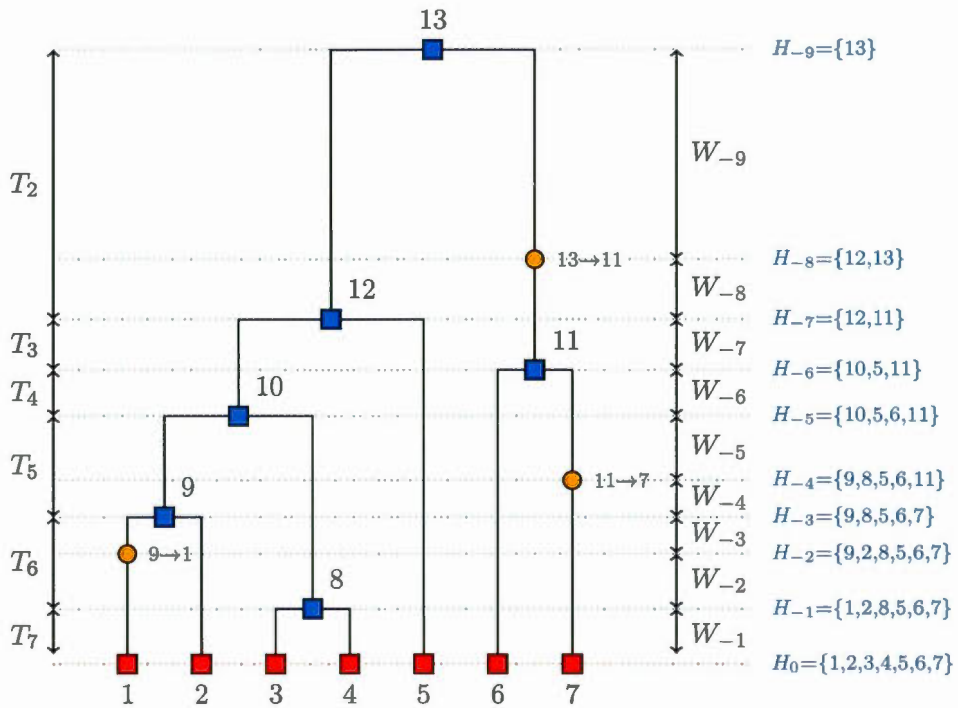


Figure 6.3: Example of a realization of the coalescent process viewed from the past to the present with $n = 7$ sequences (red squares); there are 6 coalescence events (blue squares) and 3 mutation events (orange circles).

6.3.3 An adapted version of the Stephens et Donnelly (2000) proposal distribution, where the population size is approximated by a constant per interval I_c, with known δ_c, $c = 1, 2, ..., C + 1$.

In this section, we propose to adapt the proposal distribution of Stephens et Donnelly (2000) ($S\&D$) to the case where $N_e(t)/N$ is piecewise-constant, and the ratios (but not N) are known. This piecewise-constant function can correspond to an approximation of a more general population size $N_e(t)$. The full practical context is described in Section 1, for cases where the ratios are indeed fully known,

or preliminary estimated.

We introduce some notations that will be used subsequently:

- E: the set of possible types of gene sequences;
- H_{-i}: the set of all sequences when event i occurs (coalescence or mutation) where i decreases from the present to the past in steps of 1 for each event (See Figure 6.3);
- $\mathcal{H} = \{H_0, H_{-1}, ..., H_{-\tau}\}$: a history of sequences where $H_0 = \mathcal{D}$, τ is the total number of events in the history $\mathcal{H}$, and $H_{-\tau}$ is a singleton (the MRCA);
- $\mathbf{P}$: the mutation transition matrix.

Note that H_{-i} are viewed as states of a Markov process starting at the genetic type $H_{-\tau} \in E$ and ending with $H_0 \in E$.

We showed in the previous section that the Poisson processes that describe the number of events (coalescence or mutation) are non-homogeneous because the rates $\lambda(t)$ depend on time. Thus, it is necessary to include the instants v and v' in computing the jump probabilities *forward*, and the estimate of their *backward* counterparts. We consider that:

- wt_{-i}: is the next event time;
- $v = \sum_{\ell=-i+1}^{-1} wt_\ell$: is the observed cumulated time (time measured from the present to the past) until the last event. Let $v \in I_{\ell_1}$, $\ell_1 = 1, 2, ..., C$ such as $I_{\ell_1} = [b_{\ell_1 - 1}; b_{\ell_1})$;
- the time $v' = v + wt_{-i}$, is such that $v' \in I_{\ell_2} = [b_{\ell_2 - 1}; b_{\ell_2})$, with $\ell_2 \geq \ell_1$.

Let $D_{i,v} = \{H_i, v\}$ and $D_{i-1,v'} = \{H_{i-1}, v'\}$, where:

- $D_{i,v} = \{H_i, v\}$ represents the set of all sequences present in the population after the i^{th} event at time v (measured from the present to the past);
- $D_{i-1,v'} = \{H_{i-1}, v'\}$ represents the set of all sequences present in the population after the $(i+1)^{\text{th}}$ event at time v' (measured from the present to the past).

Thus, we define :

1. $p_\delta(D_{i,v} \mid D_{i-1,v'})$, the probability to pass from the state $D_{i-1,v'} = \{H_{i-1}, v'\}$ to the state $D_{i,v} = \{H_i, v\}$ (from the past to the present);
2. $\hat{q}_\delta(D_{i-1,v'} \mid D_{i,v})$, the estimated probability to pass from $D_{i,v} = \{H_i, v\}$ to $D_{i-1,v'}$=$\{H_{i-1}, v'\}$ (from the present to the past). This last probability is used for simulating the genealogies $\mathcal{G}^{(j)}$, $j = 1, ..., J$.

Since the Poisson process that describes the number of events (coalescence or mutation) is non-homogeneous, the probability that the next event is a coalescence or a mutation is different if we pass from $D_{i,v}$ to $D_{i-1,v'}$ (*backwards*) or from $D_{i-1,v'}$ to $D_{i,v}$ (*forwards*). Consider first the backward moves.

- Starting from the state $D_{i,v}$ with $v \in I_{\ell_1}$, the event $D_{i-1,v'}$ with $v' \in I_{\ell_2}$ is a coalescence with probability

$$\frac{\lambda_{\text{coa},\, n_i}(v')}{\lambda_{\text{coa},\, n_i}(v') + \lambda_{\text{mut},\, n_i}(v')} = \frac{\binom{n_i}{2}\frac{1}{\delta_{\ell_2}}}{\binom{n_i}{2}\frac{1}{\delta_{\ell_2}} + \frac{n_i\theta}{2}} = \frac{n_i - 1}{n_i - 1 + \theta\delta_{\ell_2}} \tag{6.23}$$

  and a mutation with probability

$$1 - \frac{n_i - 1}{n_i - 1 + \theta\delta_{\ell_2}} = \frac{\theta\delta_{\ell_2}}{n_i - 1 + \theta\delta_{\ell_2}}, \tag{6.24}$$

where n_i, represents the number of sequences at the state $D_{i,v}$ and n_{i-1} represents the number of sequences at the state $D_{i-1,v'}$.

- However, if we consider the forward process and moves from $D_{i-1,v'}$ to $D_{i,v}$ with $v' \in I_{\ell_2}$ and $v \in I_{\ell_1}$, the probabilities change. For a chosen sequence in E, we have a split with probability

$$\frac{\lambda_{\text{coa},\, n_{i-1}}(v)}{\lambda_{\text{coa},\, n_{i-1}}(v) + \lambda_{\text{mut},\, n_{i-1}}(v)} = \frac{\binom{n_{i-1}}{2}\frac{1}{\delta_{\ell_1}}}{\binom{n_{i-1}}{2}\frac{1}{\delta_{\ell_1}} + \frac{n_{i-1}\theta}{2}} = \frac{n_{i-1} - 1}{n_{i-1} - 1 + \theta\delta_{\ell_1}}, \tag{6.25}$$

and a mutation into another sequence in E with probability

$$1 - \frac{n_{i-1} - 1}{n_{i-1} - 1 + \theta\delta_{\ell_1}} = \frac{\theta\delta_{\ell_1}}{n_{i-1} - 1 + \theta\delta_{\ell_1}}, \tag{6.26}$$

where n_{i-1}, represents the number of sequences at the state $D_{i-1,v'}$.

Based on equations (6.25) and (6.26), we can compute the forward probability to pass from the state $D_{i-1,v'}$ to the state $D_{i,v}$ in the presence of n_{i-1} sequences:

$$p_\delta(D_{i,v} \mid D_{i-1,v'}) = \begin{cases} \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{\delta_{\ell_1}\theta}{n_{i-1} - 1 + \delta_{\ell_1}\theta} P_{\alpha\beta} & \text{if } M_\alpha^\beta, v \in I_{\ell_1}, v' \in I_{\ell_2} \\ \frac{n_{i-1}^{(\alpha)}}{n_{i-1}} \frac{n_{i-1} - 1}{n_{i-1} - 1 + \delta_{\ell_1}\theta} & \text{if } C_\alpha^\alpha, v \in I_{\ell_1}, v' \in I_{\ell_2} \\ 0 & \text{otherwise}, \end{cases} \tag{6.27}$$

where:

- the times $v < v'$ are measured from the present to the past;
- $n_{i-1}^{(\alpha)}$ represents the number of sequences of type α in $D_{i-1,v'} = \{H_{i-1}, v'\}$;
- M_{α}^{β} denotes a mutation of a gene sequence from type α to type β according to $\mathbf{P}$;
- C_{α}^{α} denotes a coalescence (a split actually) of two gene sequences of type α;
- $P_{\alpha\beta}$ is the probability of a DNA sequence of type α to mutate to a DNA sequence of type β.

From equation (6.27), the factor

$$\frac{\delta_{\ell_1}\theta}{(n_{i-1} - 1 + \delta_{\ell_1}\theta)},$$

represents the probability that the next event is a mutation, see (6.26). Further,

$$\frac{n_{i-1} - 1}{n_{i-1} - 1 + \delta_{\ell_1}\theta},$$

represents the probability that the next event is a coalescence (see equation (6.25)). It can be shown that the sum of the first and the second line of the equation (6.27), summing over all sequences β to which the selected sequence α can mutate, gives $\frac{n_{i-1}^{(\alpha)}}{n_{i-1}}$.

In estimation, one progresses backward in time and, in what follows, we consider a new proposal distribution, whose fundamental difference with that of $(S\&D)$, is that the coalescence rate depends on the I_c, $c = 1, 2, ..., C + 1$ interval where the transition takes place.

We note that the special case where $\delta_c = 1$ for every $c = 1, 2, ...C + 1$, is exactly the $(S\&D)$ proposal distribution. Therefore, the probability to pass from a state to another one differs according to the intervals to which v and v' belong, as described by equation (6.27) for the *forward* probabilities.

Proposition 2. *Consider the backward process, where the transition probability to pass from the state $D_{i,v}$ to the state $D_{i-1,v'}$ (time measured from the present to the past) is given by :*

$$\hat{q}_\delta(D_{i-1,v'} \mid D_{i,v}) = \begin{cases} C_{\delta_{\ell_2}}^{-1} \dfrac{\delta_{\ell_2}\theta}{2} n_i^{(\alpha)} \dfrac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} P_{\beta\alpha} & \text{if } M_\beta^\alpha,\ v \in I_{\ell_1}, v' \in I_{\ell_2} \\ C_{\delta_{\ell_2}}^{-1} \binom{n_i^{(\alpha)}}{2} \dfrac{1}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} & \text{if } C_\alpha^\alpha,\ v \in I_{\ell_1}, v' \in I_{\ell_2} \\ 0 & \text{otherwise,} \end{cases} \tag{6.28}$$

where

- *$H_i - \alpha$ represents the set of all sequences in H_i without the selected sequence of type α;*
- *the proportionality constant $C_{\delta_{\ell_2}}$ is given by (see formula (6.18))*

$$C_{\delta_{\ell_2}} = \frac{n_i(n_i - 1 + \delta_{\ell_2}\theta)}{2\delta_{\ell_2}};$$

- *for a sample $\mathcal{D}$ of n sequences with $n^{(\alpha)}$ sequences of type α,*

$$\hat{\pi}_{\delta_{\ell_2}}(\beta \mid \mathcal{D}) = \sum_{\alpha \in E} \sum_{m=0}^{\infty} \frac{n^{(\alpha)}}{n} \left(\frac{\delta_{\ell_2}\theta}{n + \delta_{\ell_2}\theta} \right)^m \frac{n}{n + \delta_{\ell_2}\theta} (P^m)_{\alpha\beta} \tag{6.29}$$

is an approximation to the conditional distribution of the type of the next sampled sequence, given the types present in $\mathcal{D}$.

Let $p_m(\alpha)$ be the probability that a chosen sequence α is obtained by mutation from a sequence of some type β, and let $p_c(\alpha)$ be the probability that a chosen sequence α coalesces with another sequence of the same type.

Then the following holds

$$p_m(\alpha) + p_c(\alpha) = \frac{n_i^{(\alpha)}}{n_i},\ i = 1, 2, ... \tag{6.30}$$

Proof.

The main issue is to prove that, indeed, (6.28) defines a probability distribution, i.e, once α is chosen, equation (6.30) holds.

Before proceeding, note that (6.29) implies that (see $S\&D$)

$$\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid \mathcal{D}) = \sum_{\beta \in E} \hat{\pi}_{\delta_{\ell_2}}(\beta \mid \mathcal{D}) \left(\frac{\delta_{\ell_2}\theta}{n + \delta_{\ell_2}\theta} P_{\beta\alpha} + \frac{n^{(\alpha)}}{n + \delta_{\ell_2}\theta} \right). \tag{6.31}$$

Indeed, it can be shown as in $S\&D$ that the distribution $\hat{\pi}_{\delta_{\ell_2}}(. \mid \mathcal{D})$ has the form:

$$\hat{\pi}_{\delta_{\ell_2}}(\beta \mid \mathcal{D}) = \sum_{\alpha \in E} \frac{n^{(\alpha)}}{n} M_{\alpha\beta}^{(n,\delta_{\ell_2})}, \tag{6.32}$$

where

- $M^{(n,\delta_{\ell_2})} = (1 - \gamma_{n,\delta_{\ell_2}})(I - \gamma_{n,\delta_{\ell_2}}\mathbf{P})^{-1}$;
- $\gamma_{n,\delta_{\ell_2}} = \dfrac{\delta_{\ell_2}\theta}{n + \delta_{\ell_2}\theta}$.

Given (6.28), we have

$$p_m(\alpha) = C_{\delta_{\ell_2}}^{-1} \sum_{\beta \in E} \frac{\delta_{\ell_2}\theta}{2} n_i^{(\alpha)} \frac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} P_{\beta\alpha} \tag{6.33}$$

and

$$p_c(\alpha) = C_{\delta_{\ell_2}}^{-1} \binom{n_i^{(\alpha)}}{2} \frac{1}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)}. \tag{6.34}$$

Equation (6.31) gives for the special case $\mathcal{D} = H_i - \alpha$ (and therefore a total of $n_i - 1$ sequences with $n_i^{(\alpha)} - 1$ of type α)

$$\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha) = \sum_{\beta \in E} \hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha) \left(\frac{\delta_{\ell_2}\theta}{n_i + \delta_{\ell_2}\theta - 1} P_{\beta\alpha} + \frac{n_i^{(\alpha)} - 1}{n_i + \delta_{\ell_2}\theta - 1} \right). \tag{6.35}$$

By dividing both sides of equality (6.35) with $\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)$, we have

$$1 = \sum_{\beta \in E} \frac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} \left(\frac{\delta_{\ell_2}\theta}{n_i + \delta_{\ell_2}\theta - 1} P_{\beta\alpha} + \frac{n_i^{(\alpha)} - 1}{n_i + \delta_{\ell_2}\theta - 1} \right). \tag{6.36}$$

Further, by multiplying both sides of equation (6.36) by $n_i^{(\alpha)}/n_i$, we obtain

$$\begin{aligned} \frac{n_i^{(\alpha)}}{n_i} = & \sum_{\beta \in E} \frac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} \left(\frac{n_i^{(\alpha)}\delta_{\ell_2}\theta/2}{n_i(n_i + \delta_{\ell_2}\theta - 1)/2} P_{\beta\alpha} \right) \\ & + \sum_{\beta \in E} \frac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} \left(\frac{n_i^{(\alpha)}(n_i^{(\alpha)} - 1)/2}{n_i(n_i + \delta_{\ell_2}\theta - 1)/2} \cdot \frac{\delta_{\ell_2}}{\delta_{\ell_2}} \right). \end{aligned} \tag{6.37}$$

Let $C_{\delta_{\ell_2}} = \dfrac{n_i(n_i - 1 + \delta_{\ell_2}\theta)}{2\delta_{\ell_2}}$ be the proportionality constant, which depends on the number n_i of sequences present at $D_{i,v}$, and the interval I_c, $c = 1, 2, .., C + 1$ that contains $v' = wt_i + v$; then, one can rewrite (6.37)

$$\frac{n_i^{(\alpha)}}{n_i} = \sum_{\beta \in E} \frac{\hat{\pi}_{\delta_{\ell_2}}(\beta \mid H_i - \alpha)}{\hat{\pi}_{\delta_{\ell_2}}(\alpha \mid H_i - \alpha)} \left(n_i^{(\alpha)} \frac{\theta}{2} C_{\delta_{\ell_2}}^{-1} P_{\beta\alpha} + \binom{n_i^{(\alpha)}}{2} \frac{C_{\delta_{\ell_2}}^{-1}}{\delta_{\ell_2}} \right) \tag{6.38}$$

which gives

$$\frac{n_i^{(\alpha)}}{n_i} = p_m(\alpha) + p_c(\alpha), \tag{6.39}$$

as stated. ■

Remark. Formula (6.29) reflects the fact that, with n_i sequences present, if the next event is at time $v' \in I_{\ell_2}$, the probability that m mutations have occured before one coalescence is given by a geometric distribution of parameter $n_i/(n_i + \delta_{\ell_2})$. This is a consequence of the Poisson nature of the event processes (see, e.g. Ewens, ch. 10).

6.3.4 Algorithm for simulating the genealogies backward in time

In what follows, we describe the proposed algorithm for simulating the genealogies using the distribution $\hat{q}_\delta$ of Proposition 2, with known δ_c on I_c, $c = 1, 2, ..., C+1$:

1. initialize $n_i := n$, where n is the number of sampled sequences at time $t_0 = 0$, and let $i := 0$;

2. let v be the cumulated time until the previous event, and let I_{ℓ_1} the interval where v belongs. We initialize $v := 0$;

3. in the presence of k sequences, simulate the time to the next event, Wt_{-i-1} as described in Section 6.3.2, and let wt_{-i-1} be the time obtained.

4. compute $v' := v + wt_{-i-1}$, and record the interval in $[0; +\infty)$ to which v' belongs; it is denoted I_{ℓ_2};

5. choose randomly a sequence from H_i; the chosen sequence type is denoted α; this sequence is picked with probability $n_i^{(\alpha)}/n_i$;

6. Assuming that $v' \in I_{\ell_2}$, compute the quantities $x_1 = \delta_{\ell_2}\theta \sum_{\beta \in E} \hat{\pi}(\beta \mid H_i - \alpha) P_{\beta\alpha}$ and $x_2 = n_i^{(\alpha)} - 1$.

   Then choose

- a coalescence event with probability $x_2/(x_1 + x_2)$;
- a mutation event with probability $x_1/(x_1 + x_2)$.

7. depending on the result in the previous step, we continue as follows :
   (a) if there is a coalescence event, choose another sequence of type α randomly, let $n_{i-1} := n_i - 1$, $v := v'$, and $H_{i-1} = H_i - \alpha$;
   (b) if there is a mutation event, mutate the sequence α into a sequence β chosen proportionally to the values $\delta_{\ell_2}\theta\hat{\pi}(\beta \mid H_i - \alpha)P_{\beta\alpha}$, without changing n_i, i.e., let $n_{i-1} := n_i$, $v := v'$, and $H_{i-1} = H_i - \alpha + \beta$;
8. $i := i - 1$ and continue until $n_i = 1$.

Remark. One sees that, when $v \in I_{\ell_1}$ and $v' = v+t \in I_{\ell_2}$, choosing the next event is done proportionally to the quantities x_1 and x_2 (see step 6 of the algorithm). So, for example, when $\delta_{\ell_2} < \delta_{\ell_1}$, it is more likely to have less mutation events in the interval I_{ℓ_2} than in interval I_{ℓ_1}, as the branch lengths are shorter. This is reflected in a decrease in the effective population size in I_{ℓ_2}, when time is viewed from the present to the past.

6.3.5 Estimation of the effective population size using the *calibrated skywis plot*

Reconstructing the demographic history from DNA sequences using the *calibrated skywis plot* is similar to the *skywis plot* (Ait Kaci Azzou *et al.*, 2015), but by assuming that the population size is firstly approximated by a constant δ_c per interval I_c, $c = 1, 2, ..., C + 1$, which requires a new way of simulating the genealogies as described in Section 6.3.4. The genealogies are simulated using the importance function given in Section 6.3.3), and the importance sampling weight,

$W^{(j)}$, can be inferred for each genealogy $\mathcal{G}^{(j)}$ as follows:

$$W^{(j)} = \prod_{i=0}^{-\tau^{(j)}+1} \frac{p_\delta^{(j)}(H_i \mid H_{i-1})}{\hat{q}_\delta^{(j)}(H_{i-1} \mid H_i)} = \frac{P(\mathcal{G}^{(j)})}{\widehat{Q}(\mathcal{G}^{(j)})}, \; j = 1, 2, \ldots, J, \tag{6.40}$$

where $\tau^{(j)}$ is the total number of events, and $p_\delta^{(j)}$, $\hat{q}_\delta^{(j)}$ are respectively the forward and backward probabilities as defined in (6.27), (6.28), for genealogy $\mathcal{G}^{(j)}$.

Once the genealogies are simulated, and the importance sampling weights are computed for each genealogy, we proceed as follows:

- we fix the total number of epochs, n_{cum}, i.e. the total number of time intervals where we estimate the effective population size;
- for each simulated genealogy $\mathcal{G}^{(j)}$, we compute the MRCA time, $T_{MRCA}^{(j)}$;
- we use formula (6.41) proposed by Durbin et Li (2011) in order to define epochs where estimates of the effective population size are computed.

  For a genealogy $\mathcal{G}^{(j)}, j = 1, 2, .., J$, the time cutting points are given by:

  $$t_{\text{cut},a}^{(j)} = 0.1 \cdot \exp\left(\frac{a}{n_{\text{cum}}} \cdot \log(1 + 10 \cdot T_{MRCA}^{(j)})\right) - 0.1, \quad a = 1, 2, ..., n_{\text{cum}}, \tag{6.41}$$

  where $t_{\text{cut},n_{\text{cum}}}^{(j)} = T_{MRCA}^{(j)}$.

- For each *epoch* a, $a = 1, ..., n_{\text{cum}}$, we compute the estimate $\widehat{N}e_a$ as the weighted mean of $\widehat{N}e_a^{(j)}$ over the J simulated genealogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(J)}$, where $\widehat{N}e_a^{(j)}$ is computed using equation (6.5).

6.3.6 *Iterative calibrated skywis plot* method

Sometimes, in order to estimate the relative effective population size between two successive sampling times, there is no available information, like the viral load,

e.g.. In such a case, it is not possible to apply the *calibrated skywis plot* method as described in the previous section. Thus, in what follows, we propose a new iterative procedure in which the standard *skywis plot* is applied at the first iteration.

The main steps of this *iterative calibrated skywis plot* are as follows. At iteration "0" one uses the standard *skywis plot* estimate, which gives an estimated time to the Most Recent Common Ancestor, $\hat{T}^{(0)}_{\mathrm{MRCA}}$, and an estimated $\hat{N}_e^{(0)}(\cdot)$. This allows to create a first approximating function, which is piecewise constant; indeed, given the $\hat{T}^{(0)}_{\mathrm{MRCA}}$ one computes the boundaries $b_1^{(1)}, b_2^{(1)}, ..., b_C^{(1)}$ by appropriately dividing the time axis from 0 to the $\hat{T}^{(0)}_{\mathrm{MRCA}}$; further, the estimate $\hat{N}_e^{(0)}(\cdot)$ is used to get the values $\hat{\delta}_c^{(1)}$, $c = 1, 2, ..., C+1$. Next, one can apply the *calibrated skywis plot* at iteration "1" on this piecewise constant function. The procedure continues in a similar way, where at iteration i one applies the *calibrated skywis plot* to an approximating function computed from the values obtained at the previous iteration, $i-1$. In our present implementation of the general method we use the following values.

- The boundaries $b_1^{(i)}, b_2^{(i)}, ..., b_C^{(i)}$ are obtained by dividing the interval $[0, \hat{T}^{(i-1)}_{MRCA})$ according to the formula (Durbin et Li, 2011)

$$b_c^{(i)} = 0.1 \cdot \exp\left(\frac{c}{C} \cdot \log(1 + 10 \cdot \hat{T}^{(i-1)}_{MRCA})\right) - 0.1,\ c = 1, 2, \ldots, C+1, \quad (6.42)$$

  where $b_{C+1}^{(i)} = \hat{T}^{(i-1)}_{MRCA}$.

- The values $\hat{\delta}_c^{(i)}, c = 1, 2, ..., C+1$ are defined as

$$\hat{\delta}_c^{(i)} = \overline{\frac{\hat{N}_e^{(i-1)}(t)}{N} I(t)_{[b_c^{(i)}; b_{c+1}^{(i)})}}, \quad (6.43)$$

  where $\hat{N}_e^{(i-1)}(\cdot)$ is the population size estimate obtained at the previous iteration.

As for the estimates $\hat{T}^{(i)}_{MRCA}$, $i = 0, 1, 2, \ldots$, they are obtained from the J estimates $\hat{T}^{(i,j)}_{MRCA}$, $j = 1, 2, ..., J$ (one for each genealogy $\mathcal{G}^{(j)}$) as

$$\hat{T}^{(i)}_{MRCA} = \sum_{j=1}^{J} w^{(i,j)} \hat{T}^{(i,j)}_{MRCA}, \tag{6.44}$$

where $w^{(i,j)}$ represents the importance weight of a genealogy $\mathcal{G}^{(j)}, j = 1, 2, \ldots, J$, at iteration $i = 0, 1, 2, \ldots$.

We note that the estimate given by Equation (6.44) makes sense because the distribution of the importance weights of genealogies is an approximation of the posterior distribution $P(\mathcal{G} \mid \mathcal{D}, \theta)$ (Stephens, 2001).

Stopping Criterion

The stopping criterion for our iterative method is based on the estimated MRCA time ($\hat{T}_{\text{MRCA}}$). A good estimate $\hat{T}_{\text{MRCA}}$ is very important for providing a good estimate of the effective population size. For example, a short time to the MRCA is indicative of a rapid expansion of the population size (from the past to the present). Thus, we choose to use the following stopping criterion:

$$\frac{|\hat{T}^{(i)}_{\text{MRCA}} - \hat{T}^{(i-1)}_{\text{MRCA}}|}{\hat{T}^{(i-1)}_{\text{MRCA}}} < \epsilon, \tag{6.45}$$

that measures the relative change in the T_{MRCA} between two successive iterations.

Quality of estimation

Importance sampling uses unequally weighted observations, where, in our context, a weight for a genealogy $\mathcal{G}^{(j)}$ is given by

$$W^{(j)} = \frac{P(\mathcal{G}^{(j)})}{Q(\mathcal{G}^{(j)})}, \; j = 1, 2, ..., J. \tag{6.46}$$

The quality of an estimate based on importance sampling depends on how well the proposal distribution $Q(.)$ is in agreement with $P(\cdot)$. The Effective Sample Size (ESS) which is based on the weights $W^{(j)}, j = 1, 2, ..., J$, can be used to provide a quantitative measure of the quality of an estimator (Owen, 2009).

In order to illustrate the meaning of ESS, we consider two extreme cases:

- there is one genealogy $\mathcal{G}^{(j)}$ that has the largest weight which is much bigger than the ones of other genealogies; then, the Importance Sampling technique is equivalent to using only one genealogy, which cannot be desirable especially in our case; indeed, our method is aiming at using several genealogies in order to obtain a relative smooth plot of the change in the effective population size over time;
- the second extreme situation is when the weights $W^{(j)}$, $j = 1, 2, ..., J$ are all equal and necessarily very small. This indicates that the Importance Sampling failed as the weighted estimation reduces to simple averaging.

In the general case, and assuming that $\sum_{j=1}^{J} W^{(j)} > 0$, we use the ESS as a quality measure of an estimation procedure based on importance sampling. The ESS is given by (Owen, 2009):

$$ESS = \frac{\left(\sum_{j=1}^{J} W^{(j)}\right)^2}{\sum_{j=1}^{J} \left(W^{(j)}\right)^2}. \tag{6.47}$$

If the weights $W^{(j)}, j = 1, 2, ..., J$ are too imbalanced, the result is equivalent to averaging only ESS genealogies, where $ESS \ll J$. The point at which ESS becomes very small is hard to specify, and is specific to each application. In our case, we found empirically that we need $ESS \approx 10$ in order to have a relatively smooth plot for estimating the effective population size over time.

6.4 Results

In this section, we use simulated data in order to illustrate the ability of the newly proposed *calibrated skywis plot* to capture the demographic signal contained in the DNA sequences. We consider the case of rapid changes in population size, where the standard *skywis plot* (Ait Kaci Azzou *et al.*, 2015) performs less well.

We provide the results at the following steps:

1. estimation of the effective population size using the standard *skywis plot* (Ait Kaci Azzou *et al.*, 2015);
2. estimating the effective population size using the *calibrated skywis plot* with known $\delta_c, c = 1, ..., C+1$ (see Section 6.3) ;
3. estimating the effective population size with unknown δ_c, $c = 1, ..., C+1$. This is done via an iterative procedure, in which the *skywis plot* approach is used at the first estimation step of δ_c, and the *calibrated skywis plot* for the other iterations.

Our illustration used data produced by the *fastsimcoal* program (Excoffier et Foll, 2011), and we simulated 50 DNA sequences (under the usual assumption of no population structure, and under no recombination). We assumed that the effective population growth is exponential according to the equation

$$N_e(t) = N \exp(-\beta t),\ \beta = 10, \tag{6.48}$$

where t is measured from the present to the past. In other words, from the past to the present the population is expanding. The time units are N generations and from now on these time units are denoted N·g.

We took the following parameters:

- number of nucleotides: 1000;
- $N = N_e(0)$=10000, at time $t = 0$;
- mutation rate per nucleotide : $5 \cdot 10^{-7}$ ($\theta = 10$);
- JC69 finite sites model (Jukes et Cantor, 1969).

The *skywis plot* (see Section 6.2.2) for the DNA sequences simulated from the model described above is given by the blue line in Figure 5(a). Further, the green line represents the *skywis plot* with equal weights for all simulated genealogies (in other words, a non weighted average).[3]

Comparing the blue and the green lines is a way to visualize the performance of our importance sampling scheme. We can note the better performance of the *skywis plot* (blue line) in which only some genealogies contribute significantly to the full likelihood (genealogies that are in better agreement with the data). The blue curve gives a better indication of the rapid decrease of the effective population size from the present to the past, but doesn't follow closely the true exponential curve (red line); in reality, the effective population decreases even faster from the present to the past. Indeed, the estimated time to the Most Recent Common Ancestor ($\hat{T}_{\text{MRCA}}$) is given by $1.16\,\text{N}\cdot\text{g}$, while the true value of the T_{MRCA} is $0.51\,\text{N}\cdot\text{g}$.

Although the *skywis plot* allows us to conclude that the effective population size has been expanding, we consider that this non-parametric estimate can be improved by using the iterative method described in Section 6.3.6.

Further, assume that we know the true values δ_c, $c = 1, 2, ..., C + 1$ which are given by $N(t_c)$, where t_c is the mid-interval value of I_c; the *calibrated skywis plot*

---

[3]For Figure 6.5 (a), we simulate 10 000 genealogies using MATLAB programming language (MATLAB, 2013), with number of epochs $n_{\text{cum}} = 5$.

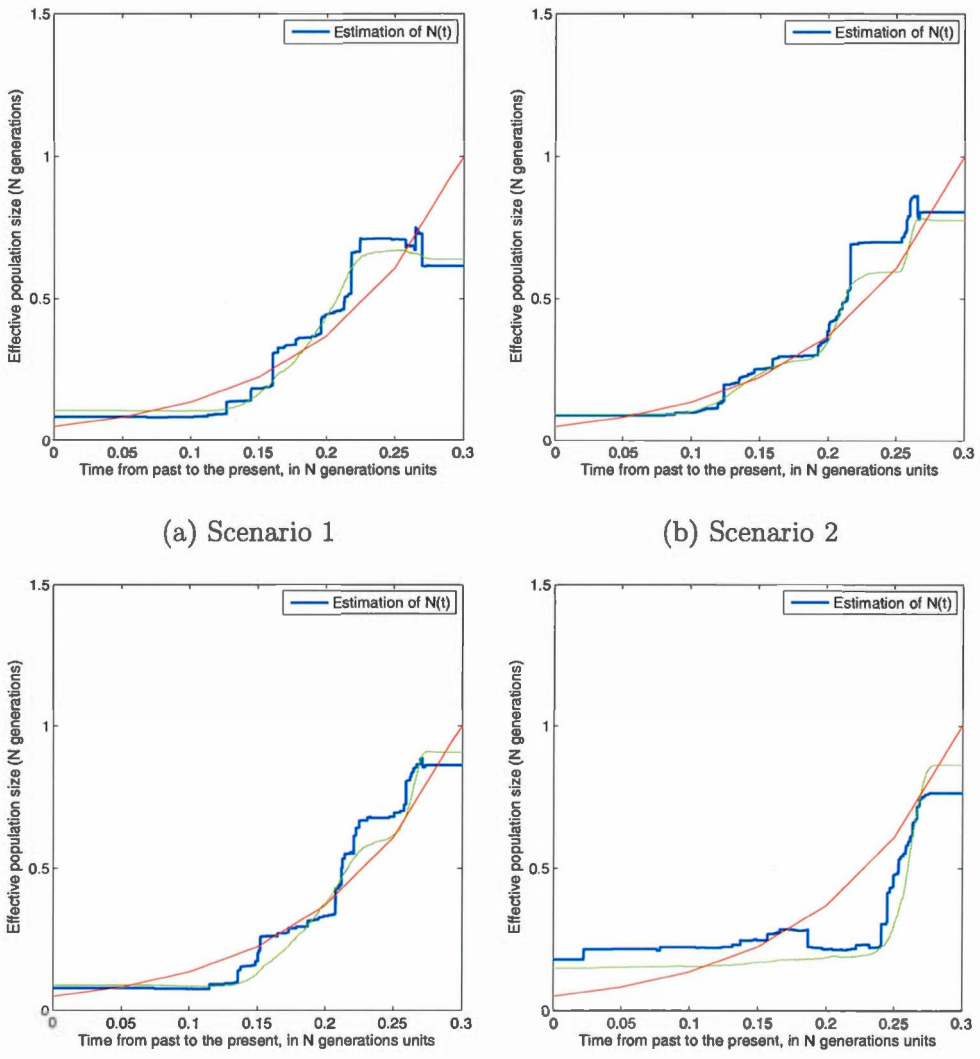
	Cutoff time values, in N·g	δ_c	ESS
Scenario 1	0.1328, 0.4419, 0.8017	0.5880, 0.0882, 0.0035, 0.0001	7
Scenario 2	0.08, 0.2302, 0.3651	0.7233, 0.2445, 0.0590, 0.0149	14
Scenario 3	0.0500 0.1500	0.8661, 0.4453, 0.0683	10
Scenario 4	0.025 0.05	0.9524, 0.7798, 0.1485	6

Table 6.1: Cutoff times, δ_c, and ESS values for 4 scenarios.

is computed for different approximating scenarios which differ in the cutoff time values as well as in the δ_c. These cutoff time values and the effective sample sizes (ESS) are summarized in Table 6.1.

The *calibrated skywis plots* for the scenarios described above are given in Figure 6.4. In all four studied scenarios, we observe, by comparing the blue and the red lines, that the *calibrated skywis plot* gives a good estimate of the effective population size. This means that compressing time by the known values of δ_c substantially improves the method of simulating our genealogies backward in time. Further, we note that the curves are smoother than in the case of the standard *skywis plot* (Ait Kaci Azzou *et al.*, 2015) which means that there are more genealogies that contribute significantly to the calculation of the estimates of the effective population size. This is confirmed by the values of ESS which in this case are greater than the ESS in the *skywis plot*, where ESS=3.

From Table 6.1 and Figure 6.4, we can note that using previous estimates of the relative population size at different points in time is more efficient when these times are quite spread apart. Finally, we can conclude that the present way of simulating the genealogies allows us to choose the genealogies that are more consistent with the data, which reduces substantially the search space, and improves the estimation of the effective population size. All these observations allow us to



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 6.4: *Calibrated skywis plot* for four scenarios and known δ_c, using 2000 simulated genealogies.

c	Interval I_c, in N·g	$\hat{\delta}_c$
1	$[0; 0.1328)$	0.9969
2	$[0.1328; 0.4419)$	0.8713
3	$[0.4419; 0.8017)$	0.7719
4	$[0.8017; 1.16)$	0.3432

Table 6.2: Values of $\hat{\delta}_c$, $c = 1, 2, 3, 4$ computed from iteration (0)

conclude that the importance function $Q(\cdot)$ introduced in Proposition 2 is good and adapted to the present data.

In what follows, we apply the *iterative calibrated skywis plot* methodology described in Section 6.3.6 in order to capture the demographic signal contained in DNA sequences without adding information like the one used to produce the plots in Figure 6.4.

In our simulation, we considered four intervals I_c, $c = 1, 2, 3, 4$. At the first iteration we applied the *calibrated skywis plot*. At the next iterations, $i = 1, 2, \ldots$ we determined the boundaries of the intervals I_c, $c = 1, 2, 3, 4$ as follows: we used the formula (6.42) to fix the bounds $b_1^{(i)}, b_2^{(i)}$ measured from the present to the past (in N·g units). The third cut off value $b_3^{(i)}$ was taken as the point which cuts the interval $[b_2^{(i)}; \hat{T}_{\text{MRCA}}^{(i-1)})$ in half.

For example, the time to the Most Recent Common Ancestor, $\hat{T}_{\text{MRCA}}^{(0)}$ for the iteration "0", which uses the standard *skywis plot* (Ait Kaci Azzou *et al.*, 2015), is estimated by $1.16\,\text{N·g}$. Further, using equation (6.42) we get the following cut off values of the cumulated time measured from the present to the past: $b_1^{(1)} = 0.1328\,\text{N·g}$, $b_2^{(1)} = 0.4419\,\text{N·g}$, and $b_3^{(1)} = 0.8017\,\text{N·g}$.

At iteration "1" the *calibrated skywis plot* is computed using the parameters given

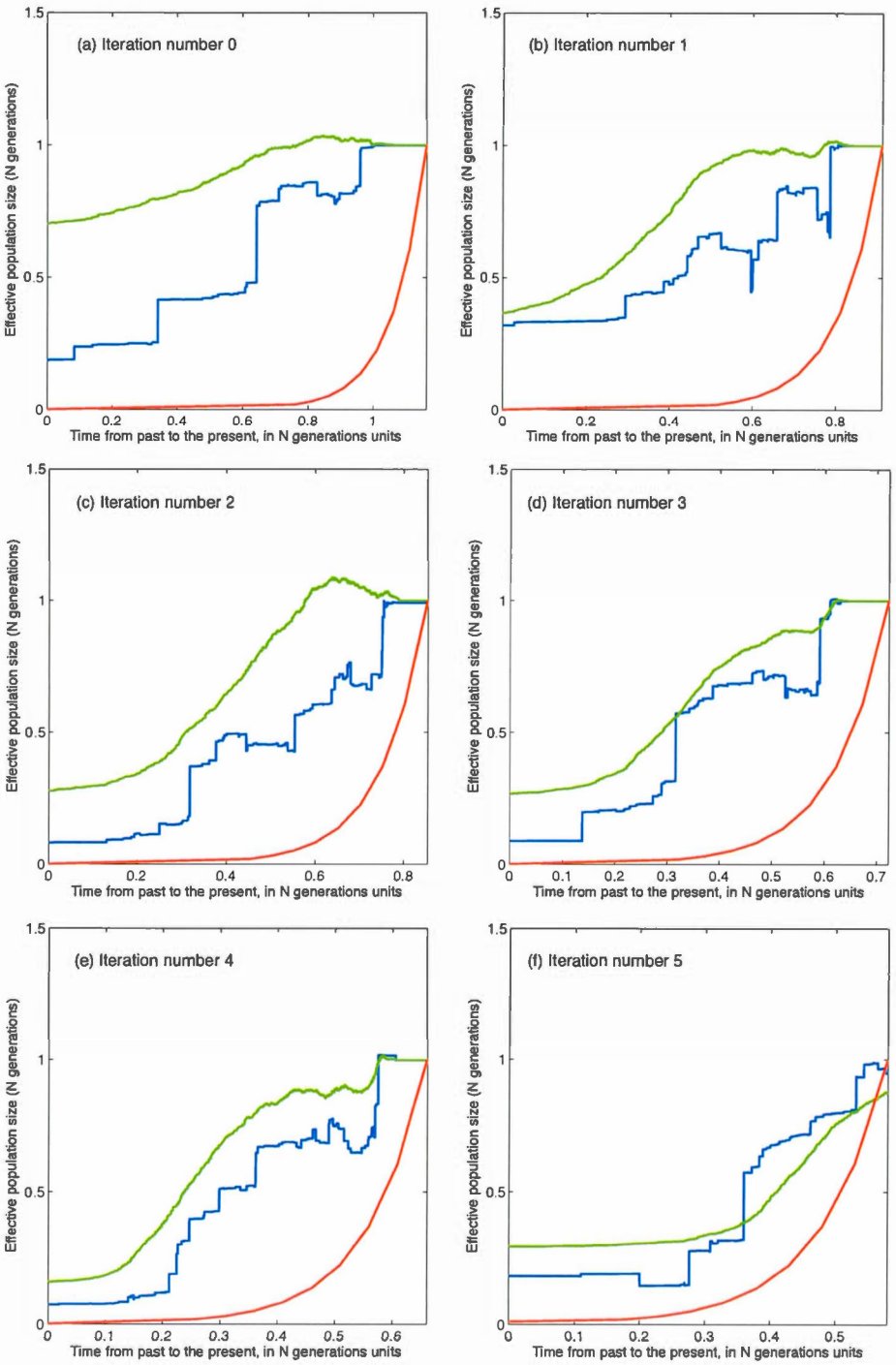


Figure 6.5: *Iterative calibrated skywis plot* with estimated δ_c (blue line); true $N(t)$ (red line), *calibrated skywis plot* with unweighted mean over genealogies (green line).

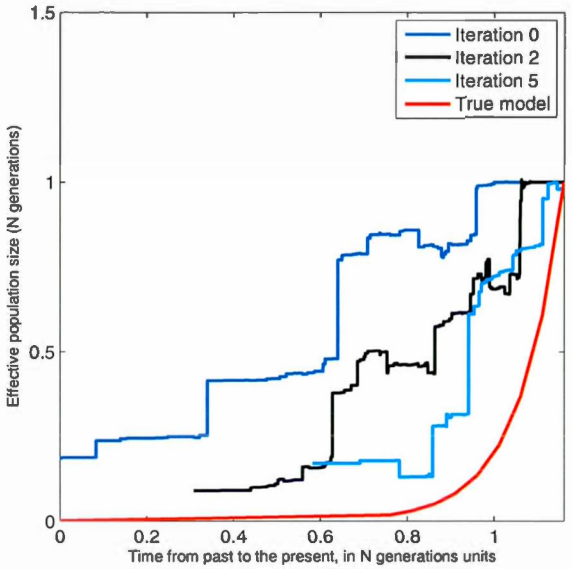


Figure 6.6: Evolution of the *iterative calibrated skywis plot*: three iterations out of six

in Table 6.2, that are based on $\hat{T}^{(0)}_{\text{MRCA}}$ and $\overline{\hat{N}^{(0)}_e(t)}$. Thus, at iteration "1", the times to coalescence events will be compressed by a factor of $\hat{\delta}_c$ for an interval I_c, $c = 1, 2, 3, 4$. For example, when simulating the genealogies from the present to the past, if the cumulated time is greater than or equal to 0.8017 N·g , the coalescence rate in presence of k sequences is $\binom{k}{2}/\hat{\delta}_4$. This statement is equivalent to saying that, in interval I_4 the time t to a coalescence equals the time t^* one would have gotten in the case of a constant population size multiplied by $\hat{\delta}_4 = 0.3432$ (i.e. $t = t^* \cdot 0.3432$).

This procedure is repeated until finding a stable estimate of the time to the MRCA, as defined by the criterion given in Equation (6.45).

Since our stopping criterion is based on the estimate $\hat{T}_{\text{MRCA}}$, Figure (6.7) gives the evolution of this estimate across iterations. Thus, we can observe that the esti-

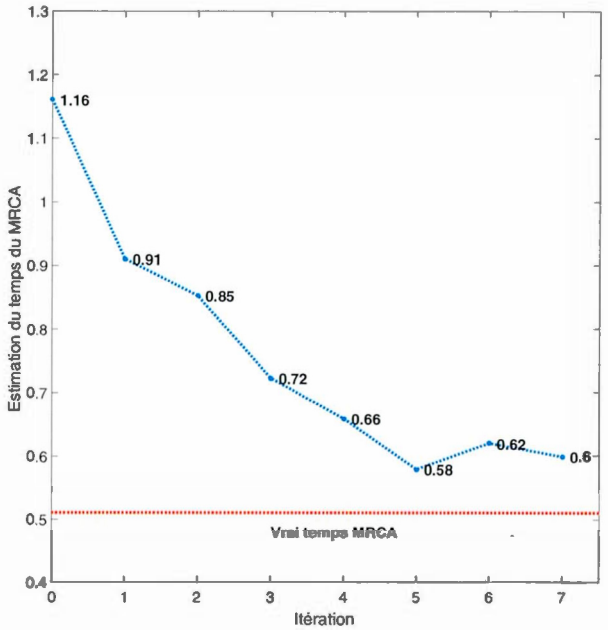


Figure 6.7: Value of the $\hat{T}_{\mathrm{MRCA}}$ for each iteration.

mate $\hat{T}_{\mathrm{MRCA}}$ is stable from iteration 5 on, where the estimated time turns around $0.6\,\mathrm{N{\cdot}g}$, which substantially improves the result given by the standard *skywis plot* where $\hat{T}_{\mathrm{MRCA}}$ is $1.16\,\mathrm{N{\cdot}g}$ (the true value is $0.5\,\mathrm{N{\cdot}g}$). Finally, in Figure 6.6 we present, for comparison, the estimates at iterations 0, 2 and 5 in one single plot. One can see that the *iterative calibrated skywis plot* improves the estimate of the effective population size, as in iteration 5 the estimated curve is close to the true value (red curve). We note that, in this example, ESS=10 at the final iteration.

6.5 Discussion

In this paper we generalize the *skywis plot* methodology introduced in Ait Kaci Azzou *et al.* (2015). The aim is to estimate the evolution over time of the effective population size from a sample of DNA sequences. Our estimates are based on the same importance sampling (IS) methodology as the *skywis plot*. Namely, starting

with a sample of DNA sequences, we generate genealogies backwards in time and the effective population size is first estimated for each *epoch* and each genealogy; further, the effective population size estimate is obtained as a weighted average of these separate estimates using the IS weights.

In order to compute our estimates we approximate the effective population size by a piecewise constant function, which is either (i) assumed known or previously estimated (*calibrated skywis* method) or (ii) estimated (*iterative calibrated skywis* method). The first case comes to assuming that mean population values δ_c are known for $C+1$ intervals I_c, while in the second case, we leave δ_c unknown, but at each iteration step we reduce ourselves to the known case by using a previous estimate.

Our *calibrated skywis plot* requires to adapt the importance function $Q(\cdot)$ of Stephens et Donnelly (2000) on one hand, and on the other hand to simulate the next event time in a novel way that takes into account the fact that the coalescent process is no longer homogeneous in time. How efficient is this IS strategy in the case of a known piecewise constant function can be observed in the resulting *calibrated skywis plots* (Figure 6.4) and in the ESS values (Table 6.1). It appears that the strategy of compressing (or stretching) times on each interval I_c by the approximating factor δ_c works quite well.

In the *iterative calibrated skywis plot* method, at each iteration i we compute a *calibrated skywis plot* estimate by resorting to the estimated values $\hat{N}_e^{(i-1)}(\cdot)$ obtained at the previous iteration. We illustrate the *iterative calibrated skywis plot* methodology on a simulated data set from a rapidly expanding exponential model. It appears that after performing a few iterations, the estimate of the effective population size is improved substantially. In particular, one gets a good estimate of the time to the most recent ancestor, which is a natural indicator of a good

estimation procedure in the demographic history context; therefore, we expect our method to be extended to other models.

Acknowledgements. This research received financial support from NSERC (Natural Sciences and Engineering Research Council of Canada).

6.6 Appendices

Appendix A : Simulation of the time to the next event presented in tabular form: piecewise constant function on C=4 intervals.

Table 6.3: Formulas for the simulation of the time to the next event, where the effective population size is piecewise constant ($C = 4$).

	$v + t_k < b_1$	$b_1 \leq v + t_k < b_2$	$b_2 \leq v + t_k < b_3$
$v < b_1$	$\frac{-\ln(1-u)}{C_{\delta_1}}$	$\frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}} - \frac{\ln(1-u)}{C_{\delta_2}}$	$-\frac{\ln(1-u)}{C_{\delta_3}} - \frac{(b_1-v)\binom{k}{2}}{\delta_1 C_{\delta_3}} - \frac{(b_2-b_1)\binom{k}{2}}{\delta_2 C_{\delta_3}} + \frac{(b_2-v)\binom{k}{2}}{\delta_3 C_{\delta_3}}$
$b_1 \leq v < b_2$	-	$\frac{\ln(1-u)}{C_{\delta_2}}$	$-\frac{\ln(1-u)}{C_{\delta_3}} + \frac{(b_2-v)\binom{k}{2}\left(\frac{1}{\delta_3}-\frac{1}{\delta_2}\right)}{C_{\delta_3}}$
$b_2 \leq v < b_3$	-	-	$-\frac{\ln(1-u)}{C_{\delta_3}}$
$v \geq b_3$	-	-	-

Table 6.4: Domains for the simulation of the next event time where the population size is piecewise constant ($C = 3$).

	$v + t_k < b_1$	$b_1 \leq v + t_k < b_2$	$b_2 \leq v + t_k < b_3$	$v + t_k \geq b_3$
$v < b_1$	$\frac{-\ln(1-u)}{C_{\delta_1}} < (b_1 - v)$	$(b_1 - v) - \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}} \leq -\frac{\ln(1-u)}{C_{\delta_2}} < (b_2 - v) - \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}}$	$\{-\frac{\ln(1-u)}{C_{\delta_3}} \geq (b_2 - v) + \frac{(b_1-v)\binom{k}{2}}{\delta_1 C_{\delta_3}} + \frac{(b_2-b_1)\binom{k}{2}}{\delta_2 C_{\delta_3}} - \frac{(b_2-v)\binom{k}{2}}{\delta_3 C_{\delta_3}}\}$ and $\{-\frac{\ln(1-u)}{C_{\delta_3}} \leq (b_3 - v) + \frac{(b_1-v)\binom{k}{2}}{\delta_1 C_{\delta_3}} + \frac{(b_2-b_1)\binom{k}{2}}{\delta_2 C_{\delta_3}} - \frac{(b_2-v)\binom{k}{2}}{\delta_3 C_{\delta_3}}\}$	$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) + \frac{(b_1-v)\binom{k}{2}}{\delta_1 C_{\delta_4}} + \frac{(b_2-b_1)\binom{k}{2}}{\delta_2 C_{\delta_4}} + \frac{(b_3-b_2)\binom{k}{2}}{\delta_3 C_{\delta_4}} - \frac{(b_3-v)\binom{k}{2}}{\delta_4 C_{\delta_4}}$
$b_1 \leq v < b_2$	-	$(b_1 - v) \leq -\frac{\ln(1-u)}{C_{\delta_2}} < (b_2 - v)$	$(b_2 - v) - \frac{(b_2-v)\binom{k}{2}\left(\frac{1}{\delta_3}-\frac{1}{\delta_2}\right)}{C_{\delta_3}} \leq -\frac{\ln(1-u)}{C_{\delta_3}} < (b_3 - v) - \frac{(b_2-v)\binom{k}{2}\left(\frac{1}{\delta_3}-\frac{1}{\delta_2}\right)}{C_{\delta_3}}$	$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) + \frac{(b_2-v)\binom{k}{2}}{\delta_2 C_{\delta_4}} + \frac{(b_3-b_2)\binom{k}{2}}{\delta_3 C_{\delta_4}} - \frac{(b_3-v)\binom{k}{2}}{\delta_4 C_{\delta_4}}$
$b_2 \leq v < b_3$	-	-	$(b_2 - v) \leq -\frac{\ln(1-u)}{C_{\delta_3}} < (b_3 - v)$	$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) - \frac{(b_3-v)\binom{k}{2}\left(\frac{1}{\delta_4}-\frac{1}{\delta_3}\right)}{C_{\delta_4}}$.
$v \geq b_3$	-	-	-	Valid at all times

Appendix B : Simulation method of the next event time where the population size is approximated by a piecewise constant function on $I_c, c = 1, 2, ...C + 1$: the general formulas.

Corollary 1.

Let v be the observed cumulated time in the $M_{Tot,k}(t)$ process defined in Section 6.3.1, and let t be a realization of the time to the next event.

Assume we use the approximation

$$\frac{\widetilde{N_e}(t)}{N} = \delta_c I_{\{b_{c-1} \leq t < b_c\}},\ c = 1, 2, ..., C + 1,$$

where $b_{C+1} = T_{MRCA}$.

Let $v \in I_\ell$, and $v + t \in I_j$, $\ell, j = 1, 2, ..., C + 1$, $j \geq \ell$. Then, all possible cases of solutions to equation (6.14) are summarized below :

1. *$\ell = 1$, and $\frac{-\ln(1-u)}{C_{\delta_1}} < (b_1 - v)$, then $j = 1$ and*

$$t = -\frac{\ln(1-u)}{C_{\delta_1}}.$$

2. *$\ell = C + 1$, we have necessarily $j = C + 1$ and*

$$t = -\frac{\ln(1-u)}{C_{\delta_{C+1}}}.$$

3. *ℓ such as $\ell \neq 1$, and $\ell \neq C + 1$, and $(b_{\ell-1} - v) \leq -\frac{\ln(1-u)}{C_{\delta_\ell}} < (b_\ell - v)$, then $j = \ell$, and*

$$t = -\frac{\ln(1-u)}{C_{\delta_j}}.$$

4. *ℓ such as $\ell \neq C$, and*

$$(b_\ell - v) - \frac{(b_\ell - v)\binom{k}{2}\left(\frac{1}{\delta_{\ell+1}} - \frac{1}{\delta_\ell}\right)}{C_{\delta_{\ell+1}}} \leq -\frac{\ln(1-u)}{C_{\delta_{\ell+1}}} < (b_{\ell+1} - v) - \frac{(b_\ell - v)\binom{k}{2}\left(\frac{1}{\delta_{\ell+1}} - \frac{1}{\delta_\ell}\right)}{C_{\delta_{\ell+1}}},$$

then $j = \ell + 1$ and :

$$t = \frac{(b_{j-1} - v)\binom{k}{2}}{C_{\delta_j}}\left(\frac{1}{\delta_j} - \frac{1}{\delta_{j-1}}\right) - \frac{\ln(1-u)}{C_{\delta_j}}.$$

5. *$\ell \neq C + 1$ and*

$$-\frac{\ln(1-u)}{C_{\delta_{C+1}}} \geq (b_C - v) + \binom{k}{2}\left(\frac{(b_\ell - v)}{\delta_\ell C_{\delta_{C+1}}} + \sum_{i=\ell+1}^{C} \frac{(b_i - b_{i-1})}{\delta_i C_{\delta_{C+1}}} - \frac{(b_C - v)}{\delta_{C+1} C_{\delta_{C+1}}}\right),$$

then $j = C + 1$, and

$$t = -\frac{\ln(1-u)}{C_{\delta_{C+1}}} - \binom{k}{2}\left\{\frac{(b_\ell - v)}{\delta_\ell C_{\delta_{C+1}}} - \sum_{i=\ell+1}^{C} \frac{(b_i - b_{i-1})}{\delta_i C_{\delta_{C+1}}} + \frac{(b_C - v)}{\delta_{C+1} C_{\delta_{C+1}}}\right\}$$

6. *If $v \in I_\ell$, then $v + t \in I_j = [b_{j-1}; b_j)$ if and only if*

$$\left\{-\frac{\ln(1-u)}{C_{\delta_j}} \geq (b_{j-1} - v) + \binom{k}{2}\left(\frac{(b_\ell - v)}{\delta_\ell C_{\delta_j}} + \sum_{i=\ell+1}^{j-1} \frac{(b_i - b_{i-1})}{\delta_i C_{\delta_j}} - \frac{(b_{j-1} - v)}{\delta_j C_{\delta_j}}\right)\right\}$$

and

$$\left\{-\frac{\ln(1-u)}{C_{\delta_j}} \leq (b_j - v) + \binom{k}{2}\left(\frac{(b_\ell - v)}{\delta_\ell C_{\delta_j}} + \sum_{i=\ell+1}^{j-1} \frac{(b_i - b_{i-1})}{\delta_i C_{\delta_j}} - \frac{(b_{j-1} - v)}{\delta_j C_{\delta_j}}\right)\right\}.$$

We have, in this case

$$t = -\frac{\ln(1-u)}{C_{\delta_j}} - \binom{k}{2}\left(\frac{(b_\ell - v)}{\delta_\ell C_{\delta_j}} - \sum_{i=\ell+1}^{j-1} \frac{(b_i - b_{i-1})}{\delta_i C_{\delta_j}} + \frac{(b_{j-1} - v)}{\delta_j C_{\delta_j}}\right).$$

■

Appendix C: Example of simulation of the next event time where the population size is approximated by a piecewise constant population size, for intervals $I_c, c = 1, 2, 3, 4$

In what follows, we develop the formulas (6.14) for the case $C = 3$.

a) Equation (6.14) for t, where $v < b_1$ and $v + t < b_1$.

In this case $\lambda_{\text{coa}}(t) = \binom{k}{2}\frac{1}{\delta_1}$. Thus, equation (6.14) is equivalent to solving:

$$\begin{aligned}\left(\int_v^{v+t}\left(\binom{k}{2}\frac{1}{\delta_1}\right)du + \frac{k\theta}{2}t\right) = -\ln(1-u) \\ \Leftrightarrow t\left(\binom{k}{2}\frac{1}{\delta_1} + \frac{k\theta}{2}t\right) = -\ln(1-u) \\ \Leftrightarrow t = \frac{-\ln(1-u)}{\binom{k}{2}\frac{1}{\delta_1} + \frac{k\theta}{2}} \\ \Leftrightarrow t = \frac{-\ln(1-u)}{C_{\delta_1}},\end{aligned} \tag{6.49}$$

where

$$C_{\delta_c} = \binom{k}{2}\frac{1}{\delta_c} + \frac{k\theta}{2}.$$

Domain of u, if $v < b_1$ and $v + t < b_1$.

We have, by relacing t with its value in (6.49)

$$v + t < b_1 \Leftrightarrow \frac{-\ln(1-u)}{C_{\delta_1}} < (b_1 - v).$$

b) Equation (6.14) for t, where $v < b_1$ and $v + t \in [b_1; b_2)$.

If $v < b_1$ and $v + t \in [b_1; b_2)$, then

$$\begin{aligned}\int_v^{v+t}\left(\binom{k}{2}\frac{1}{\nu(u)}\right)du &= \int_v^{b_1}\left(\binom{k}{2}\frac{1}{\delta_1}\right)du + \int_{b_1}^{v+t}\left(\binom{k}{2}\frac{1}{\delta_2}\right)du \\ &= \binom{k}{2}\frac{1}{\delta_1}(b_1-v) + (t+v-b_1)\binom{k}{2}\frac{1}{\delta_2}.\end{aligned} \tag{6.50}$$

In this case, the simulation of the time to the next event is equivalent to solving the equation :

$$\binom{k}{2}\frac{1}{\delta_1}(b_1-v) + (t+v-b_1)\binom{k}{2}\frac{1}{\delta_2} + \frac{k\theta}{2}t = -\ln(1-u),$$

which gives :

$$\begin{aligned}t &= \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{\binom{k}{2}\frac{1}{\delta_2}+\frac{k\theta}{2}} - \frac{\ln(1-u)}{\binom{k}{2}\frac{1}{\delta_2}+\frac{k\theta}{2}}. \\ &= \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}} - \frac{\ln(1-u)}{C_{\delta_2}}.\end{aligned} \tag{6.51}$$

Domain of u, if $v < b_1$ and $v + t \in [b_1; b_2)$.

By replacing t by its value in (6.51), we have, since $b_1 \leq v + t < b_2$

$$(b_1-v) \leq \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}} - \frac{\ln(1-u)}{C_{\delta_2}} < (b_2-v).$$

Therefore, u satisfies

$$b_1 - v - \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}} \leq -\frac{\ln(1-u)}{C_{\delta_2}} < b_2 - v - \frac{(b_1-v)\binom{k}{2}\left(\frac{1}{\delta_2}-\frac{1}{\delta_1}\right)}{C_{\delta_2}}.$$

c) Equation (6.14) for t, where $v < b_1$ and $v + t \in [b_2; b_3)$.

If $v < b_1$ and $v + t \in [b_2; b_3)$, then

$$\int_v^{v+t}\left(\binom{k}{2}\frac{1}{\nu(u)}\right)du=\int_v^{b_1}\left(\binom{k}{2}\frac{1}{\delta_1}\right)du+\int_{b_1}^{b_2}\left(\binom{k}{2}\frac{1}{\delta_2}\right)du+\int_{b_2}^{v+t}\left(\binom{k}{2}\frac{1}{\delta_3}\right)du$$
$$=\binom{k}{2}\frac{1}{\delta_1}(b_1-v)+\binom{k}{2}\frac{1}{\delta_2}(b_2-b_1)+(t+v-b_2)\binom{k}{2}\frac{1}{\delta_3}.$$

Solving the equation (6.14) in this case is equivalent to solving

$$\binom{k}{2}\frac{1}{\delta_1}(b_1-v)+\binom{k}{2}\frac{1}{\delta_2}(b_2-b_1)+(t+v-b_2)\binom{k}{2}\frac{1}{\delta_3}+\frac{k\theta}{2}t=-\ln(1-u),$$

which gives

$$t=-\frac{\ln(1-u)}{C_{\delta_3}}-\frac{(b_1-v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_3}}-\frac{(b_2-b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_3}}+\frac{(b_2-v)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_3}}. \qquad (6.52)$$

Domain of u, if $v<b_1$ and $v+t\in[b_2;b_3)$

Given (6.52), $b_2\le v+t<b_3$ is equivalent to

$$(b_2-v)\le-\frac{\ln(1-u)}{C_{\delta_3}}-\frac{(b_1-v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_3}}-\frac{(b_2-b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_3}}+\frac{(b_2-v)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_3}}<(b_3-v),$$

which gives two inequalities for u:

$$-\frac{\ln(1-u)}{C_{\delta_3}}\ge(b_2-v)+\frac{(b_1-v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_3}}+\frac{(b_2-b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_3}}-\frac{(b_2-v)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_3}}$$

and

$$-\frac{\ln(1-u)}{C_{\delta_3}}\le(b_3-v)+\frac{(b_1-v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_3}}+\frac{(b_2-b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_3}}-\frac{(b_2-v)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_3}}$$

d) Equation (6.14) for t, where $v<b_1$ and $v\ge b_3$.

In this case,

$$\int_v^{v+t} \left(\binom{k}{2} \frac{1}{\nu(u)} \right) du,$$

equals

$$\int_v^{b_1} \left(\binom{k}{2} \frac{1}{\delta_1} \right) du + \int_{b_1}^{b_2} \left(\binom{k}{2} \frac{1}{\delta_2} \right) du + \int_{b_2}^{b_3} \left(\binom{k}{2} \frac{1}{\delta_3} \right) du + \int_{b_3}^{v+t} \left(\binom{k}{2} \frac{1}{\delta_4} \right) du$$
$$= \binom{k}{2} \frac{1}{\delta_1}(b_1 - v) + \binom{k}{2} \frac{1}{\delta_2}(b_2 - b_1) + \binom{k}{2} \frac{1}{\delta_3}(b_3 - b_2) + (t + v - b_3) \binom{k}{2} \frac{1}{\delta_4}.$$

Solving equation (6.14) in this case is equivalent to solving

$$\binom{k}{2} \frac{1}{\delta_1}(b_1 - v) + \binom{k}{2} \frac{1}{\delta_2}(b_2 - b_1) + \binom{k}{2} \frac{1}{\delta_3}(b_3 - b_2) + (t + v - b_3) \binom{k}{2} \frac{1}{\delta_4} + \frac{k\theta}{2} t = -\ln(1-u),$$

which gives :

$$t = -\frac{\ln(1-u)}{C_{\delta_4}} - \frac{(b_1 - v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_4}} - \frac{(b_2 - b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_4}} - \frac{(b_3 - b_2)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_4}} + \frac{(b_3 - v)\binom{k}{2}\frac{1}{\delta_4}}{C_{\delta_4}}.$$

Domain of u, if $v < b_1$ and $v \geq b_3$.

In this case, $v + t \geq b_3$ is equivalent to :

$$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) + \frac{(b_1 - v)\binom{k}{2}\frac{1}{\delta_1}}{C_{\delta_4}} + \frac{(b_2 - b_1)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_4}} + \frac{(b_3 - b_2)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_4}} - \frac{(b_3 - v)\binom{k}{2}\frac{1}{\delta_4}}{C_{\delta_4}}.$$

e) Equation (6.14) for t, where $b_1 \leq v < b_2$ and $b_1 \leq v + t < b_2$

We have,

$$\int_v^{v+t} \left(\binom{k}{2} \frac{1}{\nu(u)} \right) du = \int_v^{v+t} \left(\binom{k}{2} \frac{1}{\delta_2} \right) du = \binom{k}{2} \frac{1}{\delta_2} t.$$

Thus, solving (6.14) in this case is equivalent to solving:

$$\binom{k}{2}\frac{1}{\delta_2}t + \frac{k\theta}{2}t = -\ln(1-u),$$

and,

$$t = -\frac{\ln(1-u)}{C_{\delta_2}}. \tag{6.53}$$

Domain of u, if $b_1 \leq v < b_2$ and $b_1 \leq v + t < b_2$.

Given (6.53), $b_1 \leq v + t < b_2$ is equivalent to :

$$(b_1 - v) \leq -\frac{\ln(1-u)}{C_{\delta_2}} < (b_2 - v).$$

f) Equation (6.14) for t, where $[b_1; b_2)$ and $v + t \in [b_2; b_3)$.

We have, in this case :

$$\begin{aligned}\int_v^{v+t} \left(\binom{k}{2}\frac{1}{\nu(u)}\right) du &= \int_v^{b_2} \left(\binom{k}{2}\frac{1}{\delta_2}\right) du + \int_{b_2}^{v+t} \left(\binom{k}{2}\frac{1}{\delta_3}\right) du \\ &= \binom{k}{2}\frac{1}{\delta_2}(b_2 - v) + (t + v - b_2)\binom{k}{2}\frac{1}{\delta_3}.\end{aligned}$$

Solving equation (6.14) in this case is equivalent to solving:

$$\binom{k}{2}\frac{1}{\delta_2}(b_2 - v) + (t + v - b_2)\binom{k}{2}\frac{1}{\delta_3} + \frac{k\theta}{2}t = -\ln(1-u),$$

which gives

$$t = -\frac{\ln(1-u)}{C_{\delta_3}} + \frac{(b_2 - v)\binom{k}{2}\left(\frac{1}{\delta_3} - \frac{1}{\delta_2}\right)}{C_{\delta_3}}. \tag{6.54}$$

Domain of u, if $[b_1; b_2)$ and $v + t \in [b_2; b_3)$.

Given (6.54), $b_2 \leq v + t < b_3$ is equivalent to :

$$(b_2 - v) - \frac{(b_2 - v)\binom{k}{2}\left(\frac{1}{\delta_3} - \frac{1}{\delta_2}\right)}{C_{\delta_3}} \leq -\frac{\ln(1-u)}{C_{\delta_3}} < (b_3 - v) - \frac{(b_2 - v)\binom{k}{2}\left(\frac{1}{\delta_3} - \frac{1}{\delta_2}\right)}{C_{\delta_3}}.$$

g) Equation (6.14) for t, where $v \in [b_1; b_2)$ and $v + t \geq b_3$.

We have

$$\begin{aligned}\int_v^{v+t} \left(\binom{k}{2}\frac{1}{\nu(u)}\right) du &= \int_v^{b_2} \left(\binom{k}{2}\frac{1}{\delta_2}\right) du + \int_{b_2}^{b_3} \left(\binom{k}{2}\frac{1}{\delta_3}\right) du + \int_{b_3}^{v+t} \left(\binom{k}{2}\frac{1}{\delta_4}\right) du \\ &= \binom{k}{2}\frac{1}{\delta_2}(b_2 - v) + \binom{k}{2}\frac{1}{\delta_3}(b_3 - b_2) + (t + v - b_3)\binom{k}{2}\frac{1}{\delta_4},\end{aligned}$$

and solving equation (6.14) in this case is equivalent to solving:

$$\binom{k}{2}\frac{1}{\delta_2}(b_2 - v) + \binom{k}{2}\frac{1}{\delta_3}(b_3 - b_2) + (t + v - b_3)\binom{k}{2}\frac{1}{\delta_4} + \frac{k\theta}{2}t = -\ln(1-u),$$

which gives

$$t = -\frac{\ln(1-u)}{C_{\delta_4}} - \frac{(b_2 - v)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_4}} - \frac{(b_3 - b_2)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_4}} + \frac{(b_3 - v)\binom{k}{2}\frac{1}{\delta_4}}{C_{\delta_4}}. \qquad (6.55)$$

Domain of u, if $v \in [b_1; b_2)$ and $v + t \geq b_3$.

Given (6.55), $v + t \geq b_3$ is equivalent to

$$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) + \frac{(b_2 - v)\binom{k}{2}\frac{1}{\delta_2}}{C_{\delta_4}} + \frac{(b_3 - b_2)\binom{k}{2}\frac{1}{\delta_3}}{C_{\delta_4}} - \frac{(b_3 - v)\binom{k}{2}\frac{1}{\delta_4}}{C_{\delta_4}}.$$

h) Equation (6.14) for t, where $b_2 \leq v < b_3$ and $b_2 \leq v + t < b_3$.

We have, in this case,

$$\int_v^{v+t} \left(\binom{k}{2} \frac{1}{\nu(u)} \right) du = \int_v^{v+t} \left(\binom{k}{2} \frac{1}{\delta_3} \right) du = \binom{k}{2} \frac{1}{\delta_3} t.$$

Solving equation (6.14) in this case is equivalent to solving

$$\binom{k}{2} \frac{1}{\delta_3} t + \frac{k\theta}{2} t = -\ln(1-u),$$

which gives

$$t = -\frac{\ln(1-u)}{C_{\delta_3}}. \tag{6.56}$$

Domain of u, if $b_2 \leq v < b_3$ and $b_2 \leq v + t < b_3$.

Given (6.56), $b_2 \leq v + t < b_3$ is equivalent to :

$$(b_2 - v) \leq -\frac{\ln(1-u)}{C_{\delta_3}} < (b_3 - v)$$

i) Equation (6.14) for t, where $[b_2; b_3)$ and $v + t \geq b_3$.

We have

$$\begin{aligned} \int_v^{v+t} \left(\binom{k}{2} \frac{1}{\nu(u)} \right) du &= \int_v^{b_3} \left(\binom{k}{2} \frac{1}{\delta_3} \right) du + \int_{b_3}^{v+t} \left(\binom{k}{2} \frac{1}{\delta_4} \right) du \\ &= \binom{k}{2} \frac{1}{\delta_3} (b_3 - v) + (t + v - b_3) \binom{k}{2} \frac{1}{\delta_4}. \end{aligned}$$

Solving equation (6.14) in this case is equivalent to solving:

$$\binom{k}{2} \frac{1}{\delta_3} (b_3 - v) + (t + v - b_3) \binom{k}{2} \frac{1}{\delta_4} + \frac{k\theta}{2} t = -\ln(1-u),$$

which gives

$$t = -\frac{\ln(1-u)}{C_{\delta_4}} + \frac{(b_3 - v)\binom{k}{2}\left(\frac{1}{\delta_4} - \frac{1}{\delta_3}\right)}{C_{\delta_4}}. \tag{6.57}$$

Domain of u, if $[b_2; b_3)$ and $v + t \geq b_3$.

Given (6.57), $v + t \geq b_3$ is equivalent to

$$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v) - \frac{(b_3 - v)\binom{k}{2}\left(\frac{1}{\delta_4} - \frac{1}{\delta_3}\right)}{C_{\delta_4}}.$$

j) Equation (6.14) for t, where $v \geq b_3$ and $v + t \geq b_3$.

We have

$$\int_v^{v+t} \left(\binom{k}{2}\frac{1}{\nu(u)}\right) du = \int_v^{v+t} \left(\binom{k}{2}\frac{1}{\delta_4}\right) du = \binom{k}{2}\frac{1}{\delta_4}t$$

Solving equation (6.14) in this case is equivalent to solving:

$$\binom{k}{2}\frac{1}{\delta_4}t + \frac{k\theta}{2}t = -\ln(1-u),$$

which gives

$$t = -\frac{\ln(1-u)}{C_{\delta_4}}. \tag{6.58}$$

Domain of u, if $v \geq b_3$ and $v + t \geq b_3$.

Finally, given (6.58), $v + t \geq b_3$ comes to :

$$-\frac{\ln(1-u)}{C_{\delta_4}} \geq (b_3 - v)$$

CONCLUSION

En nous basant sur la théorie de coalescence, nous avons proposé une nouvelle méthode non paramétrique, le *skywis plot*, qui permet d'explorer l'historique démographique d'un échantillon de séquences d'ADN. La méthode du *skywis plot* est basée sur la simulation d'un grand nombre de généalogies en utilisant un échantillonnage pondéré, où les poids résultants sont utilisés pour le calcul d'une moyenne pondérée des tailles de population effective par *époque* ; cela permet de produire de bons estimés qui détectent bien la tendance évolutive de la taille de population effective à travers le temps.

La performance de la méthode *skywis plot* pour la capture du signal démographique contenu dans les séquences contemporaines d'ADN, a été illustrée par simulation, en utilisant plusieurs scénarios démographiques pour lesquels la taille de la population effective varie de manière « modérée ». Il s'est avéré que le *skywis plot* permet de reconstruire correctement l'historique démographique récent des séquences selon plusieurs scénarios démographiques proposés. En particulier, notre méthode permet de capter les points de changement de la taille de la population effective. De plus, on a trouvé que la performance du *skywis plot* est comparable à la méthode *skyline plot bayésien* qui utilise des techniques phylogénétiques et un échantillonnage MCMC.

La méthode a été ensuite généralisée au cas d'un échantillonnage hétérochrone, où, en plus d'introduire le cadre méthodologique adéquat en adaptant la fonction d'importance de Stephens et Donnelly (2000), il a été illustré par simulation qu'il est possible, en présence de telles séquences hétérochrones, d'améliorer la perfor-

mance du *skywis plot* dans le cas d'une croissance exponentielle de la taille de la population effective. Cela est encore plus marquant à l'approche du temps de l'ancêtre commun, puisque cela permet l'ajout d'information qui est bénéfique à la méthode au moment de la reconstruction des généalogies qui se fait, rappelons-le, du présent vers le passé. La méthode *skywis plot* permet de bien reconstruire l'historique démographique dans des cas où le changement de la taille de la population n'est pas brutal. Par contre, cette approche très flexible n'arrive pas à bien capter une forte augmentation/réduction de la taille d'une population. Cela nous a amené à développer une seconde méthode, qui est étudiée dans la suite de la thèse.

Nous avons donc proposé d'améliorer la performance de la méthode *skywis plot* en améliorant le biais de l'échantillonnage pondéré, en supposant d'abord la disponibilité d'information supplémentaire sur l'estimé de la taille de la population relative à différents instants. Techniquement, cela rend le processus du nombre d'événements (coalescence ou mutation) non-homogène et a une incidence majeure sur la méthodologie utilisée pour la simulation des généalogies. Plus particulièrement, la simulation du temps du prochain événement, ainsi que la fonction d'importance utilisée, sont très affectées. Cette nouvelle méthode a été appelée, *calibrated skywis plot*, car l'estimation de la taille de la population relative aux temps d'échantillonnage nous permet de simuler différemment les temps entre deux événements en opérant un « calibrage » sur chaque intervalle inter-échantillonal.

L'hypothèse de l'existence d'une estimation au préalable de la taille de la population relative à différents moments a été ensuite levée, en proposant une procédure itérative, *iterative calibrated skywis plot*. Dans cette méthode, la taille de la population effective est ainsi approximée par une fonction en escalier, où les estimés sont réestimés après chaque itération en utilisant la méthode *calibrated skywis plot*. Ces fonctions en escalier sont utilisées pour générer les temps d'attente d'un

processus de Poisson non homogène (coalescence avec mutation) sous un modèle avec une taille de population variable. Cela nous a amené à adapter la distribution proposée de Stephens et Donnelly (2000). Comme illustration, nous avons appliqué la méthode *iterative calibrated skywis plot* sur un ensemble de données simulées à partir d'un modèle où la taille de la population effective évolue de manière exponentielle, à croissance rapide. Nous avons montré que la nouvelle méthode améliore nettement le résultat trouvé par la méthode *skywis plot.*

Dans le futur, nous prévoyons généraliser notre ensemble de méthodes *skywis plot* en incluant la recombinaison, qui induit une structure de graphe, plutôt que d'arbre. En effet, contrairement aux méthodes basées sur la phylogénétique, cela est possible, puisque les méthodes IS ont été déjà développées dans ce contexte (par exemple, Fearnhead and Donnelly, 2001). De plus, nos méthodes pourraient être appliquées à des modèles de substitution plus complexes, ce qui est plus réaliste, notamment dans le cas de virus ARN qui évoluent rapidement.

Enfin, notons que la méthode *skywis plot* a fait l'objet d'un article scientifique publié dans la revue *Frontiers in Genetics* (Ait Kaci Azzou *et al.*, 2015), tandis que les méthodes *calibrated skywis plot* et *iterative calibrated skywis plot* ont fait l'objet d'un deuxième article à paraître dans la revue *Theoretical Population Biology.*

BIBLIOGRAPHIE

Ait Kaci Azzou, S., Larribe, F. et Froda, S. (2015). A new method for estimating the demographic history from dna sequences : an importance sampling approach. *Frontiers in Genetics*, *6*(259). `http://dx.doi.org/10.3389/fgene.2015.00259`

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Donnelly, P. et Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet*, *29*, 401–421.

Drummond, A., O.G.Pybus, Rambaut, A., Forsberg, R. et Rodrigo, A. (2003a). Measurably evolving populations. *Trends in Ecology and Evolution*, *18*, 481–488.

Drummond, A., Pybus, O. et Rambaut, A. (2003b). Inference of viral evolutionary rates from molecular sequences. *Advances in Parasitology*, *54*, 331–358.

Drummond, A., Rambaut, A., Shapiro, B. et Pybus, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*, 1185–1192.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. et Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, *161*, 1307–1320.

Drummond, A. J., W.Ho, S. Y., Phillips, M. J. et Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*.

Durbin, R. et Li, H. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*, 493–496.

Ewens, W. J. (2004). *Mathematical Population Genetics*. Springer-Verlag, New York. Second Revised Edition.

Excoffier, L. et Foll, M. (2011). fastsimcoal : a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*, 1332–1334.

Fearnhead, P. et Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, *159*(3), 1299–1318.

Felsenstein, J. (1981). Evolutionary trees from dna sequences : A maximum likelihood approach. *J. Mol. Evol*, *17*, 368–376.

Felsenstein, J. (1989). Phylip - phylogeny inference package (version 3.2). *Cladistics*, *5*, 164–166.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Oxford University Press, Oxford.

Fu, Y. (1994). A phylogenetic estimator of effective population size or mutation rate. *Genetics*, *136*, 685–692.

Galtier, N., Gascuel, O. et Jean-Marie, A. (2005). Introduction to markov models in molecular evolution. *Statistical Methods in Molecular Evolution, Rasmus Nielsen, Springer*, 3–24.

Griffiths, R. C. et Tavaré, S. (1994a). Ancestral inference in population genetics. *Statistical Science*, *9*, 307–319.

Griffiths, R. C. et Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London, Series B*, *344*, 131–159.

Griffiths, R. C. et Tavaré, S. (1994c). Simulating probability distributions in the coalescent. *Theory Popln Biol*, *46*, 131–159.

Gutierrez, S., Yvon, M., Pirolles, E., Garzo, E., Fereres, A., Michalakis, Y. et Blanc, S. (2012). Circulating virus load determines the size of bottlenecks in viral populations progressing within a host. *PLOS Pathogens*, *8*(11).

Hanada, K., Suzuki, Y. et Gojobori, T. (2004). A large variation in the rates of synonymous substitution for rna viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular Biology and Evolution*, *21*, 1074–1080.

Hasegawa, M., Kishino, H. et Yano, T. (1985). Dating of human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, *22*, 160–174.

Hein, J., Schierup, M. et Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution :A Primer in Coalescent Theory.* Oxford University Press, USA.

Heled, J. et Drummond, A. (2008). Bayesian inference of population size history from multiple loci. *BMC Evol Biol, 8-15.*

Ho, S. et Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources, 11*(3), 423–434.

Holmes, E. C. (2009). *The evolution and emergence of RNA viruses.* Oxford University Press, Oxford, United Kingdom.

Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxf. Surv.Evol. Biol, 7*(1-44).

Jenkins, G., Rambaut, A., O.G.Pybus et Holmes, E. (2002). Rates of molecular evolution in rna viruses : a quantitative phylogenetic analysis. *Journal of Molecular Evolution, 54*, 152–161.

Jukes, T. et Cantor, C. (1969). Evolution of protein molecules. mammalian protein metabolis. *m.Academic Press, New York*, 21–132.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics, 61*(4), 893–903.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol, 16*, 111–120.

Kimura, M. et Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics, 49*, 725–738.

Kingman, J. (1982a). The coalescent. stochastic process and their applications. *Journal of Applied Probability, 13*, 235–248.

Kingman, J. (1982b). On the genealogy of large populations. *Journal of Applied Probability.*

Kuhner, M. K., Yamato, J. et Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics, 140*(1421-1430).

Kuhner, M. K., Yamato, J. et Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics, 149*, 429–434.

Lai, M. (1992). Rna recombination in animal and plant viruses. *Microbiological Reviews, 56*, 61–79.

Lwoff, A. (1957). The concept of virus. *J. Gen. Microbiol*, *17*, 239–253.

Maretty, L., Sibbesen, J. et K.Dialdestoro (2013). Coalescent inference from serially sampled, next-generation sequencing data from hiv patients.

MATLAB. (2013). *version 8.1.0.604 (R2013a)*. Natick, Massachusetts : The MathWorks Inc.

Minin, V., Bloomquist, E. et Suchard, M. (2008). Smooth skyride through a rough skyline : Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol*, *25*, 1459–1471.

Nordborg, M. (2001). In handbook of statistical genetics. *M. J. Bishop and C. Cannings, Eds. John Willey and Sons, Inc*, p. 179–212.

Nordborg, M. (2007). *Coalescent theory.* In Handbook of Statistical Genetics (Edited by D. J.Balding, M. J. Bishop, and C. Cannings) . John Wiley and Sons, Inc., Chichester, U.K.

Owen, A. B. (2009). *Monte Carlo theory, methods and examples.* Copyright Art Owen.

Paradis, E., Claude, J. et Strimmer, K. (2004). Ape : analyses of phylogenetics and evolution in r language. *Bioinformatics*, *20*, 289–290.

Pybus, O., Drummond, A., Nakano, T., Robertson, B. et Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis c virus in egypt : a bayesian coalescent approach. *Mol Biol Evol*, *20*, 381–387.

Pybus, O., Rambaut, A. et Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies genetics. *GENETICS*, *155*(3), 1429–1437.

R. Opgen-Rhein, L. Fahrmeir, K. S. K. (2005). Inference of demographic history from genealogical trees using reversible jump markov chain monte carlo. *BMC Evolutionary Biology*, *6*(5).

Rodrigo, A. G. et Felsenstein, J. (1999). Coalescent approaches to hiv population genetics. *Molecular Evolution*, 233–272.

Rodrigo, A. G., Shpaer, E. G., Delwart, E. L., Iversen, A. K. et et al, M. V. G. (1999). Coalescent estimates of hiv-1 generation time in vivo. *PNAS*, *96*(5), 2187–2191.

Rosenberg, N. et Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*, 380–390.

Ross, S. (1995). *Stochastic processes.* Wiley, New York, NY.

Ross, S. (2013). *Simulation.* Academic Press, 5th edition.

Slatkin, M. et Hudson, R. R. (1991). Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics*, *129*(2), 555–562.

Stephens, M. (2000). Times on trees, and the age of an allele. *Theoretical Population Biology*, *57*, 109–119.

Stephens, M. (2001). Inference under the coalescent. *In Handbook of Statistical Genetics*, 213–238.

Stephens, M. et Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society*, *62*, 605–655.

Strimmer, K. et Pybus, O. G. (2001). Exploring the demographic history of dna sequences using the generalized skyline plot. *Mol. Biol. Evol*, *18*(12), 2298–2305.

Tavaré, S. et Zeitouni, O. (2004). *Stochastic processes.* LNM 1837, J. Picard (Ed.) Springer-Verlag Berlin Heidelberg.

Wakeley, J. (2008). *Coalescent Theory : An Introduction.* Roberts and Company Publishers.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *2*(16), 97–159.