Modèles stochastisques de populations structurées en génétique des populations neutre

Pierre Pudlo et Mohammed Sedki

14 mars 2017

Abstract

Cet article décrit quelques modèles de génétique des populations sous neutralité, incluant dérive génétique et mutations. À partir du coalescent de Kingman, nous montrons comment on peut modéliser des populations structurées. Nous détaillons ces modèles en montrant comment il est possible d'écrire des algorithmes de simulations. En particulier, nous mettons en avant l'ensemble des processus latents qui rendent le calcul de la fonction de vraisemblance sur un jeu de données difficile, voire impossible.

This paper describes some population genetic models under neutrality, involving genetical drift and mutations. From Kingman's coalescent we show how structured populations can be modeled. We detail those models by showing how we can write simulation algorithms. In particular we highlight the latent processes than ban the explicit computation of the likelihood function on a dataset.

Keywords: génétique des populations, coalescent de Kingman, vraisemblance complexe, population genetics, Kingman's coalescent, intractable likelihood.

1 Introduction

La génétique des population s'intéresse à la distribution du polymorphisme génétique au sein de populations d'individus d'une même espèce. Deux mécanismes gouvernent la distribution des différentes modalités génétiques, appelées allèles, au cours du temps : un processus de mutation (essentiellement un processus markovien sur l'espace des allèles possibles) et des variations de fréquences d'une génération à l'autre dues au succès reproducteur de chaque individu. Ce processus, appelé dérive génétique, est gouverné par la structuration de la population en sous-populations et la taille de ces populations. La neutralité génétique Kimura (1968, 1983) suppose que les différents allèles possibles en une position donnée du génome (appelée locus) ne sont ni avantageux, ni délétères pour les individus qui les portent. Cette hypothèse de neutralité n'est évidemment pas réaliste en toute position du génome : il est bien connu que certains loci sont sous sélection, autrement dit que les différents allèles en cette position du génome ne sont pas tous égaux et fournissent un avantage ou un inconvénient à leurs porteurs. Ils influent alors sur les succès reproducteurs de leurs porteurs. Mais, sous neutralité, le succès reproducteur est

indépendant du type génétique, et les deux processus d'évolution mentionnés ci-dessus (mutation et dérive génétique) sont indépendants. L'enjeu du point de vue statistique est alors d'inférer l'histoire passée des populations à partir de données de polymorphisme génétique sur un échantillon d'individus collecté à la date actuelle. Le but de cet article n'est pas de présenter les méthodes d'inférence pour répondre à ces questions, mais, plus simplement, de présenter les modèles stochastiques qui permettent de relier paramètres démographiques et mutationnels aux données génétiques collectées à la date actuelle. Autrement dit, il s'agit d'expliquer des modèles qui permettent de définir une fonction de vraisemblance.

En particulier, nous présentons ici toute une large gamme de modèles stochastiques sous neutralité qui font parti du folklore de la génétique des populations, mais sont rarement décrits avec précision. La solution la plus pédagogique pour présenter ces modèles est d'exposer des schémas de simulations. Nous laissons le lecteur en déduire les formules mathématiques dont il a besoin, en particulier celles de vraisemblances complétées. En outre, nous espérons convaincre le lecteur que la fonction de vraisemblance des données n'est pas explicite, mais s'écrit comme une intégrale sur tout un processus latent qui décrit l'histoire passée des ancêtres de l'échantillon.

Résumons le contenu de ce article. Les données génétiques s'expliquent par quelques phénomènes d'évolution sous-jacents et un peu de biologie (section 2). Un scénario évolutif est une série d'événements spatio-temporels ordonnés du plus récent au plus ancien (figure 1). Ces modèles expliquent des jeux de données à l'aide d'un processus latent. La composante latente inclut une généalogie et un ensemble de génotypes ancestraux. La généalogie (section 3) est représentée par un dendrogramme dont une lignée correspond à l'ascendance d'un individu depuis l'ancêtre commun le plus récent (MRCA) des échantillonns observés. L'évolution des lignées d'une généalogie est régie par les différents événements inter-populationnels présents dans le scénario évolutif (section 3). Les génotypes d'un jeu de données dérivent de celui du MRCA via une série de mutations (section 4). Une réalisation d'un processus ponctuel positionne les mutations sur la généalogie. Selon le type de données de l'étude (section 2), un modèle mutationnel (section 4) engendre le génotype du MRCA et applique les mutations sur les branches du dendrogramme. La composante latente ajoute une dimension temporelle au modèle évolutif. Ainsi, l'interprération des paramètres de ce modèles pose des problèmes d'échelle de temps (section 5). Indépendamment de l'approche d'inférence statistique, le calcul de la vraisemblance $\ell\big(\mathbf{x}\big|\phi\big)$ nécessite une marginalisation. Celle-ci s'obtient par une intégration sur l'espace de la composante latente (section 5).

2 Biologie, évolution et données génétiques

Le but de cette partie est de décrire quelques éléments de génétique afin que le lecteur, probabiliste ou statisticien ne soit pas perdu.

Chez l'homme, le matériel génétique nucléaire de nos cellules est organisé en 23 paires de chromosomes. Chaque chromosome organise linéairement le matériel génétique le long de molécules (ou brins) d'ADN en double hélice. Il existe aussi un brin d'ADN hors du noyau, dans la mitochondrie. Seule une petite partie de ces molécules portent l'information codant pour les protéines (les exons, qui constituent moins de 2% du

génome nucléaire chez l'homme). Lors de son utilisation par nos cellules, une partie plus importante de la molécule d'ADN est transcrite, qui inclue des introns. Il convient de distinguer deux types de paires de chromosomes : les chromosomes autosomaux (22 paires chez l'homme), et les chromosomes sexuels (1 paire chez l'homme). Sauf différences dues à des mutations, les chromosomes d'une paire autosomale portent la même information génétique. Toutes les espèces dont le génome est organisé par paires sont dite diploïdes. Mais il existe des espèces avec des niveaux de ploïdie différents. On trouve ainsi quelques espèces tétraploïdes (organisation par quatre au lieu de paires) ou hexaploïdes (par six), surtout chez les végétaux. D'autres spécificités existent encore, par exemple, chez l'abeille, les femelles sont diploïdes alors que les mâles sont haploïdes (une seule copie de chaque chromosome). Quant à l'organisation de la paire de chromosomes sexuels chez l'espèce humaine, il est bien connu que les femmes sont XX, autrement dit ont deux chromosomes X identiques (sauf différences dues à des mutations) et les hommes un chromosome X et un Y. Certaines espèces fonctionnent suivant le même système XY, mais il existe d'autres systèmes de chromosomes sexuels ZW, XX-X0, etc. Enfin, le matériel génétique hors du noyau est beaucoup plus court. Par exemple, il y a un brin d'ADN dans les mitochondries de nombreuses espèces, et chez les végétaux, on trouve également des brins d'ADN dans les chloroplastes.

Reste à comprendre comme ce patrimoine génétique est hérité d'une génération à l'autre. Pour tous les brins d'ADN hors du noyau, l'héritage est très simple, il est clonal. Chez l'homme par exemple, le brin d'ADN mitochondrial est hérité par copie de celui de la mère. C'est pourquoi il a beaucoup été utilisé dans des études de génétique des populations, voir par exemple les questions d'Ève mitochondriale Majoram and Donnelly (1997); Griffths and Tavaré (1994). Pour les espèces diploïdes à reproduction sexuée, les individus portent l'information génétique nucléaire en double : une copie issue de la gamète maternelle, une copie issue de la gamète paternelle. Cet appariement n'apporte aucune information lorsque l'équilibre d'Hardy-Weinberg (Crow and Kimura, 2009, chapitre 2) est réalisé. On peut donc assimiler un individu diploïde aux deux gamètes qui l'ont engendré, c'est-à-dire à deux individus haploïdes. Dans toute la suite, nous considérons donc que les individus sont haploïdes. Entre outre, la copie issue d'une gamète d'un parent n'est pas la reproduction exacte d'un des chromosomes de ce parent. Un phénomène de recombinaison se produit lors de la méiose, c'est-à-dire la création de la gamète : le chromosome de la gamète est un mélange des deux chromosomes homologues du parent.

Pour chacun des individus, l'information génétique que l'on considère dans un jeu de données est limitée. On ne s'intéresse qu'à quelques positions particulières du génome appelées locus. À ces locus, la séquence d'ADN peut varier d'un individu à l'autre, à cause des mutations au cours de l'évolution de l'espèce. On parle alors de polymorphisme génétique. Les différentes variantes s'appellent des allèles ou des états alléliques. La constitution de notre jeu de données a nécessité de déterminer l'allèle que porte chacun des individus pour tous les locus de l'étude. Cette opération de génotypage fait appel à de la micro-biologie que nous ne décrirons pas ici, et nous admettons que le résultat fourni est exact (sans bruit ou erreur de mesure).

Enfin, dans les modèles que nous utilisons, nous considérons que les locus issus du génome nucléaire sont indépendants. Notre vraisemblance s'écrit donc comme un produit de vraisemblances du jeu de données restreint à un locus donné. Cette hypothèse cache en fait un modèle de vraisemblance composite Hudson (2001); Stumpf and McVean (2003).

Grace au brassage dû à la recombinaison génétique, cette approximation est valable dès lors que les locus sont suffisamment éloignés sur le génome (en particulier pour que les généalogies d'un locus à l'autre soient indépendantes). Signalons ici qu'il existe des processus plus complexes pour modéliser la dépendance entre locus, à partir du graphe de recombinaison ancestrale de Griffiths and Marjoram (1997).

Un jeu de données est constitué de différents échantillons d'individus. Chaque échantillon correspond à une population géographique (appelée parfois colonie ou dème). Nous numérotons les populations de 1 à D et leur donnons les labels $Pop1$ à $PopD$. Contrairement à de nombreux problèmes en statistique, l'appartenance des individus échantillonnés aux populations d'intérêt est ici connue. Par population, la taille typique d'un échantillon varie entre une quarantaine et une centaine d'individus. La taille de l'échantillon issu de la population $Popi$ est notée n_i.

Les modèles probabilistes que nous présentons ici expliquent le polymorphisme en retraçant l'évolution de l'espèce et les mutations au cours du temps. Nous devons donc détailler quelles sont les formes possibles d'information génétique pour chaque locus. En fait, il existe trois types de locus : microsatellite, séquence ou SNP (Single Nucleotide Polymorphism). Nous nous concentrons ici sur les locus microsatellites. Il s'agit d'une partie de l'ADN où un court motif (de 1 à 4 paires de base) est répété en de nombreux exemplaires. L'information que l'on retrouve dans notre jeu de données est alors la longueur totale de cette séquence (en nombre de paires de base). À cause de ces répétitions, cette portion d'ADN est fortement variable (fort polymorphisme) et a donc été beaucoup utilisé en génétique des populations. Nous décrivons deux modèles mutationnels classiques sur ce type de locus dans la section 4.

3 Généalogie d'échantillons

Cette section est composée de deux parties. Premièrement, on se restreint à un scénario très simple avec une seule population fermée à l'équilibre. On introduit l'outil fondamental pour simuler une généalogie qui est le processus appelé coalescent de Kingman (1982a,c,b). Ensuite, on s'intéresse à la généalogie de populations géographiquement structurées. Ces populations sont stucturées par des événements inter-populationnels. On décrit l'évolution des lignées d'une généalogie en présence pour d'événements inter-populationnels instantanés, qui sont la divergence et l'admixture, et d'événements persistants de migration.

Certains modèles (Wright-Fisher, Moran, etc. voir Wakeley, 2005, chapitre 3) proposent de simuler l'évolution de la population entière, du passé au présent, puis d'échantillonner la dernière génération. Pour une population de grande taille, la simulation suivant ce type de modèles est très lente. D'où la méthodologie que l'on adoptera et qui consiste à considérer seulement l'échantillon d'intérêt au lieu de la totalité de la population. On s'intéresse seulement à l'évolution des ascendants des individus de notre échantillon en remontant le temps.

3.1 Une seule population : Coalescent de Kingman

Introduisons le coalescent de Kingman (1982a,c,b), qui est l'outil fondamental pour la simulation des généalogies. Par souci de clarté, nous nous contentons d'un scénario

Algorithme 1 Coalescent de Kingman en temps naturel

Entrées : La taille de l'échantillon k, la taille efficace de la population Ne.

Tant que $k \geq 2$ **faire**

1) Simuler le temps inter-coalescent T_k suivant une loi exponentielle de paramètre $k\big(k-1\big)\Big/(2Ne)$.

2) Augmenter les longueurs des k lignées de T_k.

3) Parmi les k lignées, choisir aléatoirement deux lignées à regrouper pour former un nœud du dendrogramme.

4) $k \leftarrow k-1$.

Fin tant que

basique formé d'une seule population fermée à l'équilibre. Nous supposons que cette population n'est pas soumise à un flux extérieur de gènes et ne subit aucune variation démographique interne.

La généalogie d'un échantillon d'individus est représenté par un dendrogramme (figure 2). On génère des lignées ancestrales jusqu'à l'ancêtre commun le plus récent (MRCA en anglais). Chaque lignée d'un individu passe par une série d'ancêtres. Un évènement de coalescence se produit lorsque les lignées de deux individus se rejoignent en un noeud du dendrogramme (figure 2). La généalogie d'un échantillon de k individus est donc composée de $k-1$ évènements de coalescence. Chaque évènement décroît le nombre de lignées ancestrales de 1 jusqu'à la dernière lignée (racine du dendrogramme) qui correspond au MRCA.

Les variables $T_k, \ldots, T_2$ représentent les durées entre les événements de coalescences (figure 2) successifs. La loi de la généalogie de k individus est entièrement caractérisée par la loi du choix des lignées à chaque évènement de coalescence et la loi des durées entre évènements $T_k, \ldots, T_2$. Pour le coalescent de Kingman, les durées entres évènements de coalescences $T_k, \ldots, T_2$ sont indépendantes et T_k suit la loi exponentielle de paramètre $k\big(k-1\big)/2$.

Ici, nous détaillons la réalisation d'un événement de coalescence dans un échantillon ancestral de taille k. Pour chacun des $\binom{k}{2}$ couples de lignées en compétition pour une coalescence, on associe une horloge exponentielle de paramètre 1. L'horloge qui réalise le temps minimal désigne le couple de lignées à coalescer. En pratique, la réalisation des $\binom{k}{2}$ variables exponentielles de paramètre 1 est lent. Nous pouvons diminuer le nombre de variables aléatoires exponentielles à réaliser avec le lemme ci-dessous.

Lemme 1. *Soient $T_1, \ldots, T_\ell$ des variables aléatoires exponentielles indépendantes de paramètres respectifs $\lambda_1, \ldots, \lambda_\ell$. Alors la variable aléatoire $\inf\limits_{1 \leq i \leq \ell} T_i$ suit une loi exponentielle de paramètre $\sum_{i=1}^{\ell} \lambda_i$. Et la variable aléatoire T_k réalise cet infimum avec probabilité $\lambda_k \big/ \sum_{i=1}^{\ell} \lambda_i$.*

Ainsi, à l'aide du lemme 1, la réalisation d'un événement de coalescence revient donc à augmenter les k lignées du dendrogramme d'une réalisation exponentielle de paramètre $\binom{k}{2}$. Le couple de lignées à coalescer est choisi aléatoirement parmi les $\binom{k}{2}$ couples.

Une unité de temps coalescent s'interprète comme Ne générations, où Ne est un paramètre du modèle qui s'appelle taille efficace de la population (plus de détails en section 5). On décrit dans l'algorithme 1 le coalescent de Kingman sur un échantillon à k individus d'une population de taille efficace Ne. Dans cet algorithme, le temps est à l'échelle naturelle, et donc le taux de coalescence dans la généalogie est linéaire en Ne.

Enfin, notons que pour ne pas croiser les lignées du dendrogramme, l'ordre des individus sur les feuilles est différent l'ordre des individus dans le jeu de données. Sur l'exemple de généalogie donné dans la figure 2, les numéros en bas du dendrogramme correspondent aux numéros des individus dans le jeu de données.

3.2 Plusieurs populations structurées

Algorithme 2 Généalogie (partielle) dans une population indépendante

Entrées : La taille de l'échantillon ancestral notée k à la date t dans la population d'intéret. La taille efficace Ne de la population et les dates des événements t et t'.

Simuler T_k suivant une loi exponentielle de paramètre $k\big(k-1\big)/2Ne$.
Tant que $\big(t+T_k\big) \leq t'$ **faire**

1) Augmenter les longueurs des k lignées d'une longueur T_k.

2) Choisir aléatoirement parmi les k lignées, deux lignées à regrouper pour former un nœud du dendrogramme.

3) $k \leftarrow k-1$.

4) Simuler la durée inter-coalescence T_k suivant une loi exponentielle de paramètre $k\big(k-1\big)/2Ne$.

Fin tant que

Si $\big(t+T_k\big) > t'$ **alors**
Augmenter les lignées restantes jusqu'à la hauteur t'.
Fin si

Algorithme 3 Généalogie de populations avec migration en temps coalscent

Entrées : Les tailles des échantillons ancestraux: $k_1, \ldots, k_D$ et les taux de mirgation m_{ij}.

Pour i $=1 \rightarrow D$ **faire**

1) Associer une horloge exponentielle de paramètre $1/Ne_i$ pour chaque couple d'individus de la population i qui correspond à une coalescence potentielle.

2) Associer $D-1$ horloges exponentielles de paramètres $m_{ij}, 1 \leq j \neq i \leq D$ pour chaque individu de la population i qui correspondent à des migrations potentielles.

Fin pour
Parmi toutes les horloges en compétition, celle qui sonne en premier gagne. Si cette horloge correpond à un couple d'individus, on fait coalescer ces deux individus. Si c'est l'horloge d'un seul individu, et celle-ci est de paramètre m_{ij}, on déplace la lignée de cet individu de la population i vers la population j.

Décrivons la loi d'une généalogie pour un scénario évolutif dont la structure géographique est gouvernée par des événements inter-populationnels. On combine ces événements avec le coalescent de Kingman qui décrit la généalogie intra-populationnel. Présentons succintement les trois types d'événements inter-populationnels.

- La divergence (figure 3, (a)) est la fusion de deux populations.
- L'admixture (figure 3, (b)) est le partage d'une population en deux partie à l'instant de l'événement. Les lignées sont envoyées dans les deux autres populations. La destination d'une lignée est contrôlée par un paramètre du modèle appelé, taux d'admixture.
- La migration (figure 3, (c)) autorise le déplacement des lignées d'une population à l'autre sur une période donnée, suivant des taux par unité de temps et par gène.

La divergence et l'admixture sont des événements instantanés alors que la migration est persistante sur une période de temps.

La procédure de simulation de la généalogie dans ces trois situations simule l'évolution des lignées du dendrogramme entre deux dates d'événements inter-populationnels successifs et les changements instantanés dans les lignées de la généalogie pour ces évènements instantanés. En cas de migration sur une période donnée, l'évolution des lignées du dendrogramme sur cette période est différente des situations précédentes. Les événements de coalescences et de déplacements de lignées sont en concurrence. L'algorithme 5 résume les étapes de simulation d'une généalogie contrainte par un modèle démographique avec différents événements inter-populationnels.

Généalogie entre événements inter-populationnels instantanés L'évolution de la généalogie intra-populations entre deux dates (notées t et t' où $t' > t$ dans la figure 3 (a) et (b)) d'événements suit un coalescent de Kingman indépendant sur les lignées présentes dans chaque population. L'algorithme 2 décrit la généalogie à l'intérieur de chacune des populations. Il diffère légèrement de l'algorithme 1 décrivant l'évolution d'une population fermée à l'équilibre jusqu'au MRCA. En effet, l'algorithme 2 décrit le coalescent de Kingman sur une période $t' - t$ entre deux événements inter-populationnels. En cas de présence d'un changement de taille efficace (figure 1) dans la population à une date, il suffit de changer l'échelle de temps après cette date et donc de remplacer Ne par Ne' dans la simulation des durées inter-coalescences T_k.

Divergence À l'instant de la divergence, les lignées présentes dans les deux populations ($Pop1$ et $Pop2$ de la figure 3 (a)) sont regroupées pour former une seule population ($Pop1$ de la figure 3 (a)).

Admixture À l'instant de l'admixture, l'échantillon ancestral de $Pop3$ (voir figure 3, (b)) est partagé sur les deux autres populations ainsi: une lignée de la population $Pop3$ est envoyée dans $Pop1$ avec probabilité r et dans $Pop2$ avec probabilité $1 - r$, où r est un paramètre du modèle appelé taux d'admixture.

Migration Nous décrivons dans l'algorithme 3 l'évolution des lignées en présence d'une migration entre D populations. La migration est paramétrée dans le modèle par les taux de migration de la population i vers la population j, notés m_{ij} . L'algorithme 3 détaille toutes les horloges en compétition dont les lois sont exponentielles avec les paramètres appropriés. L'horloge qui réalise le temps minimal de ce système désigne le type d'événement, les populations et les individus consernés par l'événement. En pratique, la simulation d'une généalogie de populations structurées par une migration nécessite beaucoup moins d'horloges exponentielles que dans l'algorithme 3. Comme dans le cas de la partie 3.1, nous pouvons simplifier cet algorithme à l'aide du lemme 1. Nous donnons dans l'algorithme 4 un exemple avec $D = 2$ populations, et nous renvoyons le lecteur à (Wakeley, 2005, chapitre 5) pour le cas $D > 2$. La généalogie décrite dans cet algorithme correspond au scénario de la figure 3 (c). Cet algorithme est répété sur les populations $Pop1$ et $Pop2$ jusqu'à ce que les lignées du dendrogramme soient à la hauteur t'.

En conclusion, nous donnons le schéma général de simulation d'un dendrogramme contraint par de tels événements dans l'algorithme 5. Le coalescent de Kingman est bien la pierre de base de ces généalogies de gènes, qui est utilisé par morceaux, ou en compétition avec de la migration. De l'évolution des populations d'intérêt, nous avons extrait le seul processus utile au regard de l'échantillon grâce à cette généalogie. Reste à modéliser la dérive génétique le long de ses branches.

4 Processus mutationnels

Voyons maintenant quelle est la loi des génotypes de l'échantillon conditionnellement à une généalogie. Nous décrivons les positions des mutations comme une réalisation d'un processus ponctuel de Poisson sur les branches du dendrogramme. Ensuite nous introduisons deux modèles mutationnels sur locus microsatellite. L'application des mutations revient à faire des pas d'une chaîne de Markov associée au modèle mutationnel.

Positions des Mutations Le taux de mutation par unité de temps naturel et par individu diploïde est le paramètre μ. Ainsi, conditionnellement à une généalogie, les positions des mutations sont données par un processus ponctuel de Poisson d'intensité $\mu/2$ sur le dendrogramme. Autrement dit, sur une branche de longueur t, le nombre N de mutations suit une loi de Poisson de paramètre $\mu t/2$, et les N mutations sont uniformément réparties sur cette branche.

Modèles Mutationnels sur locus microsatellite Présentons maintenant le processus d'évolution à chaque mutation. Nous introduisons ici deux modèles mutationnels (voir Whittaker et al. (2003) et Cornuet et al. (2006)) SMM (Stepwise Mutation Model) et GSM (Generalized Mutation Model) spécifiques au locus microsatellite. Ces deux modèles mutationnels ont l'avantage d'avoir une paramétrisation simple. Les chaînes de Markov associées aux modèles GSM et SMM sont des marches aléatoires symétriques sur un intervalle de nombres entiers $[\![a; b]\!]$ de $\mathbb{N}$. Appliquer un pas de la chaîne de Markov GSM, revient à modifier le locus d'une longueur $\pm mG$, où m est la longueur du motif répété (supposée connue), G est une variable aléatoire de loi géométrique de paramètre

p et $\pm$ un signe aléatoire. En pratique, le paramètre p est de l'ordre de 0.2. Quant au modèle SMM, une mutation revient à diminuer ou augmenter (avec probabilités 1/2 et 1/2 respectivement) le locus d'une longueur de m paires de base où m est la longueur du motif répété. Il arrive qu'en appliquant les mutations, le génotype dépasse les bornes a et b de l'ensemble des états alléliques. Dans ce cas, le génotype de l'ancêtre prend comme valeur la plus proche des deux bornes a, b de l'espace des états alléliques.

Pour simuler les génotypes de l'échantillon en un locus donné, il suffit de faire évoluer le génotype du MRCA le long de la généalogie jusqu'au présent en appliquant les mutations (algorithme 6). Le génotype du MRCA provient de la loi stationnaire du modèle mutationnel. Observons le processus qui génère les génotypes des individus (figure 4). L'évolution de l'ascendance d'un individu le long d'une lignée est un processus markovien de sauts purs sur $[\![a; b]\!]$. Ici, la chaîne de Markov du modèle mutationnel est la chaîne incluse de ce processus. La chaîne incluse et le processus markovien ont la même loi stationnaire car le taux de mutation (taux de saut) ne dépend pas de l'état allélique. Les génotypes des individus observés sont donnés par les fins de trajectoires de ce processus (figure 4). Une trajectoire correspond à une lignée de la figure 4. Ces trajectoires sont corrélées et cette dépendance est décrite par les branches partagées par différentes lignées. Puisque le génotype du MRCA provient de la loi stationnaire, on déduit que la loi marginale du génotype d'un individu est cette même loi stationnaire. En revanche, la loi jointe des génotypes de tous les individus est très complexe à décrire. La structure de corrélation entre les trajectoires du processus est donnée par les branches partagées par les lignées du dendrogramme. Par exemple, sur la figure 4, les processus markoviens expliquent les génotypes des gènes 2 et 4 (représentés en rouge et vert respectivement sur la figure) ont en commun une grande partie de leurs trajectoires. La loi marginale de l'échantillon (*i.e* loi déconditionnée) n'a pas d'écriture plus simple qu'une intégrale contre la loi de la généalogie.

5 Conclusion

Vraisemblance Notons $f_\phi(\mathcal{G})$ la densité de la loi de la généalogie de gènes par rapport à une mesure de référence $\mathrm{d}\mathcal{G}$. Rappelons que nous avons décrit cette loi à partir du coalescent de Kingman dans la section 3. De même, notons $f_\phi(\mathcal{M}|\mathcal{G})$ la densité du processus mutationnel $\mathcal{M}$ sachant la généalogie $\mathcal{G}$ décrit dans la section 4. La vraisemblance du jeu de données complet $\mathbf{x}$ s'écrit donc

$$\ell\big(\mathbf{x}|\phi\big) = \prod_{i\in\{locus\}} \int_{\mathcal{M}_i\to\mathbf{x}_i} f_\phi\big(\mathcal{M}_i\big|\mathcal{G}_i\big) f_\phi\big(\mathcal{G}_i\big)\,\mathrm{d}\mathcal{G}_i\,\mathrm{d}\mathcal{M}_i, \tag{1}$$

où $\mathbf{x}_i$ est l'ensemble des données au locus i et $\mathcal{M}_i \to \mathbf{x}_i$ désigne l'ensemble des génotypes sur le dendrogramme dont les feuilles correspondent à l'échantillon observé. Nous faisons ici l'hypothèse d'indépendance entre les différents locus, voir section 2.

Cette vraisemblance ne se calcule pas facilement. L'intégrale précédente est sur l'espace des couples $\big(\mathcal{G}_i, \mathcal{M}_i\big)$ compatibles avec l'échantillon $\mathbf{x}_i$. Cet espace est de très grande dimension et comporte des directions discrètes comme les génotypes des ancêtres et des parties continues comme les différentes hauteurs dans la généalogie. En dépit de

la simplicité du coalescent de Kingman et du processus mutationnel, on ne peut espérer aucune simplification formelle dans cette intégrale.

Il existe une formule de récurrence Ethier and Griffiths (1987); Griffiths (1989) reliant la vraisemblance $\ell(\mathbf{x}_i|\phi)$ d'un locus à des vraisemblances $\ell(\mathbf{y}|\phi)$, où $\mathbf{y}$ est un échantillon mono-locus de taille inférieure ou égale à celle de $\mathbf{x}_i$. Par example, dans le cas d'une population fermée à l'équilibre, (De Iorio and Griffiths, 2004a, équation (3)) et, dans le cas de plusieurs populations avec migration, (De Iorio and Griffiths, 2004b, équation (2)). En pratique, l'utilisation de cette formule de récurrence pour calculer la vraisemblance des données est impossible. On se retrouve à devoir calculer la vraisemblance de tous les échantillons possibles de taille inférieure ou égale à celle du jeu de données. L'algorithme est alors de complexité exponentielle en la taille du jeu de données et la taille de l'espace allélique : c'est un phénomène bien connu d'explosion combinatoire.

Échelle de temps Le processus latent $(\mathcal{G}, \mathcal{M})$ introduit une dimension temporelle pour expliquer les données échantillonnées. La plupart des coordonnées du paramètre de ϕ sont exprimés dans l'échelle de cet axe. Mais les données, toutes collectées à l'époque actuelle, ne contiennent aucune information sur cette échelle. En effet, une homothétie de rapport λ sur l'axe temporel ne change pas la loi marginale des données $\mathbf{x}$ si on effectue les transformations suivantes sur les composantes de ϕ:

- les paramètres t, de type date, sont changés en $\lambda\, t$,
- les paramètres Ne, de type taille efficace, sont changés en $\lambda\, Ne$.
- les paramètres μ, de type taux de mutations, sont changés en μ/λ et
- les paramètres m, de type taux de migrations, sont changés en m/λ.

L'analyse fréquentielle se heurte donc à un problème d'identifiabilité. On peut re-paramétrer le modèle de telle sorte que ce problème disparaisse. Pour cela, il est commode d'introduire une taille efficace de référence, notée Ne_{REF}, combinaison linéaire des tailles efficaces en échelle naturelle. Les paramètres identifiables sont alors:

- $\tau_i = t_i/Ne_{REF}$,
- $\overline{Ne_i} = Ne_i/Ne_{REF}$,
- $\theta_i = 4Ne_{REF}\,\mu_i$ et
- $\overline{m_{ij}} = Ne_{REF}\, m_{ij}$,

quitte à supprimer l'une des coordonnées de type taille efficace de ϕ. Il est courant de choisir comme valeur de Ne_{REF} la somme des tailles efficaces des populations échantillonnées. Par exemple, on choisirait $Ne_{REF} = Ne_1 + Ne_2 + Ne_3 + Ne_4$ sur l'exemple de la figure 1. L'interprétation de ces paramètres identifiables pour le biologiste nécessite de choisir "*au doigt mouillé*" une échelle de temps. Ce problème ne se pose pas lorsque l'on mène une analyse bayésienne. La loi *a posteriori* sur ϕ est bien définie même si ce vecteur de paramètres n'est pas identifiable. L'échelle de temps est alors choisie à l'aide de la loi *a priori*. Cela permet d'introduire une variabilité sur le choix arbitraire de cette échelle,

codée au travers d'une loi sur des paramètres directement interprétables. C'est un argument en faveur d'une approche bayésienne, clairement mis en avant par Beaumont and Rannala (2004).

Approximation de la vraisemblance Il existe deux grandes familles d'algorithmes, soit pour échantillonner l'espace des généalogies et mutations latentes et ainsi approcher la fonction de vraisemblance, soit pour échantilloner l'espace des paramètres dans une optique bayésienne, qui suppose d'avoir défini une loi a priori sur l'espace des paramètres.

- La première classe d'algorithmes repose sur l'échantillonnage préférentiel (ou *importance sampling* en anglais). L'idée de ces méthodes est d'introduire une loi auxiliaire pour échantillonner l'espace d'intérêt et d'introduire une pondération des tirages aléatoires qui tient compte compte du rapport entre la loi cible et la loi auxiliaire. Pour approcher la vraisemblance, on s'attaque donc directement au calcul de (1) comme une intégrale contre une loi de densité. La grande difficulité est de choisir la loi auxiliaire appelée loi d'importance ou d'échantillonnage pour contrôler l'erreur de Monte-Carlo. Différentes stratégies *ad hoc* permettent de calibrer une loi d'importance dans ce contexte Stephens and Donnelly (2000); De Iorio and Griffiths (2004a,b); De Iorio et al. (2005).

- La seconde classe d'algorithmes comprend les méthodes bayésiennes approchées (ABC ou *Approximate Bayesian computation* en anglais) qui contournent le calcul de la vraisemblance en se reposant sur de nombreuses simulations suivant le (les) modèle(s). Ainsi, les méthodes ABC Beaumont et al. (2002); Marjoram et al. (2003); Marin et al. (2011) comparent des jeux de données simulés aux données observées au travaers de quantités numériques (statistiques résumées) supposées informatives pour le problème d'intérêt, à savoir le calcul de la loi a posteriori. La cible de ces méthodes de Monte Carlo est une version dégradée de la loi a posteriori : il s'agit de la distribution des paramètres sachant les statistiques résumées (et non plus tout le jeu de données). Ces méthodes fournissent des résultats moins précis que les attaques directes par échantillonnage préférentiel pour approcher la fonction de vraisemblance en chaque point où l'on souhaite l'évaluer. Mais les méthodes ABC sont beaucoup plus souples car elle ne reposent que sur notre capacité à (1) simuler suivant le modèle stochastique, et (2) capter l'information importante au travers de statistiques résumées.

Logiciels Les outils disponibles dans la littérature sont nombreux mais ne recouvrent pas toutes les questions posées en pratique. Citons quelques exemples représentatifs de logiciels de simulation et d'estimation.

- *Ms* de Hudson (2002) simule par coalescence des scénarios démographiques avec migration entre populations.

- *IBDSim* de Leblois et al. (2009) simule des scénarios en présence de migration (isolation par la distance) en temps discret, génération par génération, et échantillonne la dernière génération.

- *DIYABC* de Cornuet et al. (2008) simule des scénarios démographiques avec des événements de divergence et d'admixture mais sans présence de migration. Notons que *DIYABC*fournit une multitude de procédures d'estimation et de sélection de modèles par ABC.

- *Migraine* de Rousset and Leblois (2012) utilise l'échantillonnage préférentiel décrit ci-dessus pour estimer des surfaces de vraisemblance.

- *IM, IMa* de Nielsen and Wakeley (2001); Hey and Nielsen (2004, 2007) analysent des modèles démographiques d'isolation par migration. Ils incluent les deux approches bayésienne et fréquentielle. L'approximation de la loi *a posteriori* et de la surface de vraisemblance se fait par des méthodes MCMC.

References

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162:2025–2035.

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261.

Cornuet, J. M., Beaumont, M. A., Estoup, A., and Solignac, M. (2006). Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. *Theoretical Population Biology*, 69(2):129–144.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, 24(23):2713–2719.

Crow, J. F. and Kimura, M. (2009). *An Introduction to Population Genetics Theory*. The Blackburn Press.

De Iorio, M. and Griffiths, R. C. (2004a). Importance sampling on coalescent histories. I. *Advances in Applied Probability*, 36(2):417–433.

De Iorio, M. and Griffiths, R. C. (2004b). Importance sampling on coalescent histories. II:Subdivided population models. *Advances in Applied Probability*, 36(2):434–454.

De Iorio, M., Griffiths, R. C., Leblois, R., and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology*, 68(1):41–53.

Ethier, S. N. and Griffiths, R. C. (1987). The Infinitely-Many-Sites Model as a Measure-Valued Diffusion. *The Annals of Probability*, 15(2):515–545.

Griffiths, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-site model. *Journal of mathematical biology*, 27(6):667–680.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257.

Griffths, R. and Tavaré, S. (1994). Ancestral inference in population genetics. *Statistical Science*, 9:307–319.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, 167(2):747–760.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8):2785–2790.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications*, 13:235–248.

Kingman, J. F. C. (1982b). Exchangeability and the Evolution of Large Populations. In Koch, G. and Spizzichino, F., editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland, Amsterdam.

Kingman, J. F. C. (1982c). On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43.

Leblois, R., Estoup, A., and Rousset, F. (2009). IBDSim: A computer program to simulate genotype data under Isolation By Distance. *Molecular Ecology Ressources*, 9(1):107–109.

Majoram, P. and Donnelly, P. (1997). Human demography and the time since mitochondrial Eve. In Donnelly, P. and Tavaré, S., editors, *Progress in Population Genetics and Human Evolution*, volume 87. Springer.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2011). Approximate Bayesian computational methods. *Statistics and Computing*.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Nielsen, R. and Wakeley, J. (2001). Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach. *Genetics*, 158(2):885–896.

Rousset, F. and Leblois, R. (2012). Likelihood-Based Inferences under Isolation by Distance: Two-Dimensional Habitats and Confidence Intervals. *Molecular Biology and Evolution*, 29(3):957–973.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.

Stumpf, M. P. and McVean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nature reviews. Genetics*, 4(12):959–968.

Wakeley, J. (2005). *Coalescent Theory: An Introduction*. Roberts & Company Publishers.

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003). Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2):781–787.

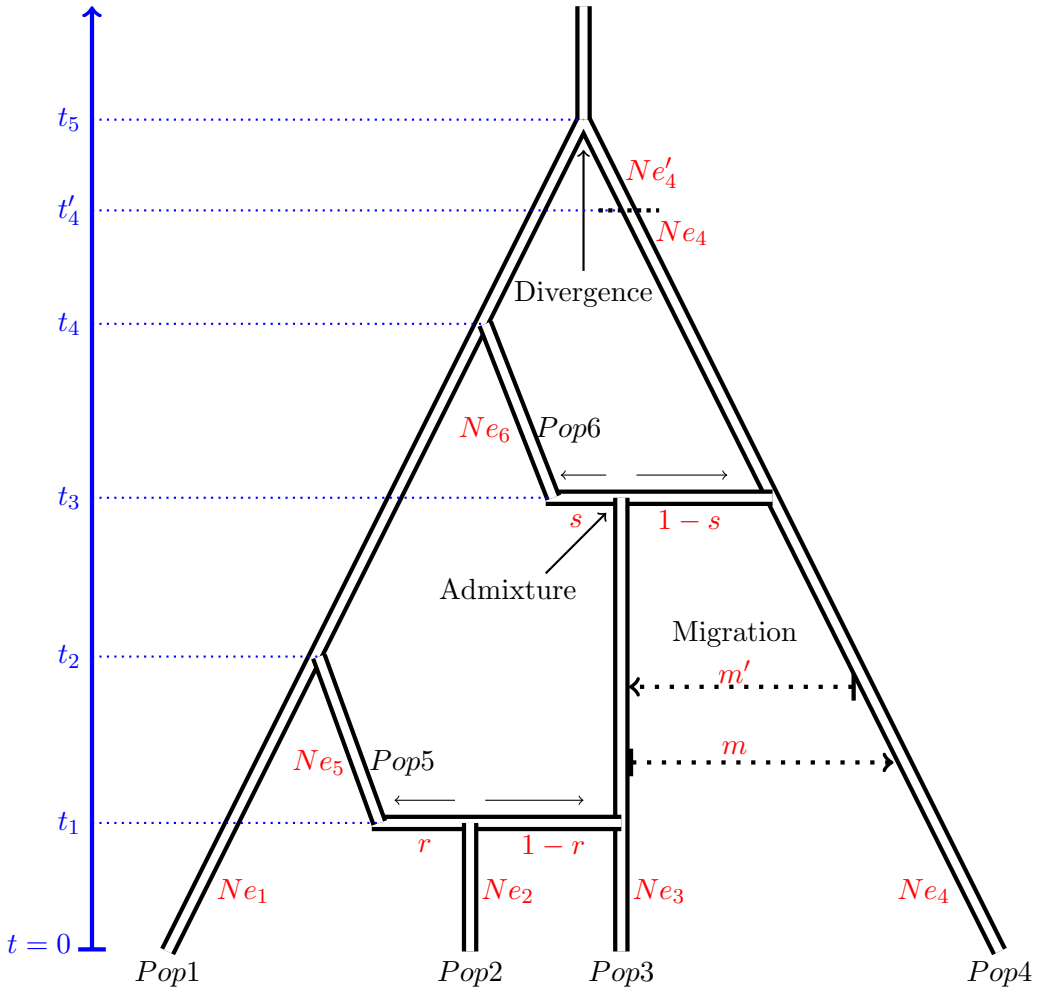

Figure 1: Exemple d'un scénario évolutif complexe composé d'évènements inter-populationnels. Ce scénario implique quatre populations échantillonnées $Pop1, \ldots, Pop4$ et deux autres populations non-observées $Pop5$ et $Pop6$. Les branches de ce schéma sont des "tubes" et le scénario démographique contraint la généalogie à rester à l'intérieur de ces "tubes". La migration entre les populations $Pop3$ et $Pop4$ sur la période $[0, t_3]$ est paramétrée par les taux de migration m et m'. Les deux évènements d'admixture sont paramétrés par les dates t_1 et t_3 ainsi que les taux d'admixture respectifs r et s. Les trois évènements restants sont des divergences, respectivement en t_2, t_4 et t_5. L'événement en t'_4 correspond à un changement de taille efficace dans la population $Pop4$.

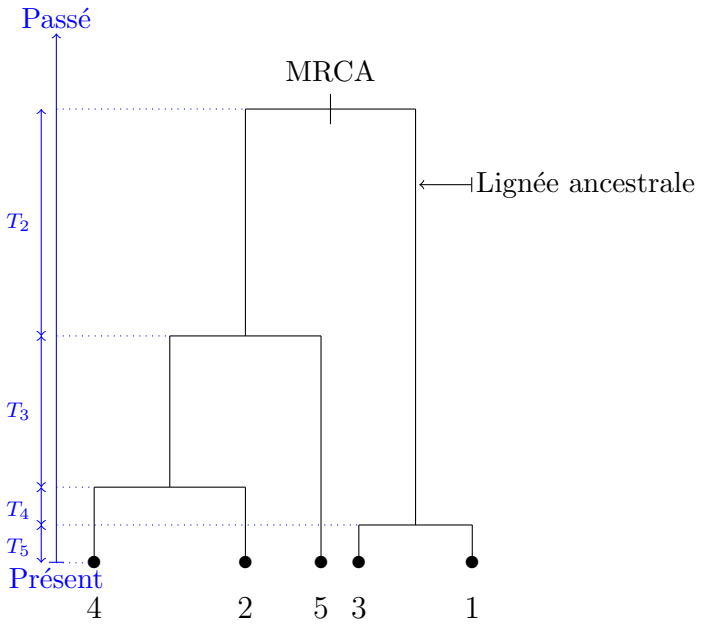

Figure 2: Exemple de généalogie de cinq individus issus d'une seule population fermée à l'équilibre. Les individus échantillonnés sont représentés par les feuilles du dendrogramme, les durées inter-coalescences $T_2, \ldots, T_5$ sont indépendantes, et T_k est de loi exponentielle de paramètre $k\big(k-1\big)/2$.

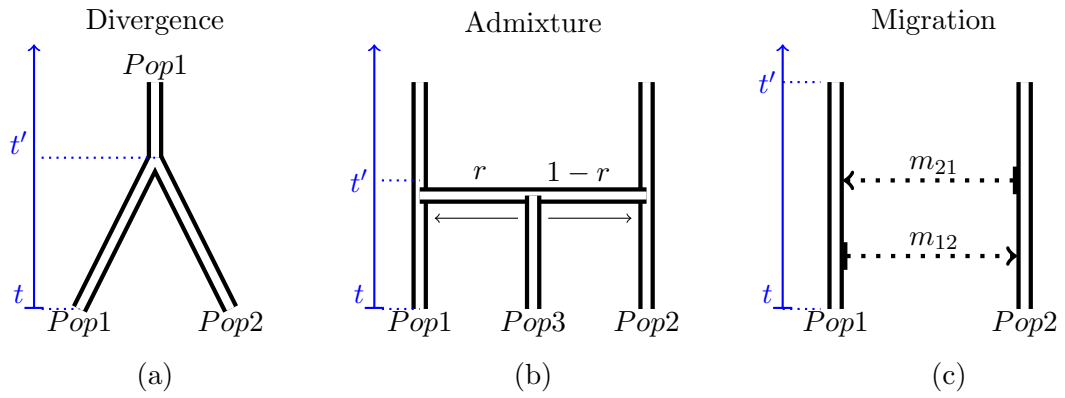

Figure 3: Représentations graphiques des trois types d'évènements inter-populationnels d'un scénario démographique. Il existe deux familles d'évènements inter-populationnels. La première famille est simple, elle correspond aux évènement inter-populationnels instantanés. C'est le cas d'une divergence ou d'une admixture. **(a)** Deux populations qui évoluent pour se fusionner dans le cas d'une divergence. **(b)** Trois populations qui évoluent en parallèle pour une admixture. Pour cette situation, chacun des tubes représente (on peut imaginer qu'il porte à l'intérieur) la généalogie de la population qui évolue indépendamment des autres suivant un coalescent de Kingman.
La deuxième correspond à la présence d'une migration.**(c)** Cette situation est légèrement plus compliquée que la précédente à cause des flux de gènes (plus d'indépendance). Ici, un seul processus évolutif gouverne les deux populations réunies. La présence de migrations entre les populations $Pop1$ et $Pop2$ implique des déplacements de lignées d'une population à l'autre et ainsi la concurrence entre les évènements de coalescence et de migration.

Algorithme 4 Généalogie de 2 populations avec migration

Entrées : Les tailles des deux échantillons ancestraux k_1 et k_2, les taux de migration m_{12} et m_{21} et les tailles efficaces des deux populations Ne_1 et Ne_2.

1) On choisit la population numéro i parmi les deux avec probabilité

$$\frac{k_i m_{ij} + \big[k_i\big(k_i-1\big)/2Ne_i\big]}{k_1 m_{12} + \big[k_1\big(k_1-1\big)/2Ne_1\big] + k_2 m_{21} + \big[k_2\big(k_2-1\big)/2Ne_2\big]}.$$

2) On choisit le type d'évènement : soit une coalescence avec probabilité

$$\frac{k_i\big(k_i-1\big)/2Ne_i}{k_i m_{ij} + \big[k_i\big(k_i-1\big)/2Ne_i\big]},$$

soit une migration avec probabilité

$$\frac{k_i m_{ij}}{k_i m_{ij} + \big[k_i\big(k_i-1\big)/2Ne_i\big]}.$$

3) Pour simuler une coalescence dans la population i:

- On simule T_c l'instant de coalscence de suivant une loi exponentielle de paramètre $k_i\big(k_i-1\big)/2Ne_i$.
- On augmente les lignées de la généalogie d'une longueur T_c.
- On tire les deux lignées à joindre uniformément parmi les k_i lignées de la population i et on applique la coalescence.
- $k_i \leftarrow k_i - 1$, et revenir en **1**.

4) Pour simuler une migration de la population i à la population j:

- On simule T_m l'instant de migration de suivant une loi exponentielle de paramètre $k_i m_{ij}/Ne_i$.
- On augmente les lignées de la généalogie d'une longueur T_m.
- On migre une lignée choisie uniformément dans la population i vers l'autre population $j \neq i$.
- $k_i \leftarrow k_i - 1$ et $k_j \leftarrow k_j + 1$, et revenir en **1**.

Algorithme 5 Généalogie contrainte par un scénario

Trier les événements inter-populationnels du plus récent au plus ancien.

Pour t allant de l'événement le plus récent au plus ancien **faire**

1) Simuler les généalogie intra-populationnel: un coalescent de Kingman indépendant par population jusqu'à t ou combiner le coalescent de Kingman avec migration dans le cas d'une migration.

2) Appliquer l'événement inter-populationnels instantané à la date t.

Fin pour

Simuler un (ou des) coalescent(s) de Kingman (avec migrations) sur la (les) dernière(s) population(s) jusqu'au MRCA.

Algorithme 6 Processus mutationnel le long d'une généalogie

Entrées : Une généalogie $\mathcal{G}$, un taux de mutation μ, et chaîne de Markov de matrice de transition Q et de la stationnaire ν.

1) Appliquer un processus ponctuel de Poisson sur les branches de $\mathcal{G}$.

2) Trier les événements de mutation sur $\mathcal{G}$ du plus ancien au plus récent.

3) Simuler suivant ν, le génotype du MRCA.

4) Parcourir $\mathcal{G}$ du MRCA jusqu'aux feuilles: construire les génotypes le long des branches.

- **si** événement de coalescence **alors** dupliquer le génotype en haut du noeud du dendrogramme.
- **si** événement de mutation **alors** appliquer un pas de la chaîne Q sur la branche (*i.e* l'individu qui porte la mutation).

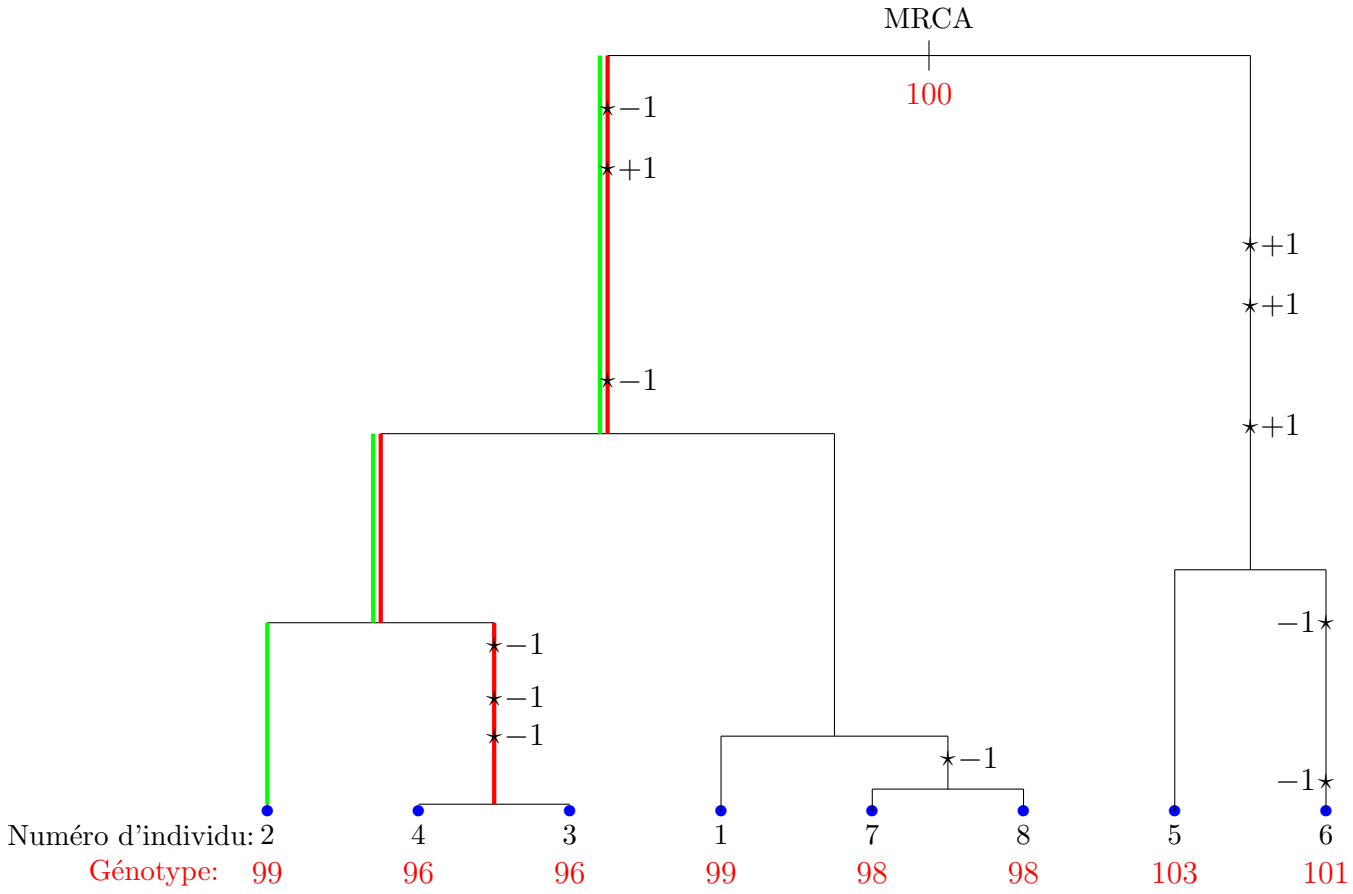

Figure 4: Exemple de simulation des génotypes d'un échantillon de huit individus en un locus microsatellite. Les mutations sont données par le modèle SMM. Les positions des mutations sont représentées par les étoiles ($\star$) sur les lignées du dendrogramme. Le génotype d'un individu (représenté par un nombre entier en rouge sur le dendrogramme) est obtenu en appliquant les mutations à partir du génotype du MRCA (100), ici, le long de la lignée de l'individu.