Statistique pour la bio-informatique

Séance 9 - Janvier 2004

Coalescence

1 Coalescence

1.1 Le coalescent de Kingman

Nous considérons l'évolution d'une population de taille finie constante égale à N, dont les générations ne se chevauchent pas. Pour simplifier, nous supposons que nous nous plaçons en un locus d'une espèce haploïde (la généralisation aux espèces diploïdes consiste à doubler la taille). Les générations sont construites en supposant que chaque descendant ($i = 1, \dots, N$) provient d'un parent pris dans $1, \dots, N$ (tirage avec remise). Tous les tirages sont indépendants.

Consultez la page suivante pour un simulateur du modèle de Wright-Fisher pour `windows`

`http ://www.handsongenetics.com/fisher.html`

Supposons que i individus sont de type A. En l'absence de mutation, le nombre de descendants de type A suivra la loi binomiale $\mathcal{B}(N, i/N)$. Nous cherchons à retracer l'historique d'un échantillon de n individus pris dans une telle population en remontant le temps. La remarque importante et que l'on peut raisonner comme si les descendants tiraient leurs parents au hasard et indépendamment de générations en générations. En identifiant l'ancêtre commun à deux individus le plus récent dans l'échantillon, puis à trois, etc on met en évidence une structure d'arbre.

Nous cherchons à caractériser la structure de cet arbre ancestral. Puisque les individus choisissent leurs parents au hasard, nous avons

$$\mathrm{P}(2 \text{ individus ont le meme parent }) = \frac{1}{N}$$

Soit T_2 le nombre de générations nécessaires pour observer la coalescence des lignées dont sont issues deux individus, c'est à dire le nombre de générations qu'il faut attendre pour que 2 individus aient un ancêtre commun. Nous avons

$$\mathrm{P}(T_2 \geq k) = (1 - 1/N)^k$$

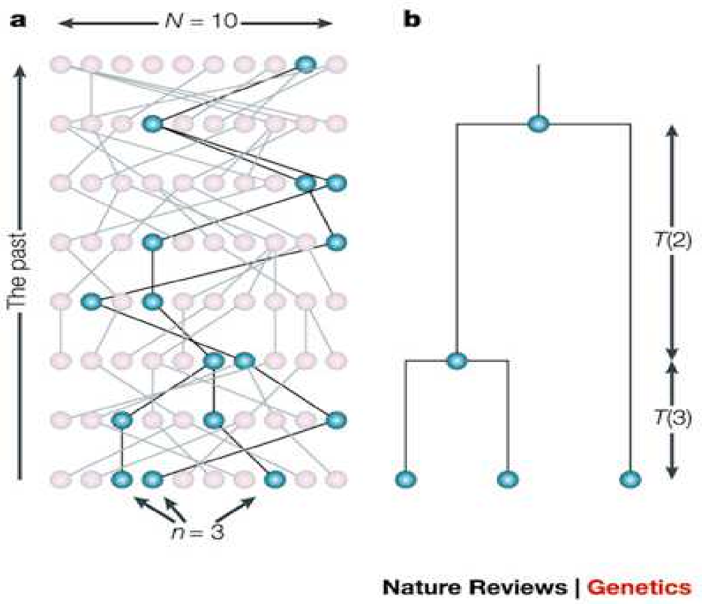

Fig. 1 – *Arbre de coalescence de $n = 3$ individus dans une population de $N = 10$ individus pour le modèle de Wright-Fisher et pour le modèle du coalescent [d'après Rosenberg et Nordborg (Nature, 2002)]*

En prenant pour unité de temps la taille de la population

$$k = Nt,$$

nous obtenons, pour $N \to \infty$,

$$(1 - 1/N)^{Nt} \to e^{-t}$$

Ainsi selon une échelle convenable, le temps qu'il faut remonter pour voir apparaître l'ancêtre commun de deux gènes est approximativement tiré selon une loi exponentielle de paramètre 1. De manière analogue, nous pouvons calculer

$$\mathrm{P}(\text{3 individus ont des parents distincts}) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right).$$

La probabilité pour qu'un échantillon de trois individus ait trois parents distincts pendant plus de k générations est alors

$$\left[(1 - \frac{1}{N})(1 - \frac{2}{N})\right]^k = \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^k$$

En changeant d'échelle de temps comme précédemment, le temps pour une coalescence vérifie

$$\mathrm{P}(T_3 \geq t) \to e^{-3t}, \quad N \to \infty.$$

Plus généralement, le nombre de parents distincts d'un échantillon de k individus peut être vu comme le nombre de cellules occupées lorsque l'on jette k boules dans N cellules

$$p_{kj} = \mathrm{P}(k \text{ individus ont } j \text{ parents distincts}) = \frac{N(N-1)\dots(N-j+1)}{N^k}\mathcal{S}_k^{(j)}$$

où $\mathcal{S}_k^{(j)}$ est le nombre de partitions d'un ensemble à k éléments en j sous-ensembles. Ce nombre est appelé *nombre de Stirling de seconde espèce.*

Le processus ancestral à taille finie est difficile à étudier. Nous allons l'étudier pour N grand et en prenant pour unité de temps N générations. Soit $A_n^N(t)$ *le nombre d'ancêtres distincts à la génération* t *d'un échantillon de taille* n. Ce processus est une chaîne de Markov sur l'ensemble $\{1, 2, \dots, n\}$ associée aux probabilités de transitions

$$\mathrm{P}(A_n^N(t+1) = j \ / \ A_n^N(t) = k) = p_{kj}$$

Pour n fixé, nous avons

$$p_{k,k-1} = C_2^k \frac{1}{N} + o(\frac{1}{N})$$

et

$$p_{k,k} = 1 - C_2^k \frac{1}{N} + o(\frac{1}{N})$$

Ainsi, lorsque N tend vers l'infini, les sauts privilégiés sont d'amplitude 1 au plus. En considérant la normalisation mentionnée ci-dessus, les durées séparant les sauts sont de loi exponentielle de paramètres C_n^2, C_{n-1}^2, ..., $C_3^2 = 3$, 1. Nous venons de donner l'idée de la démonstration du résultat suivant.

Théoreme 1.1 *(Kingman, 1982) Le processus ancestral* $A_n^N(t)$ *associé au modèle de Wright-Fisher converge vers le processus* A_n

$$A_n^N(\lfloor N.\rfloor) \to A_n(.), \quad N \to \infty$$

(convergence des lois fini-dimensionnelles) défini comme le processus de mort pure sur $n, \dots, 2$, *de condition initiale* $A_n(0) = n$ *et de taux*

$$\lambda_k = C_k^2$$

Donner la démonstration de Tavaré (2001).

Exercice. On suppose $\#S = n \geq 3$. Dénombrer les arbres de coalescence (solution : $n!(n-1)!/2^{n-1}$). Remarquer le différence avec le nombre d'arbre enracinés (les nœuds internes sont numérotés).

1.2 Temps de coalescence d'un échantillon fini

Le temps écoulé depuis l'ancêtre commun le plus récent à n gènes peut être étudié, voire estimé lorsque l'on dispose de données moléculaires, grâce au modèle de la coalescence

$$T_{\mathrm{ACPR}} = T_n + \ldots + T_2.$$

Dans cette formule, les T_k sont indépendantes de loi exponentielle de paramètres respectifs $k(k-1)/2$. Nous obtenons (exercice) que le temps moyen écoulé depuis l'ancêtre commun le plus récent est égal à

$$E[T_{\mathrm{ACPR}}] = 2\left(1 - \frac{1}{n}\right) \to 2.$$

Une remarque évidente ($E[T_2] = 1$) montre que ce temps est pratiquement le double de celui nécessaire pour trouver l'ancêtre commun de 2 individus.

En utilisant l'indépendance, nous obtenons aussi (exercice) la variance de cette variable aléatoire

$$Var[T_{\mathrm{ACPR}}] = \sum_{j=2}^{n} Var[T_j] \to \frac{8\pi^2}{6} - 12 \approx 1.16$$

À nouveau, T_2 est responsable de la part la plus importante de la variance.

1.3 Mutations

Examinons l'effet des mutations dans le modèle de Wright-Fisher. Supposons que la probabilité de mutation par gène et par génération est μ. Le nombre de mutations le long d'une branche de la généalogie suit la loi binomiale de moyenne

$$\text{longueur de la branche} \times \mu$$

où la longueur de la branche correspond au nombre de générations le long de cette branche. Dans l'approximation du coalescent, nous supposons que

$$2N\mu \to \theta.$$

Les mutations sont alors réparties sur chaque branche selon un processus de Poisson de paramètre $\theta/2$.

Laissons de coté le modèle à 2 allèles. Nous supposons aussi que *la mutation d'un gène produit toujours un nouveau gène.* Ce modèle, proposé par Kimura et Crow, s'appelle IAM : *infinite alleles model.* L'argument en faveur de cette hypothèse réaliste est simple. Si un gène comporte 500 nucléotides, le nombre de séquences possibles est

$$4^{500} = 10 \text{ avec } 300 \text{ zeros.}$$

Pour chaque séquence, $3 \times 500 = 1500$ séquences peuvent être atteinte par le changement d'un nucléotide, et la probabilité de revenir sur une séquence en 2 mutations est donc très faible (0.00066) .

Exercice. Ecrire un algorithme de simulation du coalescent avec mutations IAM.

1.4 Distribution des allèles dans un échantillon

Pour le modèle de mutation IAM, nous donnons la distribution de la configuration des allèles dans un échantillon de n individus. Nous appelons *configuration* le vecteur

$$c = (c_1, \ldots, c_n)$$

où

$$c_i = \text{nombre d'alleles presents } i \text{ fois}$$

et nous notons $|c| = c_1 + 2c_2 + \ldots + nc_n \equiv n$.

Exemple. Dès les années 60, il a été possible par la technique de l'électrophorèse de quantifier le polymorphisme à un locus donné dans une population (on parle alors de *marqueurs génétiques*). Pour des marqueurs de type allozyme chez la drosophile *D. simulans* au locus de l'Esterase C, on a alors trouvé pour un échantillon de $n = 308$ individus, 7 allèles distincts dont les comptages sont représentés ci-dessous

$$91 + 76 + 70 + 57 + 12 + 1 + 1 = 308$$

Nous avons observé la configuration c telle que $c_1 = 2$, $c_{12} = 1$, $c_{57} = 1$, $c_{70} = 1$, $c_{76} = 1$, $c_{91} = 1$ et tous les autres c_i sont nuls ($c_i = 0$ pour $i \neq 91, 76, 70, 57, 12, 1$).

Théoreme 1.2 *(Ewens, 1972) Pour le modèle de mutation IAM et un échantillon de taille n,*

$$\mathrm{P}(C_1 = c_1, \ldots, C_n = c_n) = \mathrm{I}_{|c|=n} \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}$$

où $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

La démonstration repose sur la structure du coalescent. Notons e_i le vecteur unité ayant un 1 en i

$$e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

et pour $n = |c| > 1$,

$$q(c) = \mathrm{P}(C_1 = c_1, \ldots, C_n = c_n)$$

avec la convention $q(e_1) = 1$. En regardant la généalogie de l'échantillon (dans le sens inverse du temps), l'un des deux événements suivants se réalise

- soit on rencontre une mutation,
- soit on peut remonter à l'ancêtre commun de deux individus parmi les n.

En utilisant un résultat portant sur le minimum de deux variables de loi exponentielle indépendantes de paramètres respectifs θn et C_n^2, nous trouvons que les probabilités de ces événements sont

$$\frac{\theta}{\theta+n-1} \quad , \quad \frac{n-1}{\theta+n-1}$$

Supposons que le premier événement se réalise (mutation) et que la configuration trouvée est b. On est alors passé de la configuration b à c aux feuilles. Voici comment se caractérise b

- Avec la probabilité c_1/n, la mutation se produit sur l'un des c_1 singletons : $b = c$
- Avec la probabilité $2(c_2+1)/n$, la mutation se produit pour la classe c_2 : $b = c - 2e_1 + e_2$,
- Avec la probabilité $j(c_j+1)/n$, la mutation se produit pour c_j : $b = c - e_1 - e_{j-1} + e_j$

Supposons que le second événement se réalise (coalescence). On a alors la configuration

$$b = c + e_j - e_{j+1}$$

un individu dans l'une des c_j+1 classes d'allèles représentés j fois a un descendant. Cet événement réduit le nombre de classes de taille j à c_j et accroit le nombre de classes de taille $(j+1)$ à c_{j+1}. La probabilité est $j(c_j+1)/(n-1)$.

Finalement, par la formule des probabilités totales, on obtient une formule de récurrence

$$\begin{aligned} q(c) &= \frac{\theta}{\theta+n-1}\left[\frac{c_1}{n}q(c)+\sum_{j=2}^{n}\frac{j(c_j+1)}{n}q(c-e_1-e_{j-1}+e_j)\right] \\ &\quad +\frac{n-1}{\theta+n-1}\left[\sum_{j=1}^{n-1}\frac{j(c_j+1)}{n-1}q(c+e_j-e_{j+1})\right]. \end{aligned}$$

On termine la vérification du résultat par récurrence sur n et $k_n = \sum_{i=1}^{n} c_i$ (exercice délicat). $\square$

Exercice. Ecrire un algorithme de simulation d'une configuration c tirée selon la loi d'Ewens. Voir dans le chapitre 1 de Durrett, la simulation du coalescent à l'aide de la méthode du *restaurant chinois* de Aldous.

Application de la formule d'Ewens Le génôme de l'être humain diffère peu d'un individu à l'autre. Il y pourtant des sites de l'ADN plus variables que d'autres et ces sites sont très utiles pour les études généalogiques. On appelle *SNP* (Single Nucleotide Polymorphism) un site qui est variable dans une population (tel site s'appelle site de ségrégation) et tel que la fréquence de la base la plus rare est supérieure à un pour cent. En un SNP, plus d'un pour cent de la population possède donc une différence génétique avec le reste de la population. Nous supposons que cette différence est le résultat d'une mutation unique.

La formule de Ewens permet d'estimer en particulier le nombre de SNP pour une population de taille n. En effet, considérons un échantillon de n individus dont m portent une mutation. Par convention, on dira que la base la plus fréquente est ancestrale et donc $m < n/2$. La probabilité de cette configuration est

$$\mathrm{P}(c_m = 1\,;\, c_{n-m} = 1) = \frac{n!}{\theta_{(n)}}\,\frac{\theta^2}{m(n-m)}.$$

Après simplification et en posant

$$x_m = m/n,$$

nous avons

$$\mathrm{P}(c_m = 1\,;\, c_{n-m} = 1) = \frac{\theta}{n x_m (1-x_m)}.$$

L'événement considéré est

$$A = X_m \geq 0.01$$

et nous avons

$$\mathrm{P}(A) = \sum_{0.01 \leq x_m \leq 0.5} \frac{\theta}{n x_m (1 - x_m)}.$$

Lorsque n tend vers l'infini, nous obtenons l'intégrale de Riemann

$$\mathrm{P}(A) \to \int_{0.01}^{0.5} \frac{\theta}{x(1-x)} dx \approx 1/300.$$

Pour la population humaine, les biologistes admettent la valeur $\theta \approx 1/1300$ permettant en effet de calculer l'intégrale ci-dessus.

1.5 Nombre d'allèles

La variable aléatoire

$$K_n = C_1 + \ldots + C_n$$

représente le nombre d'allèles distincts dans l'échantillon de taille n.

Théoreme 1.3 *Nous avons*

$$\forall k = 1, \ldots, n, \quad \mathrm{P}(K_n = k) = \frac{\theta^k |S_n^k|}{\theta_{(n)}}$$

où $|S_n^k|$ est le coefficient de Stirling de première espèce,

$$|S_n^k| = \text{ coeff. de } x^k \text{ dans } x(x+1) \cdots (x+n-1)$$

Commentaire. Les coefficients de Stirling de première espèce sont bien connus en combinatoire. Ils représentent le nombre de permutations de n éléments ayant exactement k cycles et peuvent être calculés par récurrence de la manière suivante

$$|S_n^k| = (n-1)|S_{n-1}^k| + |S_{n-1}^{k-1}|.$$

Démonstration. Exercice. □

Une description alternative utile de la loi de K_n est donnée (exercice) par la fonction génératrice

$$E[z^{K_n}] = \prod_{j=1}^{n} E[z^{X_j}]$$

où les X_j sont des variables de Bernoulli de paramètre

$$\mathrm{P}(X_j = 1) = \frac{\theta}{\theta + j - 1}$$

Grâce à cette représentation, nous avons

$$E[K_n] = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j} \sim \theta \log n$$

et

$$Var[K_n] = \sum_{j=0}^{n-1} \frac{\theta j}{(\theta + j)^2} \sim \theta \log n$$

De plus un TCL est valide

$$\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} \to \mathcal{N}(0, 1).$$

Exercice. Déterminer la loi conditionnelle de la configuration C sachant $K_n = k$. Dépend-elle de θ ? En déduire que l'estimateur de maximum de vraisemblance de θ est solution de

$$k = \sum c_i = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}$$

Solution abrégée. La loi conditionnelle est

$$\mathrm{P}_\theta(C \mid K = k) = \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \frac{1}{c_j!} \left(\frac{1}{j}\right)^{c_j}.$$

Outre l'expression de cette loi, le fait remarquable est qu'elle ne dépend plus de θ. On dit que K est une statistique exhaustive pour θ. Cela signifie que toute l'information nécessaire pour estimer le paramètre statistique θ à partir de l'observation d'une condiguration c se trouve dans $k = \sum_{c_j}$.

Ce résultat permet de calculer l'estimateur de maximum de vraisemblance de θ. Posons

$$\ell(\theta) = \log \mathrm{P}_\theta(c) = \log \mathrm{P}_\theta(c| \ K = k) + \log \mathrm{P}_\theta(K = k).$$

Lorsque nous avons un maximum, alors

$$\ell'(\theta) = 0 = \frac{\partial}{\partial \theta} \log \mathrm{P}_\theta(K = k).$$

Cette équation est équivalente à

$$\frac{\partial}{\partial \theta}(k \log \theta - \log \theta_{(n)}) = 0$$

et on trouve que l'estimateur de maximum de vraisemblance $\hat{\theta}$ est solution de

$$k = \sum c_i = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}.$$

Dans la théorie du maximum de vraisemblance, $\hat{\theta}$ est asymptotiquement sans biais

$$E[\hat{\theta}] \to \theta,$$

asymptotiquement gaussien, de variance minimale tendant vers la borne Cramer-Rao

$$Var[\hat{\theta}] \sim \frac{1}{\mathcal{I}(\theta)}$$

où $\mathcal{I}(\theta)$ est l'information de Fisher

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right] = E\left[\left(\frac{\partial \ell(\theta)}{\partial \theta^2}\right)\right] \sim \log n/\theta.$$

La convergence est donc nécessairement très lente.

1.6 Eve mitochondriale

L'ADN mitochondrial est hérité de la mère uniquement. Il est donc sujet à d'éventuelles mutations mais pas aux recombinaisons (hypothèse récemment critiquée dans la littérature spécialisée). De plus, les substitutions sont relativement fréquentes. Cela en fait une donnée de choix pour les études généalogiques.

Les études fondées sur l'archéologie, les fossiles et les cultures, montrent que l'Homo Sapiens est une espèce sortie d'Afrique il y a environ 100000 ans. L'espéce serait dispersé vers l'Asie, l'Europe et plus récemmement (15000 ans tout de même) vers les amériques. Cette donnée permet d'exprimer une relation du type

$$100000 \approx NE[T_{\text{ACPR}}].$$

Ainsi, une valeur raisonnable pour N est de l'ordre de 2500 car une génération à une durée de 20 ans. Cela peut paraître paradoxal que N soit plus petit que la taille de l'humanité. Cette grandeur est en fait appelée la *taille efficace* et ne serait interprétable comme une véritable taille que si la population se comportait de manière idéale comme dans le modèle de Wright-Fisher. Évidemment, ce n'est pas le cas de la population humaine.

À partir d'un échantillon d'ADN mitochondrial humain, on peut estimer directement $\theta = N\mu$. À ce sujet, La littérature biologique utilise encore souvent l'estimateur Δ_n plutôt que le nombre de sites de ségrégations. Le taux de mutation (par site et par génération) de la séquence d'ADN étudiée est de l'ordre de 10e-7 par génération. Ainsi on obtient $\theta = 1/1300$ et

$$N \approx 3000$$

qui est du même ordre que le chiffre précédent.

1.7 Estimation de θ à l'aide de l'homozygotie

Une démarche classique en génétique est d'estimer le taux de mutation en se servant de la statistique d'homozygotie

$$H_n = \mathrm{P}(\text{deux genes pris au hasard sont identiques}) = \sum_{j=1}^{k} \frac{n_j(n_j-1)}{n(n-1)}$$

où k est le nombre d'allèle et n_j l'effectif de l'allèle j. La formule d'Ewens montre que

$$E[H_n] = \mathrm{P}(c_1 = 0; c_2 = 1) = \frac{1}{1+\theta}$$

La variance est plus difficile à calculer. Steward (1976) montre que

$$Var[H_n] \to \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}$$

Cette statistique est à la base du test de neutralité de Watterson (1978).

Exercice. Déterminer la valeur de la statistique pour le jeu de données *D. Simulans*. Par simulations, calculer la p-value associée à l'hypothèse de neutralité : H_0 : Les données proviennent d'un modèle de coalescent avec mutations IAM.

Exercice. (Autre démonstration du résultat précédent). En utilisant le temps de coalescence dans un échantillon de taille $n = 2$, justifier la relation

$$E[H_n] = \mathrm{P}(N_{T_2} = 0) = \frac{1}{1+\theta}$$

où (N_t) est un processus de Poisson de paramètre $2\theta/2 = \theta$.

Exercice. Estimer θ pour le jeu de données des indiens de la tribu Nuu Chah Nulth (*cf* Tableau **??**).

1.8 Analyse des séquences d'ADN

1.8.1 Modèle à infinité de sites

Dans le modèle à nombre infini d'allèles, seul le fait que deux allèles sont différents est pris en compte. Le modèle ne se place pas au niveau de la séquence d'ADN et délaisse ainsi une partie intéressante de l'information moléculaire. Dans le modèle à infinité de sites, nous considérons des séquences arbitrairement longues, et faisons l'hypothèse que *les mutations se produisent toujours en des sites distincts* sur la séquence.

Considérons à nouveau deux séquences alignées de longueurs égales à m et notons Δ_2 le nombre de substitutions. Dans l'exemple ci-dessous,

$$\begin{array}{c} aatcgcttgatacc \\ actcgcctgataac \end{array}$$

nous avons $\Delta_2 = 3$. Notons encore μ le taux de mutation par site et $\theta \approx 2N\mu$. La probabilité que la coalescence de deux sites se produise avant une mutation est $p = 1/(1+\theta)$ et

$$\forall k = 0, 1, 2, \ldots, \quad \mathrm{P}(\Delta_2 = k) = \left(\frac{\theta}{1+\theta}\right)^k \frac{1}{1+\theta}.$$

La variable aléatoire Δ_2 suit donc la loi de Pascal de paramètre p, d'espérance et de variance

$$E[\Delta_2] = \theta \quad , \quad Var[\Delta_2] = \theta + \theta^2.$$

Etant donné un échantillon de n séquences, nous pouvons estimer θ par le nombre moyen de substitutions entre séquences prises deux à deux

$$\hat{\theta} = \Delta_n = \frac{2}{n(n-1)} \sum_{i<j} \Delta_{ij}.$$

Dans cette définition, nous avons noté Δ_{ij} le nombre de substitutions entre la séquence i et la séquence j. Nous avons de plus,

$$E[\Delta_{ij}] = E[\Delta_2] = \theta$$

et Δ_n est un estimateur sans biais

$$E[\Delta_n] = \theta$$

La variance peut être calculée, elle est égale à

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

Exercice. Estimer θ pour le jeu de données des indiens de la tribu Nuu Chah Nulth. Comparer à l'estimation dans le cadre du modèle IAM.

Exercice. (Autre démonstration du résultat précédent). En utilisant le temps de coalescence dans un échantillon de taille $n = 2$, justifier la relation

$$E[\Delta_2] = \theta E[T_2] = \theta.$$

1.8.2 Nombre de sites de ségrégation

Dans le modèle à infinité de sites, on appelle *site de ségrégation* pour n séquences tout site pour lequel un nucléotide diffère des autres pour une séquence au moins. On note S_n le nombre de tels sites. Clairement, le nombre de sites de ségrégation est exactement égale au nombre de mutations (ou plutôt de substitutions) depuis la séquence ancestrale. Dans l'exemple suivant, nous observons les séquences

$$\begin{array}{c} aaga \\ agaa \\ ggaa \end{array}$$

Il y a $S_3 = 3$ sites des ségrégation. Notons par ailleurs que $\Delta_3 = 6/3 = 2$.

Théoreme 1.4 *Soit h_{n-1} le nombre harmonique d'ordre $n-1$*

$$h_{n-1} = \sum_{j=1}^{n} \frac{1}{j}.$$

Alors, S_n/h_{n-1} est un estimateur sans biais de θ, i.e.,

$$E[S_n] = h_{n-1}\theta.$$

De plus, nous avons

$$Var[S_n/h_{n-1}] \sim \frac{\theta}{\log n} \to 0.$$

Démonstration. Conditionnellement à $L_n = \ell$ la longueur de l'arbre de coalescence, la variable S_n suit la loi de Poisson de paramètre $\theta\ell/2$. Nous avons donc

$$E[S_n] = \frac{\theta}{2} E[L_n].$$

Puisque

$$L_n = \sum_{j=2}^{n} jT_j,$$

nous avons

$$E[S_n] = \frac{\theta}{2}\sum_{j=2}^{n} jE[T_j] = \frac{\theta}{2}\sum_{j=2}^{n} \frac{2j}{(j-1)j}.$$

Pour la variance, le raisonnement est similaire. Nous avons

$$Var[S_n] =?E[Var[S_n \mid L_n]] + Var[E[S_n \mid L_n]] = \theta h_{n-1} + \frac{\theta^4}{4}Var[L_n]$$

Commentaires. On peut montrer que S_n est l'estimateur de maximum de vraisemblance, et donc asymptotiquement optimal. Des simulations montrent que S_n/h_{n-1} est un meilleur estimateur que Δ_n même pour des tailles d'échantillons petites.

Exercice. On considère les séquences

$$\begin{array}{c} aatcgaatcg \\ aaccgaacct \\ catcgaccct \end{array}$$

Ce jeu de données est-il compatible avec le modèle à infinité de sites. Calculer le nombre de sites de ségrégation et construire un arbre généalogique compatible avec ce nombre de mutation.

1.8.3 Problème

On considère la variable aléatoire K égale au nombre d'allèles distincts dans un échantillon de taille n pour le modèle à infinité d'allèles IAM de taux de mutation $\theta > 0$.

1) Montrer que la fonction génératrice

$$\forall |z| \leq 1, \quad G_K(z) = E[z^K]$$

est égale à

$$G_K(z) = \prod_{k=1}^{n} \left(\frac{k-1}{\theta + k - 1} + \frac{\theta z}{\theta + k - 1} \right).$$

2) En déduire que la loi de K est identique à la loi de la somme

$$S_n = X_1 + \ldots + X_n$$

où les X_i sont des variables aléatoires de Bernoulli indépendantes (à valeurs dans $\{0,1\}$) dont on précisera les paramètres.

3) Démontrer que

$$E_\theta[K] = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}.$$

4) Démontrer l'équivalent suivant

$$Var[K] = \sum_{j=0}^{n-1} \frac{\theta j}{(\theta + j)^2} \sim \theta \log n \quad (n \to \infty)$$

5) On observe k allèles distincts dans un échantillon. On note $\hat{\theta}$ l'estimateur des moments, solution de l'équation

$$k = E_{\hat{\theta}}[K]$$

Justifier le fait que $\hat{\theta}$ est asymptotiquement sans biais.

6) On note

$$\ell(\theta; k) = \mathrm{P}_\theta(K = k).$$

Montrer que l'information de Fisher

$$\mathcal{I}(\theta) \equiv E_\theta \left[\left(\frac{\partial}{\partial \theta} \log \ell(\theta; K) \right)^2 \right]$$

est égale à

$$\mathcal{I}(\theta) = \frac{Var[K]}{\theta^2}$$

En déduire la variance asymptotique de l'estimateur des moments.

1.9 Estimer le taux de mutation par MCMC

Estimer le taux de mutation de séquence d'ADN est un sujet devenu récemment très populaire. Dans cette démarche, la généalogie n'est pas observée et le modèle de la coalescence sert de modèle a priori.

Etant donné un échantillon de données moléculaires (séquences d'ADN mitochondrial par exemple),

$$D = s_1, \ldots, s_n$$

on souhaite estimer le paramètre de mutation

$$\theta = \mu N_e,$$

où N_e est la taille efficace de la population, c'est à dire la taille que devrait avoir la population si elle se comportait de manière idéale comme dans le modèle de Wright-Fisher. L'approche adoptée est le calcul de la vraisemblance

$$L(\theta) = p(D/\theta).$$

Toutefois, ce calcul n'est possible que lorsque l'on connait l'arbre généalogique des données T. La grandeur évaluée est alors

$$p(D/T\,;\theta)$$

et le calcul est effectué grâce à un algorithme explicite de programmation dynamique (algorithme de peeling) le long de l'arbre. Cet algorithme est strictement identique à celui utilisé dans le cadre des calculs de vraisemblance pour les arbres phylogénétiques.

La difficulté est que la phylogénie est inconnue, il s'agit d'une donnée manquante. Nous devons intégrer sur tous les arbres imaginables, en nous servant d'un modèle raisonnable. Le modèle de référence est la structure de coalescence

$$p(T) = p(t_2, \ldots, t_n) = \prod_{k=2}^{n} \exp(-\lambda_k t_k).$$

où $\lambda_k = k(k-1)/2$ et les t_k sont les durées de loi exponentielle séparant les événements de coalescence. On doit alors intégrer suivant toutes les généalogies pondérées par leur probabilités d'occurrence dans le modèle de coalescence

$$L(\theta) = \int p(D/T\,;\theta)p(T)dT.$$

Le calcul de cette intégrale est irréalisable pour de grandes valeurs de n et la pratique est d'utiliser une méthode de Monte Carlo par chaîne de Markov. Une application directe de la méthode de Monte Carlo au calcul de la vraisemblance pourrait être justifiée (il est facile de simuler des arbres selon la loi du coalescent). Malheureusement, cette méthode est très inefficace en pratique car les arbres tirés ont en général une valeur $p(D/T)$ négligeable. La méthode utilisée est plus subtile et repose sur l'échantillonnage d'importance (Importance Sampling). Pour commencer, une remarque utile est que le changement de variables

$$t' = \theta t$$

permet de réécrire l'intégrale de la manière suivante

$$L(\theta) = \int p(D/T)p(T/\theta)dT.$$

où $p(D/T)$ ne dépend plus que de la structure de l'arbre et

$$p(T/\theta) = \frac{1}{\theta^{n-1}} \exp(-\sum_{k=1}^{n-1} \lambda_k \frac{t_k}{\theta}).$$

Notons que pour toute distribution $q(T)$, nous avons

$$L(\theta) = \int p(D/T) \frac{p(T/\theta)}{q(T)} q(T) dT.$$

Des simulations selon $q(T)$ permettent alors d'estimer

$$L(\theta) \approx \frac{1}{M} \sum_i p(D/T^i) \frac{p(T^i/\theta)}{q(T^i)}.$$

Le choix pertinent pour $q(T)$ est donné par la formule de Bayes (exercice)

$$q(T) = p(T/D).$$

Bien que nous soyons capable de simuler la loi $p(T/D)$, il est impossible de connaître l'expression analytique de cette loi. On recourt alors à des calculs de vraisemblance relative

$$\frac{L(\theta)}{L(\theta_0)} \approx \frac{1}{M} \sum_i \frac{p(T^i/\theta)}{p(T^i/\theta_0)}$$

où les T_i sont tirés selon la loi

$$q(T) = p(T/D, \theta_0)$$

dont l'expression est proportionnelle à $p(D/T)p(T/\theta_0)$.

Les tirages s'effectuent selon la méthode de Monte Carlo par chaîne de Markov. Pour implanter la méthode dite de *Metropolis-Hasting*, la première étape est de trouver un *noyau d'exploration* $q(T, T')$ tel que

$$p(T/\theta_0)q(T, T') = p(T'/\theta_0)q(T', T).$$

Il y a de nombreux choix possibles décrits un peu plus loin. En un pas de l'algorithme de Metropolis-Hasting, nous générons un nouvel arbre T' étant donné l'arbre courant T selon les probabilités

$$\mathrm{prob(T, T')} = \begin{cases} q(T, T') & \text{if } \frac{p(D/T')}{p(D/T)} > 1 \\ q(T, T')p(D/T')/p(D/T) & \text{sinon.} \end{cases}$$

Cela revient à proposer un nouvel arbre selon le noyau d'exploration $q(T, T')$ puis appliquer une règle d'acceptation/rejet favorisant les arbres qui augmentent la probabilité $p(D/T)$. La chaîne de Markov résultant de cette algorithme est réversible

$$p(T/D)\text{prob}(T, T') = p(T'/D)\text{prob}(T', T).$$

Modulo des hypothèses faibles sur $q(T, T')$, l'algorithme peut fournir un ensemble de variables faiblement corrélées et tirées selon la loi désirée. L'implantation se fait comme suit

1. *Tirer T' selon la loi $q(T, T')$;*
2. *Calculer $p(D/T')$ grâce à un algorithme de "peeling"*
3. *Si $p(D/T')/p(D/T) > 1$ accepter l'arbre T'*
4. *Sinon, accepter T' avec la probabilité $p(D/T)/p(D/T')$, ou rester en T avec la probabilité $1 - p(D/T)/p(D/T')$.*
5. *Aller en 1.*

Le problème délicat est la construction du noyau d'exploration. Une méthode est d'utiliser l'échantillonneur de Gibbs consistant à utiliser la loi conditionnelle d'une variable sachant toutes les autres. Pour simuler un couple de loi $p(x, y)$ par exemple, l'échantillonneur de Gibbs part de x puis simule y selon $p(y/x)$ puis x selon $p(x/y)$, etc (exercice : vérifier que la chaîne produite est réversible).

Dans le cas des arbres, une stratégie consiste à supprimer un noeud de l'arbre, simuler le temps de coalescence correspondant à ce noeud conditionnellement aux autres temps qui restent inchangés, et enfin rebrancher le noeud au hasard avec son nouveau temps.

Exercice. Notons $Y_k = \sum_{j=k}^{n} T_j$ l'instant correspondant à la coalescence $n - k + 1$ (voir dessin). Déterminer la loi conditionnelle de Y_k sachant $Y_j = y_j$, $j \neq k$. Montrer quelle est identique à la loi conditionnelle d'une variable aléatoire $Y \sim \mathcal{E}(\lambda_k - \lambda_{k-1})$ sachant $Y \in (y_{k+1}, y_{k-1})$ (les λ_k sont rangés par ordre croissant tandis que les y_k sont décroissants).