THE COALESCENT
Lectures given at the
CIMPA–IRD–UCAD–school
Dakar, Sénégal, February 2011

Étienne Pardoux

Contents

Chapter 1

Kingman's coalescent

1.1 The Wright–Fisher model

Consider a population of fixed size N, which evolves in discrete generations. Each individual of generation k chooses his father uniformly among the individuals of the previous generation, independently of the choices of the other individuals.

Looking backward in time, if we sample n individuals in the present population, we want to describe at which generation any two of those had the same common ancestor, until we reach the most recent common ancestor of the sample.

1.2 Cannings' model

We can generalize the Wright–Fisher model as follows. Suppose at each generation, we label the N individuals randomly. For $r \geq 0$, $1 \leq i \leq N$, let ν_i^r denote the number of offsprings in generation $r+1$ of the i–th individual from generation r. Clearly those r. v.'s must satisfy the requirement that

$$\nu_1^r + \cdots + \nu_N^r = N.$$

Cannings' model stipulates moreover that

$$\nu^r,\ r \geq 0 \ \text{ are i. i. d. copies of } \nu,$$

and that the law of ν is exchangeable, i. e.

$$(\nu_1, \ldots, \nu_N) \simeq (\nu_{\pi(1)}, \ldots, \nu_{\pi(N)}), \forall \pi \in S_N.$$

The above conditions imply that $\mathbb{E}\nu_1 = 1$. To avoid the trivial case where $\mathbb{P}(\nu_1 = \cdots = \nu_N = 1) = 1$, we assume that $\mathrm{Var}(\nu_1) > 0$. A particular case of Cannings' model is the Wright–Fisher model, in which ν is multinomial.

1.3 Looking backward in time

Consider a population of fixed size N, which has been reproducing for ever according to Cannings' model. We sample $n < N$ individuals from the present generation, and label them $1, 2, \ldots, n$. For each $r \geq 0$, we introduce the equivalence relation on the set $\{1, \ldots, n\}$: $i \sim_r j$ if the individuals i and j have the same ancestor r generations back in the past. Denote this equivalence relation by $R_r^{N,n}$. For $r \geq 0$, $R_r^{N,n}$ is a random equivalence relation, which can be described by its associated equivalence classes, which is a random partition of $(1, \ldots, n)$. Thus $\{R_r^{N,n};\ r \geq 0\}$ is a Markov chain with values in the set $\mathcal{E}_n$ of the partitions of $(1, \ldots, n)$, which starts from the trivial *finest* partition $(\{1\}, \ldots \{n\})$, and eventually reaches the *coarsest* partition consisting of the set $\{1, \ldots, n\}$ alone. We denote by $P_{\xi,\eta}^{N,n}$ the transition matrix of that chain.

The probability that two individuals in today's population have the same ancestor in the previous generation is

$$c_N = \frac{\sum_{i=1}^N \mathbb{E}\left[\binom{\nu_i}{2}\right]}{\binom{N}{2}} = \frac{\sum_{i=1}^N \mathbb{E}[\nu_i(\nu_i - 1)]}{N(N-1)} = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)]}{N-1}.$$

Provided that $c_N \to 0$ as $N \to \infty$, if $r = t/c_N$,

$$\mathbb{P}(1 \not\sim_r 2) = (1 - c_N)^r \approx e^{-t}.$$

This suggests to consider

$$\mathcal{R}_t^{N,n} := R_{[t/c_N]}^{N,n},\ t \geq 0.$$

1.4 Kingman's coalescent

Let $\{\mathcal{R}^n_t;\ t \geq 0\}$ be a continuous time $\mathcal{E}_n$–valued jump Markov process with the rate matrix given by (for $\eta \neq \xi$)

$$Q_{\xi\eta} = \begin{cases} 1 & \text{, if } \eta \text{ is obtained from } \xi \text{ by merging exactly two classes,} \\ 0 & \text{, otherwise.} \end{cases} \tag{1.4.1}$$

This is Kingman's n coalescent. In order for $\mathcal{R}^{N,n}$ to converge to Kingman's coalescent, we certainly need that merges of 3 or more lineages are asymptotically negligible. The probability that three individuals in today's population have the same ancestor in the previous generation is

$$d_N := \frac{\sum_{i=1}^N \mathbb{E}\left[\binom{\nu_i}{3}\right]}{\binom{N}{3}} = \frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{(N-1)(N-2)}.$$

Exercise 1.4.1. *Compute c_N and d_N in the Wright–Fisher model, as well as in the model where at each generation a common father of all individuals of the next generation is chosen uniformly in the present generation.*

Theorem 1.4.2. *$\mathcal{R}^{N,n} \Rightarrow \mathcal{R}^n$ in $D(\mathbb{R}_+; \mathcal{E}_n)$ iff, as $N \to \infty$, both*

$$\begin{cases} c_N \to 0, \\ \dfrac{d_N}{c_N} \to 0. \end{cases} \tag{1.4.2}$$

Remark 1.4.3. Non-constant population size *This result assumes in an essential way that the size of the population is constant in time. What is the effect of modifying the population size ? Assume (that is true in particular for the Wright–Fisher model) that $\mathbb{E}[\nu_1(\nu_1-1)] \to c > 0$ as $N \to \infty$. In that case our theorem says roughly that for large N, $R^{N,n}_{Nt/c} \simeq \mathcal{R}^n_t$. Then for any $x > 0$, we have similarly that $R^{xN,n}_{xNt/c} \simeq \mathcal{R}^n_t$. In other words, $R^{xN,n}_{Nt/c} \simeq \mathcal{R}^n_{t/x}$. This means that if we multiply the size of the population by a factor x, we should accelerate time by a factor $1/x$, or, what is exactly the same, multiply the pairwise coalescence rate by the factor $1/x$. This argument can be justified in the case of a varying population size. The rule is to multiply at each time t the pairwise coalescence rate by 1 over the "renormalized population size".*

PROOF: The sufficiency will follow from the standard Lemma 1.4.4 below and the fact that (1.4.2) implies that

$$P_{\xi,\eta}^{N,n} = \delta_{\xi,\eta} + c_N Q_{\xi,\eta} + \circ(c_N),$$

where the error term is small, uniformly with respect to ξ, $\eta \in \mathcal{E}_n$. It follows from exchangeability that for any $f : \{0, 1, \ldots, N\} \to \mathbb{R}_+$,

$$\begin{aligned}(N-1)\mathbb{E}[\nu_2 f(\nu_1)] &= \sum_{j=2}^{N} \mathbb{E}[\nu_j f(\nu_1)] \\ &= \mathbb{E}[(N-\nu_1) f(\nu_1)] \\ &\leq N\mathbb{E}[f(\nu_1)],\end{aligned}$$

hence

$$\mathbb{E}[\nu_2 f(\nu_1)] \leq \frac{N}{N-1}\mathbb{E}[f(\nu_1)]. \tag{1.4.3}$$

From the Markov inequality and (1.4.2), with the notations $(\nu)_2 = \nu(\nu-1)$, $(\nu)_3 = \nu(\nu-1)(\nu-2)$, if $\varepsilon N \geq 2$,

$$\begin{aligned}\mathbb{P}(\nu_1 > \varepsilon N) &\leq \frac{\mathbb{E}[(\nu_1)_3]}{(\varepsilon N)_3} \\ &= \frac{\circ(N\mathbb{E}[(\nu_1)_2])}{\varepsilon^3 N^3},\end{aligned}$$

consequently

$$\mathbb{P}(\nu_1 > \varepsilon N) \leq \varepsilon^{-3} \circ (c_N/N). \tag{1.4.4}$$

Next

$$\begin{aligned}\mathbb{E}[(\nu_1)_2(\nu_2)_2] &\leq \varepsilon N\mathbb{E}[(\nu_1)_2\nu_2; \nu_2 \leq \varepsilon N] + N^2\mathbb{E}[(\nu_1)_2; \nu_2 > \varepsilon N] \\ &\leq \varepsilon N\mathbb{E}[(\nu_1)_2\nu_2] + N^3\mathbb{E}[\nu_1; \nu_2 > \varepsilon N] \\ &\leq \varepsilon N\frac{N}{N-1}\mathbb{E}[(\nu_1)_2] + N^3\frac{N}{N-1}\mathbb{P}(\nu_2 > \varepsilon N),\end{aligned}$$

where we have used (1.4.3) twice in the last inequality. Combining this with (1.4.4), we conclude that for all $\varepsilon > 0$,

$$\begin{aligned}\limsup_{N\to\infty} \frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{N\mathbb{E}[(\nu_1)_2]} &\leq \varepsilon + \limsup_{N\to\infty} \frac{\mathbb{P}(\nu_1 > \varepsilon N)}{c_N/N} \\ &= \varepsilon.\end{aligned}$$

Let $I_1, \dots, I_n$ denote the parents of n ordered randomly chosen individuals of a given generation. We have the following identities

$$\begin{aligned}
\mathbb{P}(I_1 = I_2) &= c_N \\
\mathbb{P}(I_1 = I_2 = I_3) &= d_N \\
\mathbb{P}(I_1 = I_2 \neq I_3 = I_4) &= \frac{\sum_{1 \leq i < j \leq N} \mathbb{E}\left[\binom{\nu_i}{2}\binom{\nu_j}{2}\right]}{\binom{N}{4}} \\
&= 3\frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{(N-2)(N-3)}.
\end{aligned}$$

Hence we deduce from the last estimate that

$$\lim_{N\to\infty} \frac{\mathbb{P}(I_1 = I_2 \neq I_3 = I_4)}{\mathbb{P}(I_1 = I_2)} = 0, \tag{1.4.5}$$

while (1.4.2) tells us that

$$\lim_{N\to\infty} \frac{\mathbb{P}(I_1 = I_2 = I_3)}{\mathbb{P}(I_1 = I_2)} = 0. \tag{1.4.6}$$

We now conclude, using (1.4.5) and (1.4.6). Let $\xi = (C_{11}, C_{12}, C_2, \dots, C_a)$ and $\eta = (C_1, C_2, \dots, C_a)$, where $C_1 = C_{11} \cup C_{12}$. We have

$$\begin{aligned}
\mathbb{P}(I_1 = I_2) &- \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq m \leq a+1;\ I_m = I_1\}) \\
&- \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq \ell < m \leq a+1;\ I_\ell = I_m \neq I_1\}) \\
&\leq P_{\xi,\eta}^{N,n} \leq \mathbb{P}(I_1 = I_2).
\end{aligned}$$

From (1.4.6),

$$\begin{aligned}
\mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq m \leq a+1;\ I_m = I_1\}) &\leq (a-1)\mathbb{P}(I_1 = I_2 = I_3) \\
&= \circ(\mathbb{P}(I_1 = I_2)),
\end{aligned}$$

and from (1.4.5),

$$\begin{aligned}
\mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq \ell < m \leq a+1;\ I_\ell = I_m \neq I_1\}) &\leq \binom{a-1}{2}\mathbb{P}(I_1 = I_2 \neq I_3 = I_4) \\
&= \circ(\mathbb{P}(I_1 = I_2))
\end{aligned}$$

We have proved that for such a pair (ξ, η), $P^{N,n}_{\xi,\eta} = c_N + \circ(c_N)$. If η' is obtained from ξ by merging more than two classes, then there must be at least either a triple merger or two double mergers, hence from (1.4.6), (1.4.5), $P^{N,n}_{\xi,\eta'} = \circ(c_N)$. Finally, since $|\mathcal{E}_n| < \infty$ and $\sum_{\eta\in\mathcal{E}_n} P^{N,n}_{\xi,\eta} = 1$,

$$\begin{aligned} P^{N,n}_{\xi,\xi} &= 1 - \binom{|\xi|}{2} c_N + \circ(c_N) \\ &= 1 + Q_{\xi,\xi} c_N + \circ(c_N). \end{aligned}$$

□

It remains to prove :

Lemma 1.4.4. *Let E be a finite set and $\{X_t,\ t \geq 0\}$ a continuous time E–valued jump Markov process, with generator $Q = (Q_{x,y})_{x,y\in E}$. Let for each $N \in \mathbb{N}$ X^N be a discrete time Markov chain with transition matrix satisfying*

$$P^N_{x,y} = \delta_{x,y} + c_N Q_{x,y} + \circ(c_N),\ x, y \in E,$$

where $c_N \to 0$, as $N \to \infty$. Then whenever $X^N_0 \Rightarrow X_0$,

$$\{X^N_{[t/c_N]},\ t \geq 0\} \Rightarrow \{X_t,\ t \geq 0\} \quad \text{in } D([0, +\infty); E).$$

Proof: The fact that for any $x, y \in E$, $s, t > 0$,

$$\mathbb{P}(X^N_{[(t+s)/c_N]} = y | X^N_{[t/c_N]} = x) \to \mathbb{P}(X_{t+s} = y | X_t = x),$$

together with the Markov property, implies the convergence of finite dimensional distributions. Indeed this follows easily from the fact that

$$\begin{aligned} \mathbb{P}(X^N_{[(t+s)/c_N]} = y | X^N_{[t/c_N]} = x) &= \left(P^N\right)^{s/c_N}_{xy} \\ &= (I + c_N Q + \circ(c_N))^{s/c_N}_{xy} \\ &= (e^{c_N Q} + \circ(c_N))^{s/c_N}_{xy} \\ &\to (e^{sQ})_{xy} \end{aligned}$$

It remains to prove tightness in $D([0, \infty); E)$. This follows essentially from the fact that the probability that X^N jumps more than once in an interval of length δ is of the order $\circ(\delta)$, uniformly in N. We skip the details. □

Let $\{\mathcal{R}^n_t;\ t \geq 0\}$ start from the trivial partition of $(1, \ldots, n)$. For $2 \leq k \leq n$, let T_k denote the length of the time interval during which there are

k branches alive. From the Markov property of the coalescent, and the form of the generator, we deduce that

$$\begin{aligned} &T_n, T_{n-1}, \ldots, T_2 \text{ are independent,} \\ &T_k \simeq \mathcal{E}\mathrm{xp}\left(\binom{k}{2}\right),\ 2 \leq k \leq n, \end{aligned}$$

and consequently the expected time till the Most Recent Common Ancestor in the sample is

$$\begin{aligned} \sum_{k=2}^{n} \frac{2}{k(k-1)} &= 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) \\ &= 2\left(1 - \frac{1}{n}\right). \end{aligned}$$

For $n' > n$, denote by d_n the restriction to $\mathcal{E}_n$ of an element of $\mathcal{E}_{n'}$. Kingman's n–coalescents have the consistency property that

$$d_n\left(\{\mathcal{R}_t^{n'},\ t \geq 0\}\right) \simeq \{\mathcal{R}_t^n,\ t \geq 0\}.$$

This, together with the fact that $\sum_{k\geq 2} T_k < \infty$ a. s., since the series of the expectations converges, allows us to define Kingman's coalescent $\{\mathcal{R}_t,\ t \geq 0\}$ as the limit $\lim_{n\to\infty}\{\mathcal{R}_t^n,\ t \geq 0\}$. It is readily seen that Kingman's coalescent *comes down from infinity*, in the sense that, while $\mathcal{R}_0$ is the trivial partition of $\mathbb{N}^*$, hence $|\mathcal{R}_0| = \infty$, $|\mathcal{R}_t| < \infty$, $\forall t > 0$.

1.5 The height and the length of Kingman's coalescent

The *height* of Kingman's n–coalescent is the r. v.

$$H_n = \sum_{k=2}^{n} T_k,$$

where the T_k are as above. This prescribes the law of H_n, which does not obey any simple formula. Note that

$$\mathbb{E}(H_n) = 2\left(1 - \frac{1}{n}\right), \quad \mathrm{Var}(H_n) = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2}.$$

$\mathbb{E}(H_n) \to 2$ as $n \to \infty$, and $\sup_n \mathrm{Var}(H_n) < \infty$.

The *length* of Kingman's n–coalescent (i. e. the sum of the lengths of the branches of this tree) is the r. v.

$$L_n = \sum_{k=2}^{n} kT_k = \sum_{k=2}^{n} U_k,$$

where the U_k are independent, U_k is an $\mathrm{Exp}((k-1)/2)$ r. v. The distribution function of L_n is given by

Proposition 1.5.1. *For all $x \geq 0$,*

$$\mathbb{P}(L_n \leq x) = (1 - e^{-x/2})^{n-1}.$$

This Proposition follows from the fact that the law of L_n is that of the sup over $n-1$ i. i. d. $\mathrm{Exp}(1/2)$ r. v.'s, which is a consequence of the

Proposition 1.5.2. *Let $V_1, V_2, \ldots, V_n$ be i. i. d. Exp(λ) r. v.'s, and $V_{(1)} < V_{(2)} < \cdots < V_{(n)}$ denote the same random sequence, but arranged in increasing order. Then $V_{(1)}, V_{(2)} - V_{(1)}, \ldots, V_{(n)} - V_{(n-1)}$ are independent exponential r. v.'s with respective parameters $n\lambda, (n-1)\lambda, \ldots, \lambda$.*

Proof: For any Borel measurable function $f : \mathbb{R}_+^n \to \mathbb{R}_+$,

$$\begin{aligned}
&\mathbb{E} f(V_{(1)}, V_{(2)} - V_{(1)}, \ldots, V_{(n)} - V_{(n-1)}) \\
&\quad = n! \mathbb{E}\left[f(V_1, V_2 - V_1, \ldots, V_n - V_{n-1}); V_1 < V_2 < \cdots < V_n\right] \\
&\quad = n! \int_{0<x_1<x_2<\cdots<x_n} f(x_1, x_2 - x_1, \ldots, x_n - x_{n-1}) \lambda^n e^{-\lambda \sum_{k=1}^n x_k} dx_1 dx_2 \cdots dx_n \\
&\quad = \prod_{k=1}^{n} (k\lambda) \int_0^\infty \cdots \int_0^\infty f(y_1, y_2, \ldots, y_n) \prod_{k=1}^{n} e^{-k\lambda y_{n+1-k}} dy_1 dy_2 \cdots dy_n.
\end{aligned}$$

The result follows. □

Exercise 1.5.3. *A Yule tree of rate λ is a random tree which develops as follows. Let T_k, $k \geq 1$ be independent r. v.'s, T_k being exponential with parameter λk. For $0 \leq t < T_1$, the tree has a unique branch issued from the root. At time T_1 this branch splits into 2. For $T_1 \leq t < T_1 + T_2$, there are two branches. At time $T_1 + T_2$, we choose one of the two branches with equal probability, and that branch splits into 2, etc...*

Deduce from Proposition 1.5.2 that the law of the number Y_t of branches of the tree at time t is geometric with parameter $e^{-\lambda t}$, in the sense that for $k \geq 0$,

$$\mathbb{P}(Y_t \geq k) = (1 - e^{-\lambda t})^k.$$

1.5.1 More on the length of Kingman's coalescent

The following result will be useful for analysing the effect of mutations.

Proposition 1.5.4. *The expected total length of edges in the N coalescent supporting exactly i leaves is $2/i$, $1 \leq i \leq N-1$.*

It is remarkable that this quantity does not depend upon N. As N increases, there are more such branches, but they tend to be shorter.
PROOF: We shall use the notation

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!}, & \text{if } n > k; \\ 1, & \text{if } n = k; \\ 0, & \text{if } n < k. \end{cases}$$

Let us first establish the elementary identity

$$\binom{n}{j-1} = \binom{n+1}{j} - \binom{n}{j}, \tag{1.5.1}$$

which follows from

$$\begin{aligned} \binom{n}{j-1} + \binom{n}{j} &= \frac{n!}{(j-1)!(n-j+1)!} + \frac{n!}{j!(n-j)!} \\ &= \frac{n!j + n!(n-j+1)}{j!(n-j+1)!} \\ &= \frac{(n+1)!}{j!(n+1-j)!}. \end{aligned}$$

Consider a portion of an edge of the coalescent, while there are k active lineages. What is the probability of a configuration where this edge supports exactly i leaves ? Since while going down towards the leaves each new split concerns any of the active lineages with equal probability, it is not hard to see that this probability equals

$$\frac{(i-1)!\ (k-1)(k(k+1)\cdots(N-i-1)}{k(k+1)\cdots(N-1)}.$$

The number of such configurations is $\binom{N-k}{i-1}$, and each of the k active lineages has the same probability of supporting exactly i leaves. Multiplying by $\frac{2}{k(k-1)}$, the mean length of time when there are k active lineages, and summing upon k, we obtain that if $L_{N,i}$ denotes the total length of all those edges which support exactly i leaves,

$$\begin{aligned}
\mathbb{E}[L_{N,i}] &= \sum_{k=2}^{N} \frac{\binom{N-k}{i-1}}{\binom{N-1}{i}} \frac{k-1}{i} \frac{2}{k-1} \\
&= \frac{2}{i} \frac{1}{\binom{N-1}{i}} \sum_{k=2}^{N} \binom{N-k}{i-1} \\
&= \frac{2}{i} \frac{1}{\binom{N-1}{i}} \sum_{k=2}^{N} \left[\binom{N-(k-1)}{i} - \binom{N-k}{i} \right] \\
&= \frac{2}{i} \frac{1}{\binom{N-1}{i}} \left[\sum_{k=1}^{N-1} \binom{N-k)}{i} - \sum_{k=2}^{N} \binom{N-k}{i} \right] \\
&= \frac{2}{i},
\end{aligned}$$

where we have used (1.5.1) at the third step. □

1.6 Kingman's unlabeled n–coalescent

We have introduced Kingman's n–coalescent as a process denoted as $\{\mathcal{R}_t^n;\ t \geq 0\}$, with values in the set $\mathcal{E}_n$ of partitions of the set $\{1, 2, \ldots, n\}$. Let $R_t^n := |\mathcal{R}_t^n|$ denote the number of blocks of the partition $\mathcal{R}_t^n$. It is easily seen from the above considerations that $\{R_t^n;\ t \geq 0\}$ is again a Markov process, in fact a pure death process, with death rate $\binom{k}{2}$ while $R_t = k$. Clearly the process R_t^n carries less information than the process $\mathcal{R}_t^n$.

There is in fact an intermediate coalescent process, called Kingman's unlabeled coalescent, which records how many individuals at time t are the

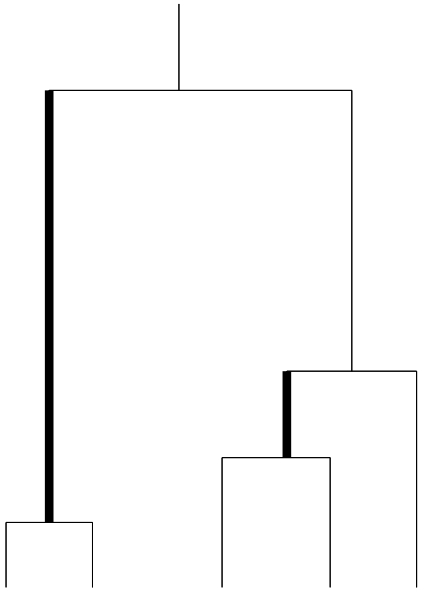

Figure 1.1: The edges supporting exactly 2 leaves

ancestors of j individuals from the sample at time 0, $1 \leq j \leq n$, without specifying which individual are their descendants. In other words, we record how many active lineages at time t subtend j leaves, for $j = 1, \ldots, n$. We denote this process by $\{\mathbf{R}_t^n,\ t \geq 0\}$.

$\mathbf{R}_t^n$ takes its values in the set $\mathbb{F} = \cup_{i=1}^n \mathbb{F}_n^{(i)}$, where for each $1 \leq i \leq n$,

$$\mathbb{F}_n^{(i)} = \left\{ f_i = (f_{i,1}, f_{i,2}, \ldots, f_{i,n}) \in \mathbb{Z}_+^n;\ \sum_{j=1}^n j f_{i,j} = n,\ \sum_{j=1}^n f_{i,j} = i \right\}.$$

The process $\mathbf{R}_t^n$ starts from $\mathbf{R}_0^n = (n, 0, \ldots, 0)$ and reaches eventually the value $(0, \ldots, 0, 1)$ when the sample finds its MRCA. Those are the unique points in $\mathbb{F}_n^{(n)}$ and $\mathbb{F}_n^{(1)}$ respectively. The process $\mathbf{R}_t^n$ jumps first from $(n, 0, \ldots, 0)$ to $(n-2, 1, 0, \ldots, 0)$, then to a point in $\mathbb{F}_n^{(n-2)}$, etc... Not all trajectories are possible. Let

$$\mathbb{G}_n = \left\{ f = (f_n, f_{n-1}, \ldots, f_1);\ f_i \in \mathbb{F}_n^{(i)}, 1 \leq i \leq n \text{ and } f_{i-1} \prec f_i, 2 \leq i \leq n \right\},$$

where for $f_i \in \mathbb{F}_n^{(i)}$, $f_{i'} \in \mathbb{F}_n^{(i-1)}$,

$$f_{i'} \prec f_i \Leftrightarrow \text{there exists } 1 \leq k, \ell < n \text{ such that } f_{i'} = f_i - e_k - e_\ell + e_{k+\ell},$$

with the notation $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, the unique nonzero coordinate of e_k being the kth coordinate. The sequence of the successive states visited by the jump Markov process $\mathbf{R}_t^n$ (i. e. its embedded chain) belongs a. s. to $\mathbb{G}_n$. Let $(U_n^n, U_{n-1}^n, \ldots, U_1^n)$ denote that sequence. We have the

Proposition 1.6.1. *For any $f \in \mathbb{G}_n$,*

$$\mathbb{P}((U_n^n, U_{n-1}^n, \ldots, U_1^n) = f) = \prod_{j=n}^2 P_{f_j, f_{j-1}},$$

where

$$P_{f_j, f_{j-1}} = \begin{cases} f_{j,k} f_{j,\ell} \binom{j}{2}^{-1}, & \text{if } f_{j-1} = f_j - e_k - e_\ell + e_{k+\ell},\ k \neq \ell; \\ \binom{f_{j,k}}{2} \binom{j}{2}^{-1}, & \text{if } f_{j-1} = f_j - 2e_k + e_{2k}; \\ 0, & \text{otherwise.} \end{cases}$$

PROOF: It is not hard to see that indeed $(U_n^n, U_{n-1}^n, \ldots, U_1^n)$ is a non homogeneous Markov chain, which has the transition probability described in the statement of the proposition. □

We can further describe the marginal laws of the U_k^n's.

Proposition 1.6.2. *For any $1 \leq j \leq n$, any $f_j \in \mathbb{F}_n^{(j)}$,*

$$\mathbb{P}(U_j^n = f_j) = \frac{i!}{\prod_{k=1}^j f_{j,k}!} \binom{n-1}{j-1}^{-1}.$$

The following follows readily from Propositions 1.6.1 and 1.6.2 and Bayes formula.

Proposition 1.6.3.

$$\mathbb{P}((U_n^n, U_{n-1}^n, \ldots, U_1^n) = f) = \prod_{j=n}^{2} Q_{f_{j-1}, f_j},$$

where for eah $1 \leq j \leq n$,

$$Q_{f_{j-1},f_j} = \begin{cases} \frac{2f_{j-1,k+\ell}}{n-j+1}, & \text{if } f_j = f_{j-1} + e_k + e_\ell - e_{k+\ell},\ k \neq \ell; \\ \frac{f_{j-1,2k}}{n-j+1}, & \text{if } f_j = f_{j-1} + 2e_k - e_{2k}; \\ 0, & \text{otherwise.} \end{cases}$$

Remark 1.6.4. *The relevance of the unlabeled coalescent comes from the following fact. In the framework of the infinitely many sites mutation model which we shall describe below, the so–called Site Frequency Spectrum records how many individuals have been jointly affected by each given mutation, without recording who was affected by which mutation. This gives us informations about the trajectory back in time of the unlabeled coalescent of our given sample. The refer the interested reader to [16] for more details about how to exploit that information in a Bayesian context.*

Chapter 2

Mutations : the infinitely many alleles model

Suppose now that mutations arise on each branch of the coalescence tree, according to a Poisson process with parameter $\theta/2$, see Figure 2.1. Assume that each mutation gives birth to a new type, different for all the others. For instance we may assume that the different types are i. i. d. r. v.'s following the uniform law on $[0, 1]$. We want to record the different types in a sample drawn at present time, we can as well "kill" the lineages which hit a mutation while going backward in time, which changes Figure 2.1 into Figure 2.2, which we can as well change into Figure 2.3. The killed coalescent can be produced by the following procedure : *Any pair of active classes merges at rate 1, any active class is killed at rate* $\theta/2$. When a class is killed, all its elements are assigned the same (different from all other classes) type. Finish when there are no classes left. Note that we add a mutation at the root of the tree.

2.1 Hoppe's urn

Assume that there are k active classes in the killed coalescent described above. Then the probability that the next (backward in time) event is a

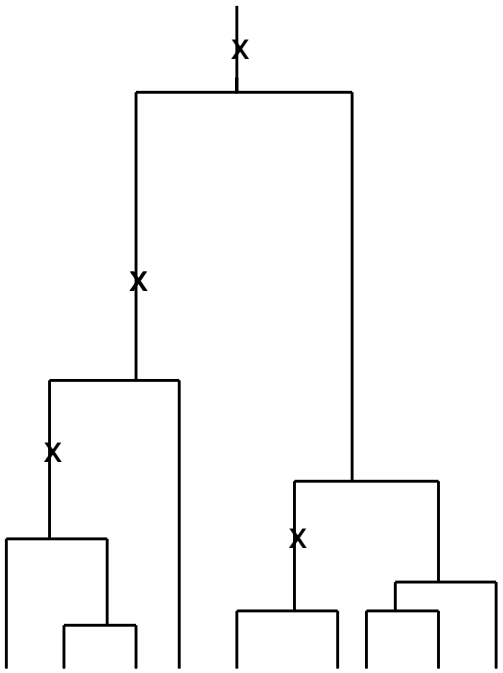

Figure 2.1: The coalescent with mutations

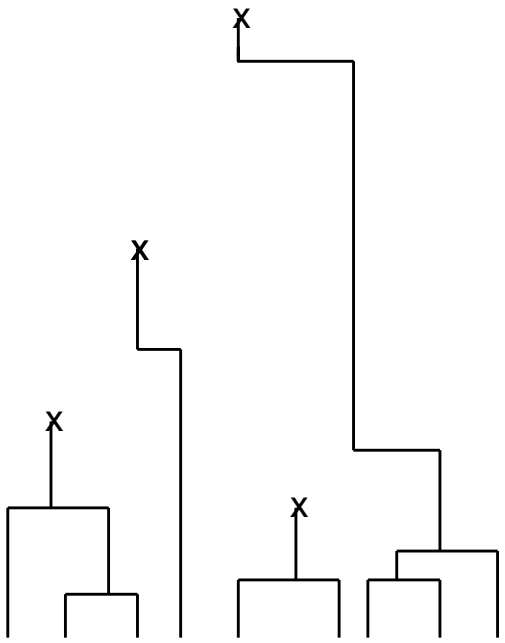

Figure 2.2: The lineages are killed above the mutations

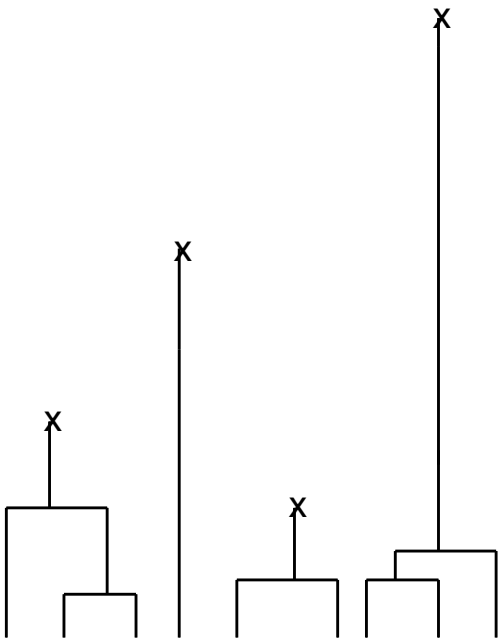

Figure 2.3: Equivalent to Figure 2.2

coalescence is

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

and the probability that that event is a mutation (i. e. a killing) is

$$\frac{k\frac{\theta}{2}}{\binom{k}{2}+k\frac{\theta}{2}} = \frac{\theta}{k-1+\theta}.$$

Moreover, given the type of event, all possible coalescence (resp. mutations) are equally likely. The history of a sample of size n is described by n events $e_n, e_{n-1}, \ldots, e_1 \in \{\texttt{coal}, \texttt{mut}\}$. Note that the event e_k happens just before (forward in time) k lineages are active, and each of those events corresponds backward in time to the reduction by one of the number of active lineages. The probability to observe a particular sequence is thus

$$\frac{\prod_{k=1}^{n}\left(\theta \mathbf{1}_{\{e_k=\texttt{mut}\}} + (k-1)\mathbf{1}_{\{e_k=\texttt{coal}\}}\right)}{\prod_{k=1}^{n}(k-1+\theta)}. \tag{2.1.1}$$

Hoppe [8] noted that one can generate this sequence *forward in time* using the following urn model.

Hoppe's urn model. We start with an urn containing one unique black ball of mass θ. At each step, a ball is drawn from the urn, with probability proportional to its mass. If the drawn ball is black return it to the urn, together with a ball of mass 1, of a new, not previously used, colour; if the drawn ball is coloured, return it together with another ball of mass 1 of the same colour.

At the k–th step, there are k balls, more precisely $k-1$ coloured balls, plus the black (so called *mutation*) ball. The probability to pick the black ball is thus $\theta/(k-1+\theta)$ while the probablity to pick a coloured ball is $(k-1)/(k-1+\theta)$. If we define

$$e_k = \begin{cases} \texttt{mut}, & \text{if in the } k\text{–step the black ball is drawn,} \\ \texttt{coal}, & \text{otherwise.} \end{cases}$$

Clearly the probability to observe a particular sequence $(e_1, \ldots, e_n)$ is given by (2.1.1). Moreover, given that $e_k = \texttt{coal}$, each of the $k-1$ present coloured balls is equally likely to be picked.

Consequently, the distribution of the family sizes generated by the n coloured balls in Hoppe's urn after n steps is the same as the one induced by the n–coalescent in the infinitely–many–alleles mutation model.

2.2 Ewens' sampling formula

Theorem 2.2.1. *Let $b_1, \ldots, b_n \in \mathbb{N}$ be such that $\sum_{j=1}^n j b_j = n$. The probability of observing b_j different types, each with j representatives, $(j = 1, \ldots, n)$ in a sample of size n is given by (here $k = \sum_{j=1}^n b_j$)*

$$\frac{n!}{1^{b_1} 2^{b_2} \cdots n^{b_n}} \cdot \frac{1}{b_1! b_2! \cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}. \tag{2.2.1}$$

PROOF: We shall prove that the distribution of the type spectrum $(B_1, \ldots, B_n)$ in a sample of size n is the product of the measures $\mathrm{Poi}(\theta/j)$, $j = 1, \ldots, n$, conditionned on $\sum_{j=1}^n j B_j = n$.

We start from the statement at the very end of the previous section. Now we describe another way of constructing the output of Hope's urn after n steps. Consider on $\mathbb{R}_+$ a Poisson process of immigrants with parameter θ. Each new immigrant starts immediately upon arrival to develop a Yule tree of parameter 1 (that is a new branch appears after a waiting time which is exponential with parameter 1, .. when they are k branches alive, a $k+1$–st appears after a waiting time which is exponential with parameter k, etc.., the successive waiting times being mutually independent and independent of the time of arrival of the founder of the tree), and moreover the various trees are mutually independent. It follows from Exercise 1.5.3 that the number of branches of a Yule tree at time t (the tree being started at time 0) is geometric with parameter e^{-t}.

We can describe this model as follows. Consider the Markov process $\{Y_t,\ t \geq 0\}$ with values in the subset E of $\mathbb{N}^\infty$ consisting of those sequences whose only a finite number of components are non zero. $Y_0 = (0, 0, \ldots)$, immigrants enter at rate θ, the first immigrant creates the first tree, the second immigrant creates the second tree, etc...Each tree is a Yule tree, with develops independently of the other trees and of the arrivals of new immmigrants. $Y_t = (Y_t^1, Y_t^2, \ldots)$, where Y_t^k denotes the number of branches at time t of the k–th Yule tree. Define

$$|Y_t| = \sum_{k \geq 1} Y_t^k$$

the total number of branches of all the trees at time t. $|Y_t|$ is a birth Markov process, the waiting time for the next birth when $|Y_t| = k$ being exponential with parameter $\theta + k$.

It is easily seen that what we have just constructed is exactly a continuous time embedding of Hope's urn (just compute at each time $s < t$ what is the probability that the next event is the arrival of a new immigrant, or the appearance of a new branch on an existing tree). Hence the output of Hoppe's urn has the same law as the set Y_t of Yule trees which this construction produces, if we look at it at the time when $|Y_t|$ reaches the value n. It follows from Exercise 2.2.2 below that this law is the same as the law of Y_t, conditionned upon $|Y_t| = n$, for all $t > 0$.

This continuous time model can be considered as a Poisson process on $[0, t] \times \mathbb{N}$, with the intensity measure $\theta ds \times \mathcal{G}(e^{-(t-s)})$, where for $0 < p < 1$, $\mathcal{G}(p)$ denotes the geometric measure of parameter p. A point of this Poisson process is a pair (s, j), where $s \in [0, t]$ and $j \geq 1$. The point (s, j) corresponds to an immigrant which has appeared at time s, and whose associated Yule tree at time t has exactly j branches. Now for $j \geq 1$, let $Z_j(t)$ denote the number of points of the above Poisson process whose second component equals j. It follows from well–known properties of Poisson processes that the $Z_j(t)$, $j \geq 1$ are mutually independent r. v.'s, the law of $Z_j(t)$ being Poisson with parameter

$$\int_0^t \theta e^{-(t-s)} \left(1 - e^{-(t-s)}\right)^{j-1} ds = \frac{\theta}{j}(1 - e^{-t})^j.$$

The above arguments show that the probability of observing b_j different types, each with j representatives, $(j = 1, \ldots, n)$ in a sample of size n equals

$$\mathbb{P}\left(Z_1(t) = b_1, \ldots, Z_n(t) = b_n \left| \sum_{j=1}^n j Z_j(t) = n \right. \right).$$

This is true for any $t > 0$. We can as well let $t \to \infty$, and we deduce that the same probability equals

$$\mathbb{P}\left(Z_1 = b_1, \ldots, Z_n = b_n \left| \sum_{j=1}^n j Z_j = n \right. \right),$$

where $Z_1, \ldots, Z_n$ are independent, and for each $1 \leq j \leq n$, the law of Z_j is

Poisson with parameter θ/j. This quantity is equal to

$$C(n,\theta)\prod_{j=1}^{n} e^{-\theta/j}\frac{(\theta/j)^{b_j}}{b_j!},$$

where the normalization constant satisfies

$$C(n,\theta)^{-1} = \mathbb{P}\left(\sum_{j=1}^{n} jB_j = n\right).$$

The result is proved, provided we check that

$$C(n,\theta) = \frac{n!\exp[\theta\sum_{j=1}^{n} 1/j]}{\theta(\theta+1)\cdots(\theta+n-1)}.$$

This will be done below in Lemma 2.2.3. Note however that we have already identified the Ewens sampling formula up to a normalization constant. □

Exercise 2.2.2. *Let $\{X_t,\ t\geq 0\}$ be a continuous time jump–Markov process, which takes values in a countable set E. Let $T_0 = 0$ and T_n, $n\geq 1$ denote the n–th jump time of X_t. Let $\{Z_n,\ n\geq 0\}$ denote the associated embedded Markov chain, i. e. $Z_0 = X_0$, and for all $n\geq 1$, $Z_n = X_{T_n}$. We know that there exists a function $q : E\to(0,\infty)$ such that for each $n\geq 0$,the law of $T_{n+1}-T_n$ is exponential with parameter $q(Z_n)$. Suppose that there exists a function $h:\mathbb{N}\to(0,\infty)$ such that $q(Z_n) = h(n)$, $n\geq 0$. Conclude that the sequences $\{T_n,\ n\geq 1\}$ and $\{Z_n,\ n\geq 1\}$ are mutually independent. Why is this last property not true in general ?*

Apply this result to the process $\{Y_t,\ t\geq 0\}$ from the previous proof. Show that the condition on q is satisfied here with $h(n) = \theta+n$. Prove that for all $t>0$, the law of Z_n equals the conditional law of Y_t, given that $|Y_t| = n$.

We finally prove the

Lemma 2.2.3. *If $B_1,\ldots,B_n$ are independent, each B_j being Poisson with parameter θ/j, then*

$$\mathbb{P}\left(\sum_{j=1}^{n} jB_j = n\right) = \frac{\theta(\theta+1)\cdots(\theta+n-1)}{n!\exp[\theta\sum_{j=1}^{n} 1/j]}.$$

PROOF: The left hand side of the identity to be established equals

$$\sum_{k_1,\ldots,k_n;\sum jk_j=n} e^{-\theta/j}\,(\theta/j)^{k_j}\,/k_j! = \exp[-\theta\sum_{j=1}^{n} 1/j]\sum_k \alpha(n,k)\theta^k,$$

where

$$\alpha(n,k) = \sum_{k_1,\ldots,k_n;\sum k_j=k,\ \sum jk_j=n} \left(\prod_{j=1}^{n} j^{k_j}k_j!\right)^{-1}.$$

It remains to show that

$$\theta(\theta+1)\cdots(\theta+n-1) = n!\sum_{k=1}^{n}\alpha(n,k)\theta^k.$$

Let $s(n,k) = n!\alpha(n,k)$. Splitting the last factor in the above left hand side into θ plus $n-1$, we deduce that

$$s(n,k) = s(n-1,k-1) + (n-1)s(n-1,k).$$

This shows that $s(n,k)$ can be interpreted as the number of permutations of $\{1,\ldots,n\}$ which contain exactly k cycles. Now that number is given by

$$\begin{aligned} s(n,k) &= \sum_{k_1,\ldots,k_n,\sum k_j=k,\sum jk_j=n} \frac{n!}{\prod_{j=1}^{n}(jk_j)!} \times \prod_{j=1}^{n}\left(\frac{(jk_j)!}{(j!)^{k_j}}\frac{1}{k_j!}\left[(j-1)!\right]^{k_j}\right) \\ &= n! \sum_{k_1,\ldots,k_n,\sum k_j=k,\sum jk_j=n} \prod_{j=1}^{n}\frac{1}{j^{k_j}k_j!}. \end{aligned}$$

Indeed in the above formula,

$$\frac{n!}{\prod_{j=1}^{n}(jk_j)!}$$

is the number of possibilities of choosing the elements for the cycles of size j, j varying from 1 to n,

$$\frac{(jk_j)!}{(j!)^{k_j}}\frac{1}{k_j!}$$

is the number of ways in which one can distribute the jk_j elements in the k_j cycles of size j, and

$$\left[(j-1)!\right]^{k_j}$$

is the number of different possible orderings of the elements in the k_j cycles of size j. $\square$

We now define K_n to be the number of different types oberved in a sample of size n, or equivalently the number of different colours in Hoppe's urn after n steps. Then

$$K_n = X_1 + \cdots + X_n,$$

where

$$X_k = \mathbf{1}_{A_k}, \quad A_k = \{\text{the black ball is drawn at the } k\text{–th step}\},$$

consequently the events $A_1, \ldots, A_n$ are independent, with $\mathbb{P}(A_k) = \theta/(\theta + k - 1)$, $1 \leq k \leq n$. Consequently

$$\begin{aligned}
\mathbb{E}K_n &= \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \simeq \theta \log(n), \\
\mathrm{Var}(K_n) &= \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \cdot \frac{i-1}{\theta + i - 1} \simeq \theta \log(n), \\
\frac{K_n - \mathbb{E}K_n}{\sqrt{\mathrm{Var}(K_n)}} &\Rightarrow N(0,1), \quad \text{as } n \to \infty.
\end{aligned}$$

Exercise 2.2.4. *Prove the last assertion, via a characteristic function computation.*

Chapter 3

Mutations : the infintely many sites model

We now assume that each new mutation hits a new site, different from the sites hit by all other mutations. This is a reasonable assumption if the genomes under consideration are huge. A mathematical idealized model of the infinitely many sites model is to assume that the various mutations are i. i. d. random variables, all uniform on the interval $[0, 1]$. Again mutations arrise according to a Poisson process along the branches of Kingman's coalescent tree, with intensity $\theta/2$.

3.1 The number of segregating sites

Let S_n denote the number of sites in the genome where the various individuals in the sample of size n do not coincide. This is the total number of sites hit by a mutation, i. e. the total number of mutations. Conditionnally upon L_n, S_n is Poisson with parameter $\theta L_n/2$. Consequently

$$\begin{aligned}\mathbb{E}S_n &= \mathbb{E}\left[\mathbb{E}(S_n|L_n)\right] \\ &= \frac{\theta}{2}\mathbb{E}L_n \\ &= \theta\sum_{j=1}^{n-1}\frac{1}{j}.\end{aligned}$$

Let $a_n := \sum_{j=1}^{n-1} 1/j$. Watterson's estimator of θ is the unbiased estimator

$$\hat{\theta}_W = \frac{S_n}{a_n}.$$

Let us now compute the variance of $\hat{\theta}_W$. We have

$$S_n = \sum_{k=2}^{n} S_{n,k},$$

where $S_{n,k}$ is the number of mutations which hit one of the ancestors of the sample, while there were k lineages ancestral to the sample. The $S_{n,k}$'s are independent, and if T_k is the duration of time during which there were k lineages active in the genealogy of the sample, the conditional law of $S_{n,k}$, given T_k, is Poisson with parameter $\theta k T_k/2$. Now, with a_n defined as above and $b_n = \sum_{j=1}^{n-1} j^{-2}$,

$$\begin{aligned}
\mathrm{Var}(S_n) &= \sum_{k=2}^{n} \mathrm{Var}(S_{n,k}),\\
\mathbb{E}[S_{n,k}^2] &= \mathbb{E}\left[\mathbb{E}(S_{n,k}^2|T_k)\right],\\
\mathbb{E}(S_{n,k}^2|T_k) &= \left(\frac{\theta}{2}kT_k\right)^2 + \frac{\theta}{2}kT_k\\
\mathbb{E}[S_{n,k}^2] &= \frac{\theta}{k-1} + 2\left(\frac{\theta}{k-1}\right)^2,\\
\mathbb{E}(S_{n,k}) &= \frac{\theta}{k-1}\\
\mathrm{Var}(S_{n,k}) &= \frac{\theta}{k-1} + \left(\frac{\theta}{k-1}\right)^2,\\
\mathrm{Var}(S_n) &= \theta a_n + \theta^2 b_n,\\
\mathrm{Var}\left(\hat{\theta}_W\right) &= \frac{\theta}{a_n} + \theta^2\frac{b_n}{a_n^2}.
\end{aligned}$$

We see that $\mathrm{Var}\left(\hat{\theta}_W\right) \to 0$, as $n \to \infty$.

3.2 Pairwise mismatches

For $1 \le i \ne j \le n$, let Π_{ij} denote the number of mismatches between the genome i and the genome j, which is the number of mutations which has hit

either i of j, but not both jointly. Tajima's estimator is

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

We have

$$\begin{aligned} \mathbb{E}[\pi_n] &= \frac{2}{n(n-1)} \sum_{i<j} \mathbb{E}[\Pi_{ij}] \\ &= \mathbb{E}[\Pi_{12}] \\ &= \theta \mathbb{E}[T_2] \\ &= \theta, \end{aligned}$$

so that Tajima's estimator is unbiased. In order to compute te variance of π_n, let us first compute the law of Π_{12}. Π_{12} is the number of mutations on either branch 1 or 2, which happen before those two lineages coalesce. Following the lineages back in time, mutations on the two lineages happen at rate θ, and coalescence comes at rate 1. Hence at any time before the coalescence, the next event is a mutation with probability $\theta/(\theta+1)$. Consequently for $k \geq 0$,

$$\mathbb{P}(\Pi_{12} = k) = \left(\frac{\theta}{\theta+1}\right)^k \frac{1}{\theta+1}.$$

This is a geometric distribution starting at 0 (sometimes called the "shifted geometric" distribution). Standard results yield

$$\mathbb{E}[\Pi_{12}] = \theta, \quad \mathrm{Var}(\Pi_{12}) = \theta + \theta^2.$$

From this we deduce

Lemma 3.2.1. *(Tajima) We have*

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

PROOF: Note that

$$\pi_n^2 = \frac{4}{n^2(n-1)^2} \sum_{i_1<j_1} \sum_{i_2<j_2} \Pi_{i_1 j_1} \Pi_{i_2 j_2}.$$

In this double sum, there are three types of terms

$\frac{n(n-1)}{2}$ terms with $i_1 = i_2$, $j_1 = j_2$,

$n(n-1)(n-2)$ terms with $i_1 = i_2$, $j_1 \neq j_2$, or $i_1 = j_2$, or $j_1 = i_2$,

$\frac{n(n-1)(n-2)(n-3)}{4}$ terms with $\{i_1, j_1\} \cap \{i_2, j_2\} = \emptyset$.

Define with distinct indices i, j, k, ℓ

$$\begin{aligned} U_2 &= \mathbb{E}(\Pi_{ij}^2) - \theta^2, \\ U_3 &= \mathbb{E}(\Pi_{ij}\Pi_{ik}) - \theta^2, \\ U_4 &= \mathbb{E}(\Pi_{ij}\Pi_{k\ell}) - \theta^2. \end{aligned}$$

With these notations we have

$$\mathrm{Var}(\pi_n) = \frac{2}{n(n-1)} \left(U_2 + 2(n-2)U_3 + \frac{(n-2)(n-3)}{2} U_4 \right).$$

The above computations yield $U_2 = \theta + \theta^2$. Tajima's strategy consists in computing $\mathrm{Var}(\pi_3)$ and $\mathrm{Var}(\pi_4)$, and use the last formula to deduce U_3 and U_4. We refer the reader to Tajima's original paper (1983) or Durrett [7] for the details. □

We note that $\mathrm{Var}(\hat{\theta}_T) \to \frac{1}{3}\theta + \frac{2}{9}\theta^2$ as $n \to \infty$.

3.3 Tajima's D test statistics

We have seen two unbiased estimates of the same parameter θ. It is expected that the difference between those two estimates should be small. Tajima has introduced a normalized version of that difference, namely the quantity

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{e_1 S_n + e_2 S_n(S_n - 1)}},$$

where

$$\begin{aligned} e_1 &= \frac{n+1}{3a_n(n-1)} - \frac{1}{a_n^2}, \\ e_2 &= \frac{1}{a_n^2 + b_n} \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2} \right). \end{aligned}$$

The motivation for this choice of the denominator in the formula for D is that the variance of the numerator equals $e_1\theta + e_2\theta^2$, see [7].

Tajima showed that the distribution of D is close to a beta distribution. $|D| \leq 2$ should be interpreted as the fact that the data confirm that the genealogy of our sample is well represented by Kingman's coalescent. Can should be deduce if $D > 2$, and if $D < -2$?

Two extreme violations of Kingman's coalescent can be imagined. In the first one, all lineages diverged at an initial time, and evolved independently. In that case any single mutation is counted $n-1$ times in the sum of the Π_{ij}'s. Consequently

$$\hat{\theta}_T - \hat{\theta}_W = \frac{2}{n}S_n - \frac{S_n}{a_n} < 0$$

as soon as $n > 2$. Suppose now that all but the last coalescence have happened very near the present time, and that the two long branches of the tree support each $n/2$ of the lineages. Then if we assume that all mutations happen of one of the two long branches,

$$\hat{\theta}_T - \hat{\theta}_W = \frac{n}{2(n-1)}S_n - \frac{S_n}{a_n} > 0.$$

Note that departure from Kingman's coalescent can be in particular the effect of variable population size, or selection.

3.4 Two final remarks

In these short notes, we have neglected two very important aspects of population genetics

Remark 3.4.1. Selection *So far we have assumed that all mutations are neutral, i. e. that there is no advantage nor disadvantage associated to them. In the case of selective mutations (i. e. mutations which gives a selective advantage – or disadvantage – to those who carry it), the coalescent process is modified by the mutation, or in other words there is an interaction between the process of mutations and the coalescent.*

Remark 3.4.2. Recombinations *One important aspect of the genetics of most species is recombinations. The rate of recombinations for human beings is higher than the rate of mutations.*

Going back to the MRCA, besides coalescence events, we have recombination events, which means that a genome splits into two parts, each one "recombining" with a complementary part from another genome. Since our sample is small compared to the total population size, we can assume that all recombinations are done with a genome which does not contain ancestral material to the sample. Taking into account recombinations means that Kingman's coalescent tree should be replaced by an ancestral recombination graph. While there are k ancestral to the sample, recombinations happen at rate $k\rho/2$, while coalescences happen at rate $k(k-1)/2$. The number of ancestors to our sample follows a birth and death process, with birth rate $k\rho/2$ and death rate $k(k-1)/2$. This is a bit simplified, since in that way we may follow lineages which do not contain any genomic material ancestral to the sample. At any rate, this process reaches eventually 1, which means that the MRCA of the sample has been found.

Another way of describing recombinations is to note that Kingman's coalescent tree is different from one locus of the genome to another one. It is in fact possible to describe the evolution of the coalescent tree along the genome, see Leocard, Pardoux [11].

The Ewens sampling formula is still correct at any particular locus. The various allelic distributions at various loci are conditionally independent given the ancestral recombination graph, but their joint law is still unknown, except for very small samples.

Finally let us comment on the interaction between recombinations and selection. Suppose that an advantageous mutation appears at a particular locus (which we call below the "selective locus") in one individual of the population. If that mutation happens to get fixed in the population, at the end of the period of fixation (called the selective sweep), all individuals carry that same allele at the advantageous locus. Because recombinations happen during the sweep, the alleles at neutral loci may differ among individuals in the population. However, if the sweep is rather short, a certain number of alleles at neutral loci close to the selective one are identical in all individuals of the population (and identical to the particular alleles which were carried by the individual who experienced the selective mutation). This is called "genetic hitchhiking", and can be used to detect positive selection.

Bibliography

[1] David Aldous, Exchangeability and related topics, in *Ecole d'Ete St Flour 1983* Lecture Notes in Math. **1117**, 1–198, Springer 1985.

[2] Patrick Billingsley, *Convergence of probability measures*, 2d ed., Wiley 1999.

[3] Patrick Billingsley, *Probability and measures*, 3d ed. Wiley 1995.

[4] Matthias Birkner, Stochastic models from population biology, lecture notes for a course at TU Berlin, summer 2005 http://www.wias-berlin.de/people/birkner/smpb-30.6.05.pdf

[5] L. Breiman : *Probability*, Addison–Wesley, 1968. New edition SIAM 1992.

[6] P. Donnelly, T. Kurtz, A countable representation of the Fleming–Viot measure–valued diffusion, *Annals Probab.* **24**, 698–742, 1996.

[7] Rick Durrett, *Probability models for DNA sequence evolution*, Probability and its applications, Springer 2002.

[8] Fred Hoppe, Polya–like urns and the Ewens sampling formula, *J. Math. Biol.* **20**, 91–94, 1984.

[9] J. F. C. Kingman, The coalescent, *Stoch. Proc. Appl.* **13**, 235–248, 1982.

[10] Amaury Lambert, Population dynamics and random genealogies, *Stoch. Models* **24** 45–163.

[11] Stéphanie Leocard, Etienne Pardoux, Evolution of the ancestral recombination graph along the genome in case of a selective sweep, *J. Math. Biology*, in press.

[12] Russell Lyons, Yuval Peres, *Probability on trees and networks*, a book in progress, http://mypage.iu.edu/ rdlyons/prbtree/prbtree.html

[13] Etienne Pardoux, *Processus de Markov et applications*, Dunod, 2007; Engl. translation *Markov processes and applications. Algorithms, networks, genome and finance*, Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester; Dunod, Paris, 2008.

[14] Philip Protter, *Stochastic integration and differential equations*, 2nd Edition, Applications of Mathematics **21**, Springer, Berlin, 2004.

[15] Daniel Revuz, Marc Yor, *Continuous martingales and Brownian motion*, 3rd Edition Springer 1999.

[16] R. Sainudiin, K.Thornton, J. Harlow, J. Booth, M. Stillman, R. Yoshida, R. Griffiths, G. Mc Vean, P. Donnelly, Experiments with the site frequency spectrum, *Bull. Math. Biol.* Online First, 23 December 2010.