1 Selection

We have defined *nucleotide diversity*, denoted by π, as the proportion of nucleotides that differ between two randomly chosen sequences. We have shown that $\mathbf{E}[\pi] = \theta = 4N_e\mu$, where μ can be estimated directly. Therefore, there is a means of estimating N_e. For example, it appears that diversity in the population of *Drosophila* is about ten times greater than diversity in human population.
Abundant species have more genetic diversity than less abundant species but the relationship is not linear. Therefore, we have to consider other phenomena :

- *Population bottleneck* : dramatic reduction of population size followed by rapid expansion,
- *Natural selection.*

1.1 Wright-Fisher Model with selection

Definition 1.1. *The* fitness *of an individual is the number of offsprings it leaves. The* fitness *of a gene is the number of copies it leaves. The* fitness *of an allele is the average fitness of genes of that allelic type.*

Definition 1.2 ((Wright-Fisher Model with selection))**.** *In a panmictic, haploïd population of constant size N, where individuals are of types a and A, if generation at time t consists of k individuals of type a and $N-k$ of type A then, according to the* Wright-Fisher Model with selection*, the generation at time $t+1$ is formed by sampling independantly with replacement with,*

$$\mathbf{P}(a \text{ sampled}) = \frac{k(1+s)}{k(1+s)+N-k}. \tag{1}$$

s is called selection coefficient*. We say that a, A have* relative fitness $1+s:1$.

- *If $s > 0$: a is said to be* beneficial*,*
- *If $s < 0$: a is said to be* deterious*.*

Biologists think of an infinite pool of potential offsrings (from which new generation are sampled, proportions being dictated by (1)).

The next step is to add mutations to this model : a proportion μ_1 of the pool of a gametes mutate to A. Conversely, a proportion μ_2 of A gametes mutate to a. This leads to the following definition :

Definition 1.3 ((Wright-Fisher Model with selection and mutation))**.** *If there are k individuals of type a among parents (and $N-k$ individuals of type A), then proportion of the potential offspring that are of type a after selection and mutation is*

$$\psi_k = \frac{k(1+s)(1-\mu_1)}{k(1+s)+N-k} + \frac{(N-k)\mu_2}{k(1+s)+N-k}. \tag{2}$$

Number of offsprings is then $\sim Bin(N, \psi_k)$.

In order to get a more manageable model, we have to pass to a diffusion approximation. To obtain a non-trivial limit, we suppose that $\alpha = Ns$, $\nu_1 = N\mu_1$, $\nu_2 = N\mu_2$ and we count in units of size N.

Lemma 1.1. *As $N \to \infty$, rescaled Wright-Fisher with selection and mutation converges to a one-dimensional diffusion with drift $\mu(p) = \alpha p(1-p) - \nu_1 p + \nu_2(1-p)$, and variance $\sigma^2(p) = p(1-p)$.*

Proof. Let $\delta_t = \frac{1}{N}$ be the time between two generations (in rescaled time). As in neutral case, for all $k \geq 3$, $\mathbf{E}((p_{1/N} - p)^k | p) = O(\frac{1}{N^2})$. If current proportion of a alleles is p, the the actual number is $k \equiv Np$. We have then : $\mathbf{E}((p_{1/N} - p)|p) = \frac{1}{N}(N\psi_k - k)$. But,

$$\begin{aligned}
N\psi_k - k &= \frac{Nk(1+\frac{\alpha}{N})(1-\frac{\nu_1}{N})}{N + \frac{\alpha k}{N}} + \frac{(N-k)\nu_2}{N+\frac{\alpha k}{N}} - k, \\
&= \frac{1}{N+\frac{\alpha k}{N}}\left(Nk(1+\frac{\alpha}{N})(1-\frac{\nu_1}{N}) + (N-k)\nu_2 - kN - \frac{\alpha k^2}{N}\right), \\
&= \frac{N}{N+\frac{\alpha k}{N}}\left(\frac{\alpha k}{N} - \frac{\nu_1 k}{N} + \nu_2 - \frac{\nu_2 k}{N} - \alpha\frac{k^2}{N^2} - \alpha\frac{\nu_1 k}{N^2}\right), \\
&= \alpha p - \nu_1 p + \nu_2 - \nu_2 p - \alpha p^2 + o\left(\frac{1}{N}\right), \\
&= \alpha p(1-p) - \nu_1 p - \nu_2(1-p) + o\left(\frac{1}{N}\right).
\end{aligned}$$

Since $\psi_k = \frac{k}{N} + O\left(\frac{1}{N}\right)$,

$$\begin{aligned}
\mathbf{E}\left((p_{1/N} - p)^2 | p\right) &= \frac{1}{N^2} N\psi_k(1-\psi_k) + O\left(\frac{1}{N^2}\right), \\
&= \frac{1}{N}p(1-p) + O\left(\frac{1}{N^2}\right).
\end{aligned}$$

Now, for $u : [0,1] \longrightarrow \mathbf{R}$ sufficiently differentiable, we have by Taylor's Theorem :

$$\begin{aligned}
\frac{d}{dt}\mathbf{E}\left[u(p_t)|p_0 = p\right]_{|t=0} &\approx N\left(\mathbf{E}\left[u(p_{1/N}) - u(p)|p_0 = p\right]\right), \\
&= N\left\{u'(p)\mathbf{E}\left[(p_{1/N} - p)|p_0 = p\right] + \frac{1}{2}u''(p)\mathbf{E}\left[(p_{1/N} - p)^2|p_0 = p\right] + O\left(\frac{1}{N^2}\right)\right\}, \\
&= u'(p)\left(\alpha p(1-p) - \nu_1 p + \nu_2(1-p)\right) + \frac{1}{2}u''(p)p(1-p) + O\left(\frac{1}{N}\right).
\end{aligned}$$

The last term converges , as $N \longrightarrow \infty$ to $u'(p)\left(\alpha p(1-p) - \nu_1 p + \nu_2(1-p)\right) + \frac{1}{2}u''(p)p(1-p)$. □

Definition 1.4. *We call this diffusion the* weak solution limit.

Lemma 1.2. *Suppose there is no mutation ($\nu_1 = \nu_2 = 0$). If initial proportion of alleles is p, the probability p_{fix} that a eventually fixes in the population is,*

$$p_{fix} = \begin{cases} \frac{1-\exp(-2\alpha)}{1-\exp(-2\alpha)} & if\ \alpha \neq 0, \\ p & if\ \alpha = 0. \end{cases} \tag{3}$$

Proof. We have previously seen that

$$p_{fix} = \frac{S(p) - S(0)}{S(1) - S(0)},$$

where S is the scale function corresponding to the diffusion found in lemma 1.1 :

$$S(x) = \int_{x_0}^{x} \exp\left(-\int_{\eta}^{y} \frac{2\mu(z)}{\sigma^2(z)} \mathrm{d}z\right) \mathrm{d}y, \tag{4}$$

and $\mu(z) = \alpha z(1-z)$, $\sigma^2(z) = z(1-z)$. That leads us to :

$$S(x) = C_1(\exp(-2\alpha x) - C_2),$$

for some constants C_1 and C_2 independant of x. The result follows.

□

Special cases :

- Deterious alleles : $s < 0$. If $|s| << 1$ and $N|s| >> 1$, $p_{fix} \approx 2|s|\exp(-2N|s|)$.
- Beneficial alleles : $s > 0$, $s << 1$, $Ns >> 1$, then $p_{fix} \approx 2s$, almost independant of population size.
- Nearly neutral alleles : if $N|s| << 1$, then a is nearly neutral and $p_{fix} \approx \frac{1}{N}$.

Summary :

- Most alleles (beneficial or deterious) are lost,
- deterious mutations are more likely to fix in small populations,
- fitness differences that are too small to be measured in a laboratory ($|s| << 1$) can still have evolutionnary impact if ($N|s| >> 1$).

We have here concentrated on genic selection. More generally in diploïd populations, different forms of selection can lead to $\mu(p) = sp(1-p)(1-2p)$.

1.2 The ancestral selection graph

To understand how it works, we use here the Moran model.

Definition 1.5. *In the Moran model for a haploïd population of size N, a rate* $\binom{N}{2}$, *a pair of individuals selected at random, one dies, the other reproduces. To incorporate selection, at additional rate s* $\binom{N}{2}$, *a another pair is chosen ; if both are the same, nothing happens ; if one is a and the other is A, A dies and a split in two. s plays the role of selection coefficient. Mutations are added as a Poisson process along the lineages.*

Lemma 1.3. *As $N \to \infty$, rescaled Moran model with selection converges to the same diffusion as in the Wright-Fisher model with selection and mutation.*

Proof. For a fixed N, the corresponding generator of Moran model with selection is given by :

$$\begin{aligned} \mathcal{L}_N f(x) &= \binom{N}{2} p(1-p)\left(f(p+\frac{1}{N}) - f(p)\right) + \binom{N}{2} p(1-p)\left(f(p-\frac{1}{N}) - f(p)\right) \\ &+ N\nu_1 p\left(f(p-\frac{1}{N}) - f(p)\right) + N\nu_2(1-p)\left(f(p+\frac{1}{N}) - f(p)\right) \\ &+ 2s\binom{N}{2} p(1-p)\left(f(p+\frac{1}{N}) - f(p)\right). \end{aligned} \tag{5}$$

That is to say, the generator corresponding to the neutral Moran model plus an extra term corresponding to the selection. We take f to be twice continuously differentiable. By Taylor's theorem, the last term of the right hand side of (5) is equal to $\alpha p(1-p)f'(p) + O\left(\frac{1}{N}\right)$. As $N \longrightarrow \infty$, $\mathcal{L}_N f(x)$ converges to $\mathcal{L}f(x)$ where :

$$\mathcal{L}f(x) = \frac{1}{2}p(1-p)f''(p) + (\nu_2 - (\nu_1+\nu_2)p)f'(p) + \alpha p(1-p)f'(p). \tag{6}$$

□

To construct ancestry of sample, we trace back neutral arrows affecting two individuals in ancestry result in coalescences as we have seen before. Potential selective events hitting individuals in ancestry lead to a bench in the ancestry.

Each individual are in $N-1$ pairs, each a hit by a potential selective event at rate $s = \frac{\alpha}{N}$. So, if there is currently k ancestral lineages, we go from k to $k-1$ at rate $\begin{pmatrix} k \\ 2 \end{pmatrix}$, and we go from k to $k+1$ at rate αk.

Definition 1.6. *The system of branching and coalescing lineages described here is called the ancestral selection graph.*

Lemma 1.4. *There is, with probability* 1*, a finite random time when the number of lineages is* 1 *for the first time.*

Remark 1.1. *The corresponding individual is called the* ultimate ancestor.

Proof. If we denote by X_t the number of lineages at time t, the corresponding embeded Markov Chain Y_n has a transition matrix P given by : $P(1,1) = 1$, $P(k,k+1) = \frac{2\alpha}{2\alpha+k-1}$, $P(k,k-1) = \frac{k-1}{2\alpha+k-1}, \forall k \geq 2$. For $k \in \mathbf{N} - \{0\}$, let $T_k := \inf\{n \in \mathbf{N}, Y_n = k\}$. We want to prove that $\mathbf{P}_k(T_1 < \infty) = 1$, for all $k \geq 1$. We have $\mathbf{P}_k(T_1 < \infty) = \mathbf{P}_k(\bigcup_N \{T_1 < T_N\}) = \lim_{N\longrightarrow\infty} \mathbf{P}_k(T_1 < T_N)$. For fixed $N \in \mathbf{N} - \{0\}$, denoting $u_k = \mathbf{P}_k(T_1 < T_N)$ and applying Markov property, we have :

$$u_k = \frac{k-1}{2\alpha+k-1}u_{k-1} + \frac{2\alpha}{2\alpha+k-1}u_{k+1}, \text{for } 2 \leq k \leq N-1,$$

and

$$u_1 = 1, u_N = 0.$$

Then for all $k \geq 2$, $u_k = \beta_k \dots \beta_2$, where β is given by $u_{N-l} = \beta_{N-l}u_{N-l-1}, 1 \leq l \leq N-2$. It is clear that $\beta_k \longrightarrow 1$ as $N \longrightarrow \infty$. So, $u_k \longrightarrow 1$, as $N \longrightarrow \infty$. □

If mutation rates ν_1, ν_2 are strictly positive, then diffusion describing allele frequencies has a stationnary distribution. Indeed, in this particular case, the density m of the speed measure is given by :

$$m(x) = Ce^{2\alpha x}x^{2\nu_2-1}(1-x)^{2\nu_1-1}.$$

In particular,

$$\int_0^1 m(x)\mathrm{d}x < \infty.$$

Therefore,

$$\psi(x)\mathrm{d}x := \frac{m(x)}{\int_0^1 m(y)\mathrm{d}y}\mathrm{d}x,$$

is a stationary measure for the diffusion.
In order to resolve the potential solution events, we sample the type of the ultimate ancestor from the stationnary distribution and work back through the ancestral selection graph. Following Neuhauser and Krone, we assume that mutations rates are equal : $\nu_1 = \nu_2$.

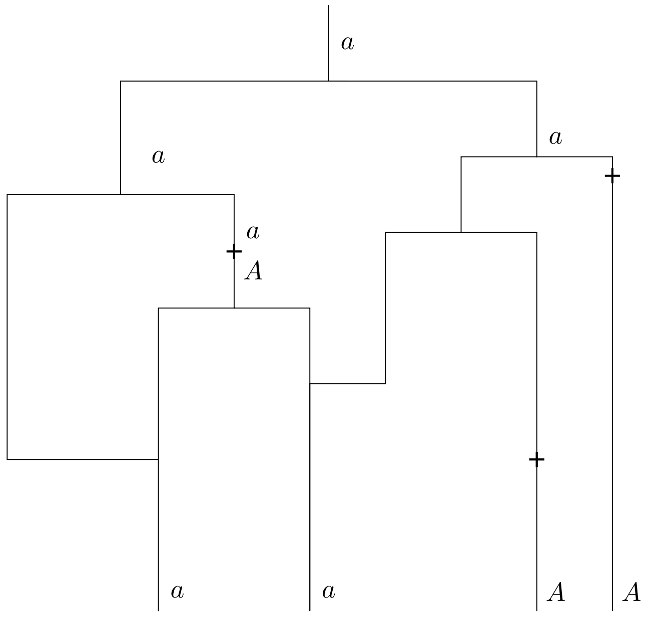

Figure 1: An example of an ancestral selection graph assuming the UA is of type a.

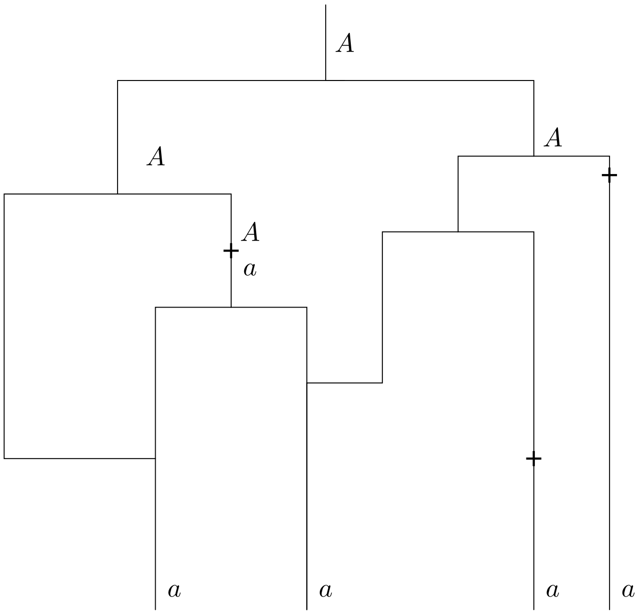

Figure 2: An example of an ancestral selection graph assuming the UA is of type A.

1.3 Adding structure to the coalescent.

In the Kingman's Coalescent, we want to discard the assumption that population size is constant. Two things led to Kingman's coalescent :

- the probability that two individuals have a common parent is $\frac{1}{N}$,
- for large N, the probability that two distinct pairs of individuals having commion parents and the probability that three or more individuals having common parents are both $O(\frac{1}{N^2})$.

We suppose now that the size of the population, t generation in the past, is $N(t)$. Then the chance p that two lineages have not coalesced by time t is such that :

$$\begin{aligned} p &= \prod_{s=1}^{t}\left(1-\frac{1}{N(s)}\right), \\ &= \exp\left(\sum_{s=1}^{t}\log\left(1-\frac{1}{N(s)}\right)\right), \\ &\approx \exp\left(-\sum_{s=1}^{t}\frac{1}{N(s)}\right), \quad \text{for } N(s) \text{ large.} \end{aligned}$$

We suppose that $N(s)$ is large and that we measure it in unit of size N (for example $M = N(0)$), and that $\frac{1}{M}N(Ms) \longrightarrow_{M\to\infty} \rho(s)$, for some nice continuous function ρ. Then, the probability q that two lineages have not coalesced by Mt is :

$$\begin{aligned} q &= \exp\left(-\sum_{s=1}^{Mt}\frac{1}{N(s)}\right), \\ &\approx \exp\left(-\sum_{s=1}^{Mt}\frac{1}{M\rho(s/M)}\right), \\ &\approx \exp\left(-\int_0^t \frac{1}{\rho(s)}\mathrm{d}s\right). \end{aligned}$$

Exactly as in derivation of Kingman's coalescent, we don't see two lineages coalescing in single generation as $M \to \infty$. So, the genealogy when population change with time in units of size M is exactly like Kingman's coalescent, except that each pair coalesces not at rate 1 but at instantaneous rate $\frac{1}{\rho(s)}$, i.e. we get a time change of Kingman's coalescent.

Secondly, we want to discard the assumption that population is panmictic. We suppose here that population is subdivided into two allelic types a, A with mutations between both. We recall here the Wright-Fisher model with mutation : if there are currently a proportion p of a alleles, then the next generation is sampled from an infinite pool of potential offsprings of which proportion $p(1-\mu_1) + (1-p)\mu_2$ are of type a and $(1-p)(1-\mu_2) + p\mu_1$ are of type A. We now suppose, that we sample n_1 individuals of type a and n_2 individuals of type A. The chance that two of the n_1 a individuals have a common parent is $\approx \left(\begin{array}{c} n_1 \\ 2 \end{array}\right)\frac{1}{N(p(1-\mu_1)+\mu_2(1-p))} = \left(\begin{array}{c} n_1 \\ 2 \end{array}\right)\frac{1}{Np} + O\left(\frac{1}{N^2}\right)$. In the same way, the chance of coalescence of two of the n_2 individuals of type A is $\frac{1}{N(1-p)} + O(\frac{1}{N^2})$. Of the type a gametes a proportion $\frac{\mu_2(1-p)}{p(1-\mu_1)+\mu_2(1-p)} = \frac{\nu_2(1-p)}{N(1-p)} + O(\frac{1}{N^2})$ arises through mutation from A gametes.

Similarly a proportion $\frac{\nu_1 p}{N(1-p)} + O(\frac{1}{N^2})$ of type A gametes arises through mutation from type a. Following a single lineage of type a, we trace back random time (which in units of size N is approximately exponential with instantaneous rate $\frac{\nu_2(1-p)}{p}$) until the ancestor's type change to A. The probability that there is mutation and coalescence of ancestral lineages in a single generation is $O(\frac{1}{N^2})$. So we do not see this event in limit as $N \longrightarrow \infty$. The process $\rho(t)$ that determines the frequency of a individuals as we trace back in time under rhe coalescent rescaling converges to time reversal of the Wright-Fisher diffusion.