Inférence en génétique des populations II.

François Rousset & Raphaël Leblois

M2 Biostatistiques 2015–2016

Outline

Likelihood calculations in the Wright-Fisher model

- A diffusion approach
- A coalescent approach

Molecular markers

Wright-Fisher model of genetic drift

- N haploid parents => gene copies (! N diploid individuals = $2N$ gene copies)
- Two alleles A,a with counts $X, N - X$
- Each parent produces a large (ideally infinite) number of juveniles
- Regulation: all juveniles compete for N breeding positions in the next generation

$$X(t+1) \sim \text{Binomial}\left[N, \pi = x(t)/N\right]$$

Mutation:

$$\mathrm{a} \underset{v}{\overset{u}{\rightleftarrows}} \mathrm{A}$$

$$\mathrm{E}(P'|P) = (1-v)P(t) + u(N - P(t))$$

$$X(t+1) \sim \text{Binomial}\left[N, \pi = \mathrm{E}(P'|P)\right]$$

Moran model of genetic drift

- N haploid parents
- Two alleles A,a with counts X, $N - X$
- Each parent produces a large (ideally infinite) number of juveniles
- All juveniles compete for 1 breeding position freed by one parent

Different variances of change in allele frequency over one "event" (check).

Sampling distribution by forward approach

How to determine the allele frequency distribution $f(p)$?

- Markov chain $\Rightarrow$ standard theory applies; but the transition matrix is hard to manipulate
- Change model to make it more tractable? Moran model
- Diffusion approximation
  - The forward Kolmogorov equation

$$\frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

  where $a(p)$ and $b(p)$ are 1st and 2nd moments of change in p per unit time.
  - The approximation of the Wright-Fisher process by a diffusion process

Intuitive explanation of the forward equation

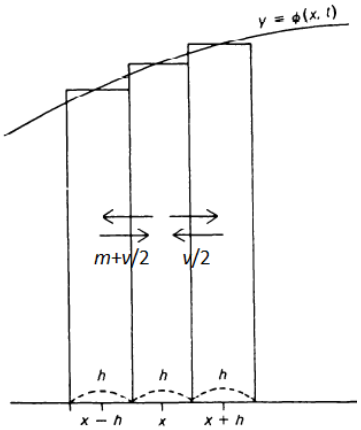


Figure 8.3.1. Diagram to show the meaning of terms in the Kolmogorov forward (Fokker–Planck) equation as applied to population genetics. (From Kimura, 1955).

$f(p)$: Probability density of allele frequency p

h: small variation of p

δ: small variation of time t

$$f(p, t+\delta)h = f(p,t)h + \frac{h\delta}{2}[v(p+h,t)f(p+h,t) - v(p,t)f(p,t)]$$
$$+ \frac{h\delta}{2}[v(p-h,t)f(p-h,t) - v(p,t)f(p,t)]$$
$$+ h\delta[m(p-h,t)f(p-h,t) - m(p,t)f(p,t)].$$

Intuitive explanation of the forward equation

$$f(p,t+\delta)h = f(p,t)h + \frac{h\delta}{2}[v(p+h,t)f(p+h,t) - v(p,t)f(p,t)]$$
$$+ \frac{h\delta}{2}[v(p-h,t)f(p-h,t) - v(p,t)f(p,t)]$$
$$+ h\delta[m(p-h,t)f(p-h,t) - m(p,t)f(p,t)].$$

$$\frac{f(p,t+\delta) - f(p,t)}{\delta} = \frac{hm(p-h,t)f(p-h,t) - hm(p,t)f(p,t)}{h}$$
$$+ \frac{1}{2}\frac{\frac{h^2 v(p+h,t)f(p+h,t) - h^2 v(p,t)f(p,t)}{h} - \frac{h^2 v(p,t)f(p,t) - h^2 v(p-h,t)f(p-h,t)}{h}}{h}$$

so that when $h \to 0$ and $\delta \to 0$

$$\frac{\partial f(p,t)}{\partial t} = -\frac{\partial M(p,t)f(p,t)}{\partial p} + \frac{1}{2}\frac{\partial^2 V(p,t)f(p,t)}{\partial p^2}$$

for $M(p,t) := hm(p,t)$ and $V(p,t) := h^2 v(p,t)$.

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.
- We approximate it as $f(p, t+\delta) \approx \int_\xi f(p, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$, hopefully ignoring only terms of order δ^2 in $f(p, t+\delta) - f(p, t)$. Then

$$
\begin{aligned}
f(p, t+\delta) &\approx \int_\xi fg - \xi \frac{\partial fg}{\partial p} + \frac{\xi^2}{2} \frac{\partial fg}{\partial p^2} + \dots \, \mathrm{d}\xi \\
&= f \int_\xi g \, \mathrm{d}\xi - \frac{\partial f \int_\xi \xi g \, \mathrm{d}\xi}{\partial p} + \frac{\partial f \int_\xi \xi^2 g \, \mathrm{d}\xi}{2 \partial p} + \dots
\end{aligned}
$$

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.
- We approximate it as $f(p, t+\delta) \approx \int_\xi f(p, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$, hopefully ignoring only terms of order δ^2 in $f(p, t+\delta) - f(p, t)$. Then

$$
\begin{aligned}
f(p, t+\delta) &\approx \int_\xi fg - \xi \frac{\partial fg}{\partial p} + \frac{\xi^2}{2} \frac{\partial fg}{\partial p^2} + \dots \, \mathrm{d}\xi \\
&= f \underbrace{\int_\xi g \, \mathrm{d}\xi}_{1} - \frac{\partial f \int_\xi \xi g \, \mathrm{d}\xi}{\partial p} + \frac{\partial f \int_\xi \xi^2 g \, \mathrm{d}\xi}{2 \partial p} + \dots
\end{aligned}
$$

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.

$$\frac{f(p, t+\delta) - f(p,t)}{\delta} = -\frac{\partial f \int_\xi \xi g \,\mathrm{d}\xi}{\delta \partial p} + \frac{\partial f \int_\xi \xi^2 g \,\mathrm{d}\xi}{2\delta \partial p} + \dots$$

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.
- Take limit of

$$\frac{f(p, t+\delta) - f(p,t)}{\delta} = -\frac{\partial f \int_\xi \xi g \,\mathrm{d}\xi}{\delta \partial p} + \frac{\partial f \int_\xi \xi^2 g \,\mathrm{d}\xi}{2\delta \partial p} + \dots$$

  as $\delta \to 0$:

$$\frac{\partial f(p,t)}{\partial t} = -\frac{\partial M(p,t) f(p,t)}{\partial p} + \frac{1}{2}\frac{\partial^2 V(p,t) f(p,t)}{\partial p^2} + \dots$$

  for $M(p,t) := \lim_{\delta\to 0} \int_\xi \xi g \,\mathrm{d}\xi/\delta$ and $\lim_{\delta\to 0} V(p,t) := \int_\xi \xi^2 g \,\mathrm{d}\xi/\delta$.
- Assume that $\lim_{\delta\to 0} \int_\xi |\xi^3| g \,\mathrm{d}\xi/\delta = 0$. Then all '...' can be neglected

Barely more rigorous sketch of proof...

... with better definitions of M and V.

- (Chapman-Kolmogorov) $f(p, t+\delta) = \int_\xi f(p-\xi, t) g(p-\xi, \xi; t, \delta)\, \mathrm{d}\xi$ where g is transition density over time interval δ.
- Take limit of

$$\frac{f(p, t+\delta) - f(p, t)}{\delta} = -\frac{\partial f \int_\xi \xi g \,\mathrm{d}\xi}{\delta \partial p} + \frac{\partial f \int_\xi \xi^2 g \,\mathrm{d}\xi}{2\delta \partial p} + \dots$$

  as $\delta \to 0$:

$$\frac{\partial f(p, t)}{\partial t} = -\frac{\partial M(p, t) f(p, t)}{\partial p} + \frac{1}{2}\frac{\partial^2 V(p, t) f(p, t)}{\partial p^2} + \dots$$

  for $M(p, t) := \lim_{\delta \to 0} \int_\xi \xi g \,\mathrm{d}\xi / \delta$ and $\lim_{\delta \to 0} V(p, t) := \int_\xi \xi^2 g \,\mathrm{d}\xi / \delta$.
- Assume that $\lim_{\delta \to 0} \int_\xi \xi^4 g \,\mathrm{d}\xi / \delta = 0$. Then all '...' can be neglected

Allele frequency distribution by diffusion approach

WF, Moran: discrete processes on allele frequency. We wish to approximate the distribution of allele frequency by a probability density $f(p, \tau)$ obtained by solution of the forward Kolmogorov equation

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

Let $X(t)$ be the number of copies of a given allele at generation t,
$X(t)/N$ is allele frequency;
$X([N\tau])/N$ is allele frequency at generation $N\tau$ (where $[x]$ denotes the greatest integer less than x);
$X([N(\tau + 1/N)])/N$ is allele frequency at generation $N\tau + 1$ or at time $(\tau + 1/N)$ for τ in units of N generations.
$\{X([N\tau] + 1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

where for the Wright-Fisher model (the drift coefficient) $a(p)$ represents expected change in one unit of τ (N generations)

$$a(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\frac{X([N\tau]+1) - X([N\tau])}{N}\middle|\, \frac{X([N\tau])}{N} = p\right]$$

and (the diffusion coefficient) $b(p)$ represents the second moment

$$b(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\left(\frac{X([N\tau]+1) - X([N\tau])}{N}\right)^2\middle|\, \frac{X([N\tau])}{N} = p\right].$$

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

where for the Wright-Fisher model (the drift coefficient) $a(p)$ represents expected change in one unit of τ (N generations)

$$a(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\frac{X([N\tau]+1) - X([N\tau])}{N}\middle|\, \frac{X([N\tau])}{N} = p\right]$$

and (the diffusion coefficient) $b(p)$ represents the second moment

$$b(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\left(\frac{X([N\tau]+1) - X([N\tau])}{N}\right)^2\middle|\, \frac{X([N\tau])}{N} = p\right].$$

Here $b(p) = ...?$ and $a(p) = ...?$

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

where for the Wright-Fisher model (the drift coefficient) $a(p)$ represents expected change in one unit of τ (N generations)

$$a(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\frac{X([N\tau]+1) - X([N\tau])}{N}\middle|\, \frac{X([N\tau])}{N} = p\right]$$

and (the diffusion coefficient) $b(p)$ represents the second moment

$$b(p) = \lim_{N\to\infty} N\,\mathrm{E}\left[\left(\frac{X([N\tau]+1) - X([N\tau])}{N}\right)^2\middle|\, \frac{X([N\tau])}{N} = p\right].$$

Here $b(p) = p(1-p)$ and $a(p) = N(-vp + u(1-p))$

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

and for the Moran model?

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

and for the Moran model?

$Var(\Delta p) = 2pq/N^2$ over one individual replacement; or "diffusion rate" is $2pq/N$ over one "generation" of N replacements;

then equivalent to WF model with population size $N/2$;

$b(p) = p(1-p)$ and $a(p) = N/2(-vp + u(1-p))$ over $N/2$ "generations"

Allele frequency distribution by diffusion approach

$$\frac{\partial f(p,\tau)}{\partial \tau} = -\frac{\partial a(p) f(p,\tau)}{\partial p} + \frac{\partial^2 b(p) f(p,\tau)}{2\partial p^2}$$

$\{X([N\tau]+1) - X([N\tau])\}/N$ is change in allele frequency over one generation ($1/N$ units of τ).

$X_N([t/c_N])/N \rightarrow$ diffusion process $Y(\tau)$ characterized by $a(p)$ and $b(p)$ in one unit of c_N generations.

Stationary distribution

- Stationarity:

$$0 = \frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

Stationary distribution

- Stationarity:

$$0 = \frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

- $$f(p) \propto \frac{exp\left(2\int^p a(x)/b(x)\,\mathrm{d}x\right)}{b(p)}$$

Stationary distribution

- Stationarity:

$$0 = \frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

- $$f(p) \propto \frac{exp\left(2\int^p a(x)/b(x)\,\mathrm{d}x\right)}{b(p)}$$

- (this is heuristic, in particular ignoring complications at the boundaries $p = 0$, $p = 1$).

Stationary distribution

- Stationarity:

$$0 = \frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

- $$f(p) \propto \frac{exp\left(2\int^p a(x)/b(x)\,\mathrm{d}x\right)}{b(p)}$$

- Frequency
  $P \sim \mathrm{Const} \quad p^{2Nu-1}(1-p)^{2Nv-1} = \mathrm{Beta}(\alpha = 2Nu, \beta = 2Nv)$

Sampling distribution

- Sample of size n

$$P_{\text{sample}} \sim \int \text{Const} \quad x^{2Nu-1}(1-x)^{2Nv-1}\text{Binomial}\left[n, \pi = x\right] \mathrm{d}x$$

$$= \text{Beta-Binomial distribution}$$

Sampling distribution

- Sample of size n

$$P_{\text{sample}} \sim \int \text{Const} \quad x^{2Nu-1}(1-x)^{2Nv-1}\text{Binomial}\left[n, \pi = x\right] \mathrm{d}x$$

$$= \text{Beta-Binomial distribution}$$

$$\text{Var}(P) = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{1+\alpha+\beta}$$

$$\text{Var}(P_{\text{sample}}) = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{n}\frac{n+\alpha+\beta}{1+\alpha+\beta}$$

Sampling distribution

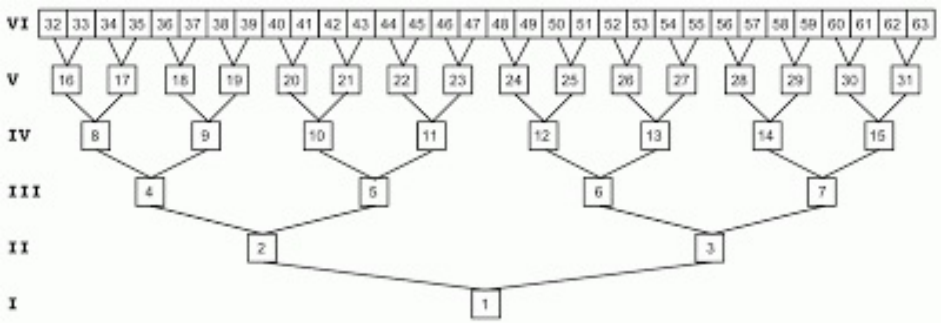
- Sample of size n

$$P_{\text{sample}} \sim \int \text{Const} \quad x^{2Nu-1}(1-x)^{2Nv-1}\text{Binomial}\left[n, \pi = x\right] \mathrm{d}x$$
$$= \text{Beta-Binomial distribution}$$

- Few models have such an explicit solution.
  In the sequel we are mostly concerned with cases where no distribution of P is known from which the distribution of P_{sample} could be derived in this way.

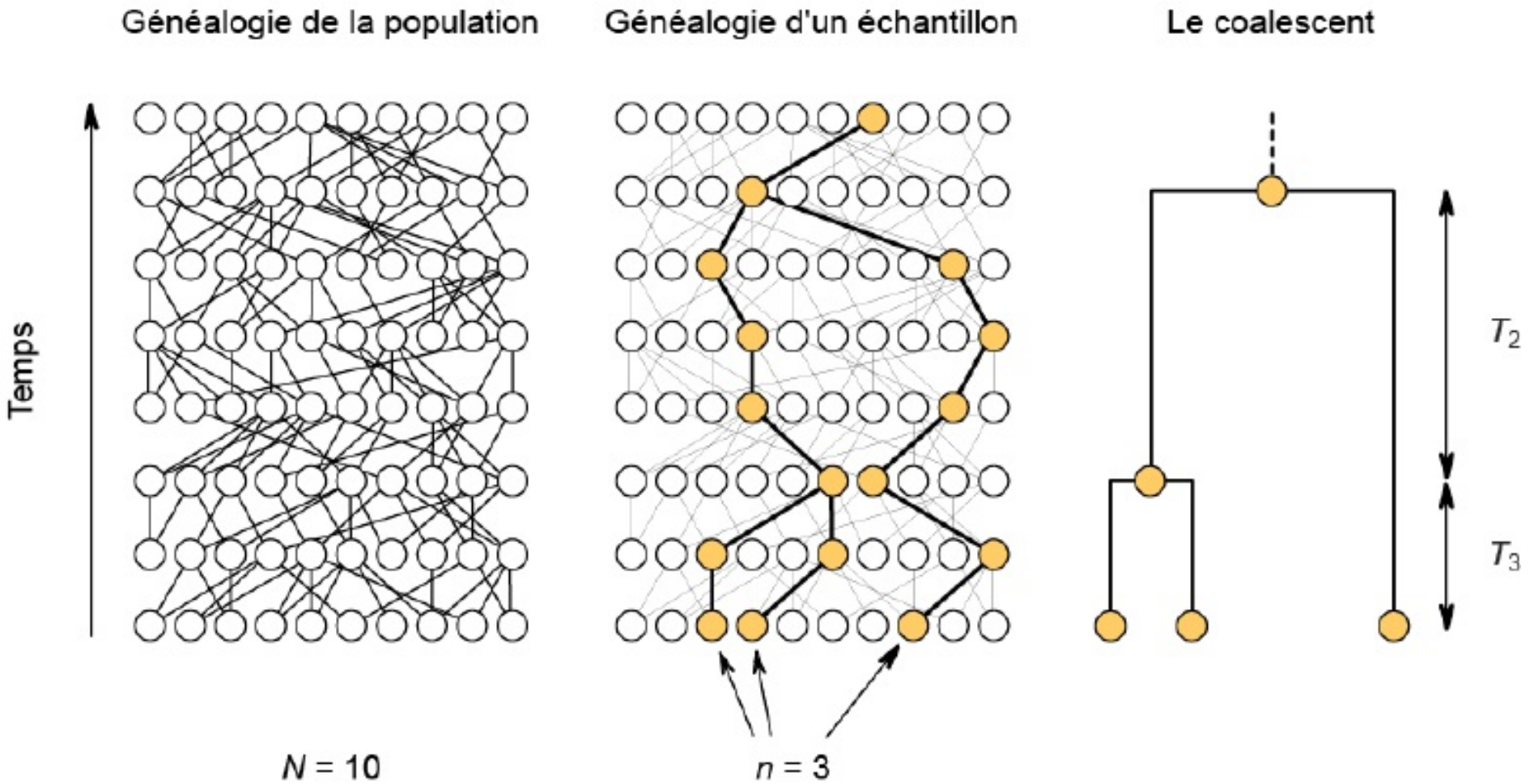
Development of the backward approach

Conventional genealogical tree vs ancestral gene tree



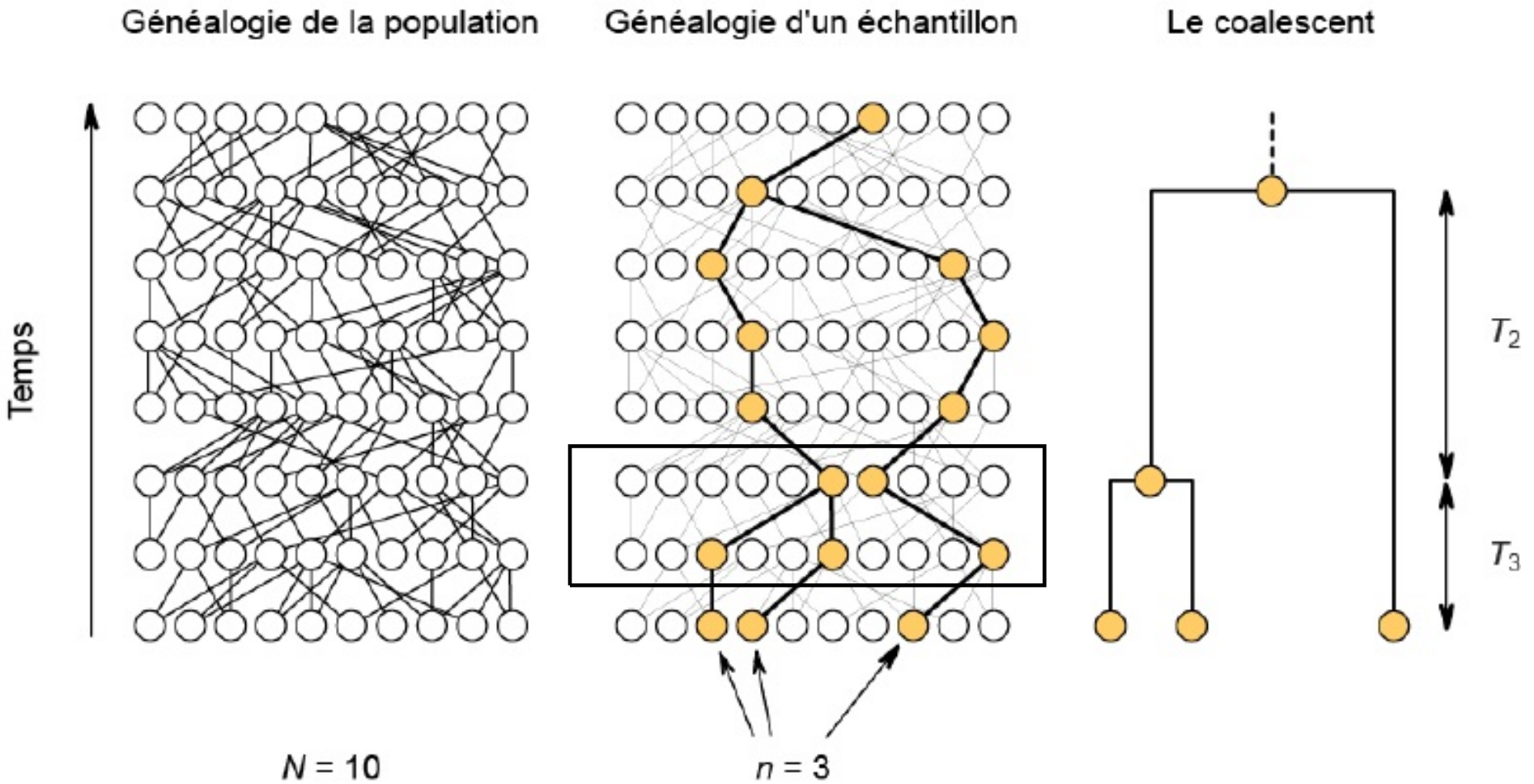
Development of the backward approach

Conventional genealogical tree vs ancestral gene tree



Development of the backward approach

Conventional genealogical tree vs ancestral gene tree



Infinite allele model

Each mutation produces an allele not pre-existing in the population

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)
- Application of coalescent arguments;

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)
- Application of coalescent arguments;

●● must come from ●, not from ●●

What are the possible ancestral types of ●●●?

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)
- Application of coalescent arguments;

●● must come from ●, not from ●●
What are the possible ancestral types of ●●●?
A sample of n genes is described by the numbers a_j of alleles found in j copies
$\mathbf{a} = (a_1 = 1, a_2 = 1, a_3 = 0)$ describes ●●● (that is, ●●●)

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)
- Application of coalescent arguments;

●● must come from ●, not from ●●
What are the possible ancestral types of ●●●?
A sample of n genes is described by the numbers a_j of alleles found in j copies
$\mathbf{a} = (a_1 = 1, a_2 = 1, a_3 = 0)$ describes ●●● (that is, ●●●)
Let $P_n(\mathbf{a})$ be the stationary proba. of sample $\mathbf{a}$ given sample size n;

Infinite allele model

Each mutation produces an allele not pre-existing in the population

- No stationary distribution of allele frequency!
- Specific diffusion tools, not considered here, were developed (frequency spectrum: probability that an existing allele is in some frequency range $p, p + \mathrm{d}p$)
- Application of coalescent arguments;

●● must come from ●, not from ●●
What are the possible ancestral types of ●●●?
A sample of n genes is described by the numbers a_j of alleles found in j copies
$\mathbf{a} = (a_1 = 1, a_2 = 1, a_3 = 0)$ describes ●●● (that is, ●●●)
Let $P_n(\mathbf{a})$ be the stationary proba. of sample $\mathbf{a}$ given sample size n;
what do $P_2(2, 0)$ and $P_2(0, 1)$ represent?

Constructing recurrences on samples

$$P_n(\mathbf{a}) = \sum_{\text{tree} \in \mathcal{T}(n)} \mathbb{P}(\text{tree}) \mathbb{P}(\mathbf{a} | \text{tree}, n) = \sum_{\text{tree} \in \mathcal{T}(n)} \mathbb{P}(\text{tree}) \pi(\mathbf{a} | \text{tree})$$

where tree is the genealogical history of the sample (but not mutation history); and π denotes probabilities in the process of adding allele types to a given tree. Over a small time interval δ:

$$P_n(\mathbf{a}) = \sum_{n_\delta} \mathbb{P}(n_\delta | n) \sum_{\mathbf{a}_\delta} \sum_{\text{tree}_\delta \in \mathcal{T}(n_\delta)} \mathbb{P}(\text{tree}_\delta) \pi(\mathbf{a}_\delta | \text{tree}_\delta) \pi(\mathbf{a} | \text{tree}_\delta, n, \mathbf{a}_\delta)$$

where $\mathbf{a}_\delta$ is the state of the ancestral sample after adding mutation to tree_δ, a random tree of a sample of size n_δ from the MRCA up to time δ.

$$P_n(\mathbf{a}) = \sum_{n_\delta} \sum_{\mathbf{a}_\delta} \mathbb{P}(n_\delta | n) P_{n_\delta}(\mathbf{a}_\delta) \pi(\mathbf{a} | n, \mathbf{a}_\delta)$$

Constructing recurrences on samples

This includes the case $\mathbf{a}_\delta = \mathbf{a}$ (no event occurred) on the RHS, leading to

$$P_n(\mathbf{a})[1 - \mathbb{P}(\text{no event})] = \sum_{n_\delta} \sum_{\mathbf{a}_\delta \neq \mathbf{a}} \mathbb{P}(n_\delta | n) P_{n_\delta}(\mathbf{a}_\delta) \pi(\mathbf{a} | n, \mathbf{a}_\delta).$$

Then $\mathbb{P}(n_\delta|n)/[1 - \mathbb{P}(\text{no event})] =: p(\mathbf{a})$ denotes the probability that the first event in the ancestry is a coalescence ($n_\delta = n - 1$), or a mutation ($n_\delta = n$). This gives a new recurrence over the time step of such a first event:

$$P_n(\mathbf{a}) = \sum_{n_\delta} \sum_{\mathbf{a}_\delta \neq \mathbf{a}} p(\mathbf{a}) P_{n_\delta}(\mathbf{a}_\delta) \pi(\mathbf{a} | n, \mathbf{a}_\delta).$$

Constructing recurrences on samples

- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability ...?

Constructing recurrences on samples

- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability $\mu/[\mu + (n-1)/2N] = \theta/(\theta + n - 1)$ for $\theta = 2N\mu$.

Constructing recurrences on samples

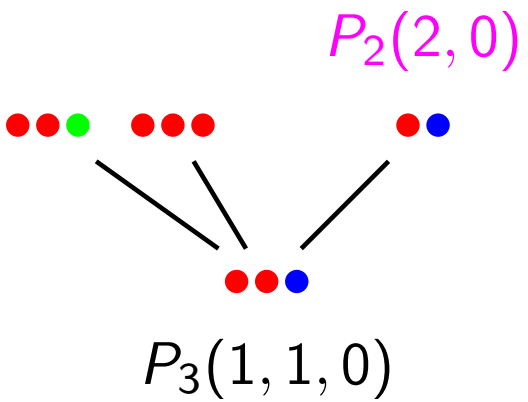
- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability $\mu/[\mu + (n-1)/2N] = \theta/(\theta + n - 1)$ for $\theta = 2N\mu$.
  Let us consider the probability $P_3(a_1 = 1, a_2 = 1, a_3 = 0)$ i.e. for sample ●●●

Constructing recurrences on samples

- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability $\mu/[\mu + (n-1)/2N] = \theta/(\theta + n - 1)$ for $\theta = 2N\mu$.
  Let us consider the probability $P_3(a_1 = 1, a_2 = 1, a_3 = 0)$ i.e. for sample ●●●
- $$P_3(1,1,0) = \frac{\theta}{2+\theta} \qquad + \frac{2}{2+\theta}$$

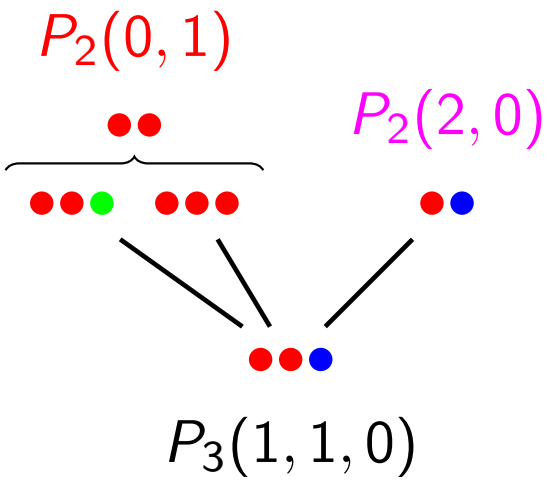
Constructing recurrences on samples

- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability $\mu/[\mu + (n-1)/2N] = \theta/(\theta + n - 1)$ for $\theta = 2N\mu$.
  Let us consider the probability $P_3(a_1 = 1, a_2 = 1, a_3 = 0)$ i.e. for sample ●●●
- $$P_3(1,1,0) = \frac{\theta}{2+\theta} + \frac{2}{2+\theta} P_2(2,0)$$



Constructing recurrences on samples

- First consider the rates of competing events affecting ancestry of n lineages: mutation at rate $n\mu$, coalescence at rate $n(n-1)/(2N)$. The first event is a mutation with probability $\mu/[\mu + (n-1)/2N] = \theta/(\theta + n - 1)$ for $\theta = 2N\mu$.
  Let us consider the probability $P_3(a_1 = 1, a_2 = 1, a_3 = 0)$ i.e. for sample ●●●
- $$P_3(1,1,0) = \frac{\theta}{2+\theta} P_2(0,1) + \frac{2}{2+\theta} P_2(2,0)$$



Sampling distribution in this model

General recursion at stationarity [with $\mathbf{e}_j = (0, \cdots, 0, 1^{j\text{th}}, 0, \cdots, 0)$]:

$$P_n(\mathbf{a}) = \frac{\theta}{n-1+\theta} P_{n-1}(\mathbf{a}-\mathbf{e}_1) + \frac{n-1}{n-1+\theta} \sum_{a_{j+1}>0} \frac{j(a_j+1)}{n-1} P_{n-1}(\mathbf{a}+\mathbf{e}_j-\mathbf{e}_{j+1})$$

where when a coalescence occurs, the descendant sample has $(\cdots, a_j, a_{j+1}, \cdots)$ hence ancestral one has $(\cdots, a_j+1, a_{j+1}-1, \cdots)$ and the probability that one of the a_j+1 alleles with j gene copies is chosen to duplicate is $j(a_j+1)/(n-1)$.
This recursion (with $P_1(1) = 1$) has a known solution: Ewens' (1972) sampling formula:

Sampling distribution in this model

General recursion at stationarity [with $\mathbf{e}_j = (0, \cdots, 0, 1^{j\text{th}}, 0, \cdots, 0)$]:

$$P_n(\mathbf{a}) = \frac{\theta}{n-1+\theta} P_{n-1}(\mathbf{a}-\mathbf{e}_1) + \frac{n-1}{n-1+\theta} \sum_{a_{j+1}>0} \frac{j(a_j+1)}{n-1} P_{n-1}(\mathbf{a}+\mathbf{e}_j-\mathbf{e}_{j+1})$$

where when a coalescence occurs, the descendant sample has $(\cdots, a_j, a_{j+1}, \cdots)$ hence ancestral one has $(\cdots, a_j+1, a_{j+1}-1, \cdots)$ and the probability that one of the a_j+1 alleles with j gene copies is chosen to duplicate is $j(a_j+1)/(n-1)$.
This recursion (with $P_1(1) = 1$) has a known solution: Ewens' (1972) sampling formula:

$$P_n(\mathbf{a}) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}$$

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

Distribution of number of alleles

- Recursion for number of alleles

$$P(K_n = k) = \frac{n-1}{n-1+\theta} P(K_{n-1} = k) + \frac{\theta}{n-1+\theta} P(K_{n-1} = k-1).$$

In other words, the probability that the nth gene is of a new type not represented in the first $n-1$ genes drawn is $\theta/(n-1+\theta)$.

Distribution of number of alleles

- Recursion for number of alleles

$$P(K_n = k) = \frac{n-1}{n-1+\theta} P(K_{n-1} = k) + \frac{\theta}{n-1+\theta} P(K_{n-1} = k-1).$$

  In other words, the probability that the nth gene is of a new type not represented in the first $n-1$ genes drawn is $\theta/(n-1+\theta)$.

- This recurrence has solution

$$P(K_n = k) = \frac{\theta^k}{\theta_{(n)}} S(n,k) = \frac{S(n,k)\theta^k}{\sum_{k=1}^{n} S(n,k)\theta^k}$$

  where $S(n,k)$ is the Stirling number of the first kind.

Stirling numbers of the first kind?

$S(n,k)$ is the coefficient of θ^k in the expansion
$\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1) = \sum_k S(n,k)\theta^k$
Thus, one can write

$$1 = \frac{\sum_k S(n,k)\theta^k}{\theta_{(n)}} = \prod_{j=1}^{n}\left(\frac{\theta}{\theta+j-1} + \frac{j-1}{\theta+j-1}\right)$$

and interpret the coefficient $S(n,k)$ as the sum of all terms that result from taking k times in the product a term of the form $\theta/(\theta+j-1)$ and $n-k$ times a term of the form $(j-1)/(\theta+j-1)$.

Then, according to the previous fact that $\theta/(k-1+\theta)$ is the probability that an additional gene is of a new type not represented in the previous k genes, $S(n,k)$ is the probability that there are k alleles in the sample.

The likelihood of θ is a function of the number of alleles

$$P_n(\mathbf{a}) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}$$

and

$$P(K_n = k) = \frac{\theta^k}{\theta_{(n)}} S(n,k)$$

imply that

$$P_n(\mathbf{a}|K_n = k) = \frac{n!}{S(n,k)} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}$$

is independent of θ.

The likelihood of θ is a function of the number of alleles

$$P_n(\mathbf{a}) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}$$

and

$$P(K_n = k) = \frac{\theta^k}{\theta_{(n)}} S(n,k)$$

imply that

$$P_n(\mathbf{a}|K_n = k) = \frac{n!}{S(n,k)} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}$$

is independent of θ.
K is sufficient for θ and $P_n(\mathbf{a}|K_n = k)$ may serve to construct a goodness-of-fit test for the WF, IAM model.

Inference in this model

- We reached a sampling distribution by a coalescent argument, not using "population" distributions

Inference in this model

- We reached a sampling distribution by a coalescent argument, not using "population" distributions
  - Standard likelihood methods for point estimation and confidence intervals can be applied
  - MLE of θ asymptotically Gaussian
  - Its variance is $O[1/\log(n)]$, not $O(n)$

Inference in this model

- We reached a sampling distribution by a coalescent argument, not using "population" distributions
  - Standard likelihood methods for point estimation and confidence intervals can be applied
  - MLE of θ asymptotically Gaussian
  - Its variance is $O[1/\log(n)]$, not $O(n)$
- But we are unable to do anything similar as soon as we change the assumptions.
  Two developments:
  - coalescent arguments used in different ways (combined with stochastic algorithms)
  - Recursions for simpler properties of samples, moment methods, ABC

Kingman's (1981) n-coalescent

Motivation Extend genealogical arguments, where "for a large class of demographic models, characterized by selective neutrality and constrained population size, the stochastic structure of the genealogy does not depend on the detail of the reproductive mechanism." (Kingman)
Large class = constant population size, and no simultaneous coalescent events; i.e.
1) continuous-time limit of WF model when $N \to \infty$ and time rescaled in units of N generations.
2) More generally when family sizes are exchangeable (e.g. Moran model).

Kingman's (1981) n-coalescent

Motivation (...)

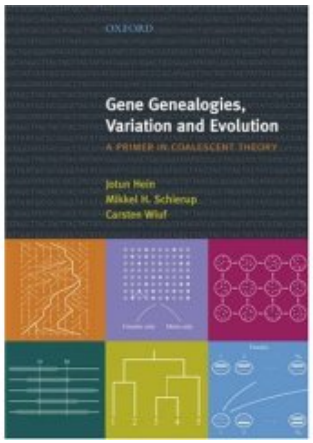
More formal definition

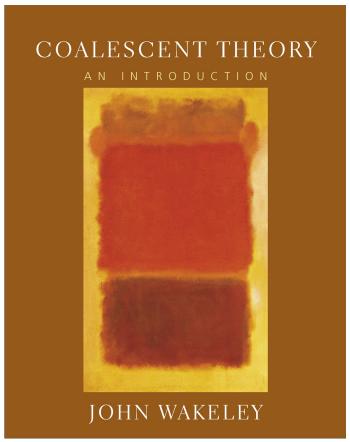
For a sample of n genes

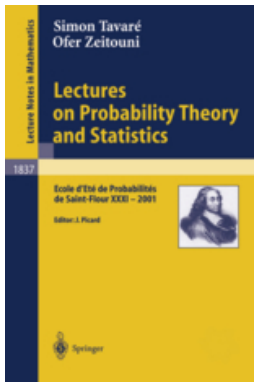
- a Markov chain whose states are equivalence relations on $\{1, 2, ..., n\}$;
- equivalence relations which contains the pair (i, j) if and only if the ith and the jth individual of this sample have a common ancestor in the rth generation;
- Let c_N be the probability that two individuals, chosen randomly without replacement from some generation, have a common ancestor one generation backwards in time;
- Then, different processes (WF, Moran) in scaled time $[t/c_N]$ converge in distribution to the coalescent process as $N \to \infty$.

References

Coalescence







(google: tavare zeitouni pdf...)

Diffusion (and more)