Coalescents: Theory and Applications

Rick Durrett

Coalescents

Moran model

- Each individual is replaced at rate 1. That is, individual \( x \) lives for an exponentially distributed amount with mean 1 and then is "replaced."
- To replace individual \( x \), we choose an individual at random from the population (including \( x \) itself) to be the parent of the new individual.

Suppose that we have two alleles \( A \) and \( a \), and let \( X_t \) be the number of copies of \( A \). The transition rates for \( X_t \) are

\[
\begin{align*}
i &\rightarrow i+1 \text{ at rate } b_i = (2N - i) \cdot \frac{i}{2N} \\
i &\rightarrow i-1 \text{ at rate } d_i = i \cdot \frac{2N - i}{2N}
\end{align*}
\]

Kingman's coalescent

**Theorem** When time is run at rate \( N \), the genealogy of a sample of size \( n \) from the Moran model converges to Kingman's coalescent.

**Proof.** If we look backwards in time, then when there are \( k \) lineages, each replacement leads to a coalescence with probability \( \frac{k-1}{2N} \). If we run time at rate \( N \), then jumps occur at rate \( N \cdot \frac{k}{2N} = \frac{k}{2} \), so the total rate of coalescence is \( k(k-1)/2 \), the right rate for Kingman's coalescent.

Directional Selection

**Fecundity selection.** Suppose \( b \)'s are born at a rate \( 1 - s \) times that of \( B \)'s.

The transition rates for \( X_t \) for the number of \( B \)'s is now:

\[
\begin{align*}
i &\rightarrow i+1 \text{ at rate } b_i = (2N - i) \cdot \frac{i}{2N} \\
i &\rightarrow i-1 \text{ at rate } d_i = i \cdot \frac{2N - i}{2N} (1 - s)
\end{align*}
\]

Embedded jump chain is a simple random walk that jumps up with probability \( p = \frac{1}{2 - s} \) and down with probability \( 1 - p \).

Started with \( X_0 = i \), \( B \) becomes fixed in the population (reaches \( 2N \)) with probability:

\[
\frac{1 - (1 - s)^i}{1 - (1 - s)^{2N}}
\]

Hitchhiking

Due to recombination, each chromosome you inherit from each parent is a mixture of their two chromosomes, with transitions between the two at points of a nonhomogeneous Poisson process.

In the absence of recombination, fixation of an allele would result in every individual in the population having a copy of the associated chromosome. With recombination, changes in allele frequency occur only near the allele that went to fixation.

Three phases of the fixation process

- While the advantageous \( B \) allele is rare, the number of \( B \)'s can be approximated by a supercritical branching process.
- While the frequency of \( B \)'s is \( \in [\epsilon, 1-\epsilon] \) there is very little randomness and it follows the solution of the logistic differential equation: \( du/dt = su(1-u) \).
- While the disadvantageous \( b \) allele is rare, the number of \( a \)'s can be approximated by a subcritical branching process.
Maynard-Smith and Haigh (1974)

Alleles $B$ and $b$ have relative fitesses 1 and 1-s, neutral locus with alleles $A$ and $a$, recombination between the two has probability $p$.

Let $p_0 = \text{frequency of } B \text{ before the sweep (1/2N)}$.

$Q_t = P(A|B)$, $R_t = P(A|b)$.

**Theorem.** Suppose $Q_0 = 0$. Under the logistic sweep model, which ignores the branching process phases 1 and 3, $Q_\infty = R_0(1 - p_0) \int_0^{2\tau} \frac{re^{-rt}}{(1 - p_0) + p_0e^{rt}} ds$

**Proof.** $R_0(1 - p_0)$ is the frequency of $A$ before the sweep. In order for a sampled individual to have the $A$ allele, its lineage must escape the sweep due to recombination.


From the previous theorem, the probability a lineage escapes from the sweep by recombination is

$p_{\text{inb}} = \int_0^{2\tau} \frac{re^{-rt}}{(1 - p_0) + p_0e^{rt}} ds$

**Theorem.** Under the logistic sweep model, if $N \to \infty$ and $r \log(2N)/s \to a$, $p_{\text{inb}} \to 1 - e^{-a}$.

Biologists rule of thumb: “hitchhiking is efficient if $r < s$ and negligible if $r \approx s$.” (should be efficient if $r \approx s/(\log(2N))$)

**Approximation 1** Let $p_{k,j}$ = probability $k$ lineages reduced to $i$ by the sweep. Under the logistic sweep model, if $N \to \infty$ with $r \ln(2N)/s \to a$ and $s(\ln N)^2 \to \infty$

then for $j \geq 2$

$$p_{k,k-j+1} \to \binom{k}{j} p^j (1 - p)^{k-j}$$

where $p = e^{-a}$

and $p_{k,k} \to (1 - p)^k + kp(1 - p)^{k-1}$.

$p$-merger. Flip coins with probability $p$ of heads for each lineage and coalesce all of those with heads. Need at least two heads to get a coalescence.

**Simulation results**

$N = 10,000, s = 0.1$. Set $r = 0.00516$ so $p_{\text{inb}} \approx 0.4$.

$p_{2\text{inb}} = P(\text{ both lineages escapes the sweep and do not coalesce}).$

$p_{2\text{cinb}} = P(\text{ both lineages escapes the sweep but coalesce}).$

$p_{1B1b} = P(\text{ one lineage escapes but the other does not}).$

$p_{22} = P(\text{ no coalescence } ) = p_{2\text{inb}} + p_{1B1b}$

<table>
<thead>
<tr>
<th>Approx.</th>
<th>0.4</th>
<th>0.16</th>
<th>0</th>
<th>0.48</th>
<th>0.64</th>
</tr>
</thead>
<tbody>
<tr>
<td>logistic ODE</td>
<td>0.39936</td>
<td>0.13814</td>
<td>0.09599</td>
<td>0.32646</td>
<td>0.46460</td>
</tr>
<tr>
<td>Moran sim</td>
<td>0.33656</td>
<td>0.10567</td>
<td>0.05488</td>
<td>0.35201</td>
<td>0.45769</td>
</tr>
<tr>
<td>Approx. 2</td>
<td>0.34065</td>
<td>0.10911</td>
<td>0.05100</td>
<td>0.36112</td>
<td>0.47203</td>
</tr>
</tbody>
</table>


Approximation 2

A stick breaking construction that leads to a coalescent with simultaneous multiple collisions.

$\begin{array}{cccccccc}
  a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\
  l_1 \\
  l_4 \\
  l_5 \\
  l_7 \\
\end{array}$

Pieces of stick are coalesced lineages that escape due to recombination. Sampled individuals = points random on (0,1). Two in the same piece coalesce. $l_4$ may be marked (×) or not (escapes sweep).
\( M = [2N_s] \) number of lineages with an infinite line of descent 
\( \xi_\ell, 2 \leq \ell \leq M \) iid Bernoulli, 1 (recombination) with prob \( r/s \).
\( W_\ell, 2 \leq \ell \leq M \) are beta\((1, \ell - 1)\) (fraction of lineages)
\( V_\ell = \xi_\ell W_\ell, T_\ell = \ell \prod_{i=\ell+1}^{M}(1 - V_i) \)
\( a_\ell = a_{\ell+1} - T_\ell, l_\ell = [a_\ell, a_{\ell+1}] \)

Error is \( O(1/\log^2 N) \) versus \( O(1/\log N) \) for approx 1
Large family sizes

The original biological motivation for Λ-coalescents is that many species have a highly variable number of offspring.

Cannings’ model Suppose that the $2N$ members of the population have offspring $(ν_1, \ldots, ν_{2N})$. The $ν_i$ are exchangeable and sum to $2N$. (Distribution depends on $N$.)

Möhle (2000). Run time at rate $2N/\text{var}(ν_i)$. Convergence to Kingman’s coalescent occurs if and only if

$$E[ν(ν−1)(ν−2)]/N^2 \to 0$$

In words, if and only if no triple mergers.


Each individual has $X_i$ offspring (independent) then $N$ are chosen to make the next generation. Part (c) of Theorem 4 shows

**Theorem.** Suppose $EX_i = μ > 1$ and $P(X_i \geq k) \sim Ck^{−α}$ with $1 < α < 2$. Then, when time is run at rate $2N/\text{var}(ν) \approx C′N^{α−1}$, the genealogical process converges to a Λ-coalescent where Λ is the beta$(2 − α, α)$ distribution, i.e.,

$$Λ(dx) \sim x^{1−α}(1−x)^{α−1}B(2−α, α)$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$, and $\Gamma(a) = \int_0^\infty x^{a−1}e^{−x}dx$ is the usual gamma function.

Genealogy when $α = 1.2$

Genealogy when $α = 1.9 \approx$ Kingman

Arnason (2004) data on cytochrome b in 1278 cod

39 mutations define 59 haplotypes (mutation patterns):
This indicates some sites were hit more than once, for if not, the number of haplotypes = 1 + the number of mutations.

Haplotypes frequencies:

696, 193, 124, 112, 29, 15, 9, 7, 6, 5(3), 4(2), 3(6), 2(7), 1(32)
Site frequency spectrum

J. Berestycki, N. Berestycki, and Schweinsberg (2006a,b).

**Theorem** Suppose we introduce mutations into the beta coalescent at rate $\theta$, and let $M_{n,k}$ be the number of mutations affecting $k$ individuals in a sample of size $n$. Then

$$\frac{M_{n,k}}{S_n} \rightarrow a_k = \frac{(2-\alpha)\Gamma(\alpha + k - 2)}{\Gamma(\alpha - 1)k!} \sim C_\alpha k^{\alpha - 3}$$

in probability as $n \rightarrow \infty$.

When $\alpha = 2$ this reduces to the $1/k$ behavior found in Kingman’s coalescent.

When $k = 1$, $a_k = 2 - \alpha$.

**Data set 2**

Boom, Boulding, and Beckenbach (1994) did a restriction enzyme digest of mtDNA on a sample of 141 Pacific Oysters from British Columbia. They found 51 segregating sites and 30 singleton mutations, resulting in an estimate of

$$\alpha = 2 - \frac{30}{51} = 1.41$$

However, this estimate is biased. If the underlying data was generated by Kingman’s coalescent, we would expect a fraction $1/\ln(141) = 0.202$ of singletons, resulting in an estimate of $\alpha = 1.8$.

**Segregating sites**

J. Berestycki, N. Berestycki, and Schweinsberg (2006a,b).

**Theorem** Suppose we introduce infinite sites mutations into the beta coalescent at rate $\theta$, and let $S_n$ be the number of segregating sites observed in a sample of size $n$. If $1 < \alpha < 2$ then

$$\frac{S_n}{n^{2-\alpha}} \rightarrow \frac{\theta \alpha(\alpha - 1)\Gamma(\alpha)}{2 - \alpha}$$

in probability as $n \rightarrow \infty$.

In Kingman’s coalescent

$$\frac{S_n}{\log n} \rightarrow \theta$$

**Simulation mean / formula : slow convergence**

Subsampling the Arnason data, $\alpha \approx 1.50$ (prev: 1.54)
Estimation results: Emilia Huerta-Sanchez

Now VIGRE postdoc, U.C. Berkeley Statistics.