A coalescent model for the effect of advantageous mutations on the genealogy of a population

by Jason Schweinsberg
University of California at San Diego

(joint work with Rick Durrett)

Outline of Talk

1. The model
2. A simple approximation
3. An improved approximation
4. Recurrent beneficial mutations
5. Applications
The model

Population has fixed size $2N$.

Consider two sites on the chromosomes:

- One site has an $A$ or $a$ allele, neither is advantageous.
- One site has a $B$ or $b$ allele, $B$ is advantageous.

At time zero, $2N - 1$ chromosomes have the $b$ allele and one has the $B$ allele.

Each individual lives for an Exponential(1) time, then is replaced.

When a new individual is born:

- The $B$ or $b$ comes from a randomly chosen parent. A replacement of a $B$ by a $b$ is rejected with probability $s$.
- With probability $1 - r$, the $A$ or $a$ comes from the same parent.
- With probability $r$, the $A$ or $a$ allele comes from a parent chosen independently at random.

Eventually, the number of $B$'s reaches 0 or $2N$. If the number of $B$'s reaches $2N$, a selective sweep occurs. The probability of a selective sweep is

$$ \frac{s}{1 - (1 - s)^{2N}} \approx s. $$

The genealogy of a sample

Sample $n$ individuals at the time $\tau$ when a selective sweep ends.

All $n$ individuals in the sample inherited their $B$ allele from the same individual at time 0.

Let $\Theta$ be a random partition of $\{1, \ldots, n\}$ such that $i$ and $j$ are in the same block of $\Theta$ if and only if the $i$th and $j$th individuals in the sample inherited their $A/a$ allele from the same individual at time zero.

Goal: to describe the distribution of the random partition $\Theta$.

$$\Theta = \{\{1, 2, 3\}, \{4\}\}.$$ 

If the $A/a$ allele of one individual comes from an individual that had the $b$ allele at time zero, we say the lineage escapes the selective sweep.
Estimating the probability $p$ of failing to escape

There is a small probability that a given lineage is affected by recombination each time there is a change in the population.

$U_k =$ number of times the number of $B$’s goes from $k$ to $k + 1$.
$D_k =$ number of times the number of $B$’s goes from $k$ to $k - 1$.
$H_k =$ number of changes that leave the number of $B$’s at $k$.

Let $p_k$, $q_k$, and $r_k$ be the probabilities of escape at times of the $U_k$, $D_k$, and $H_k$ changes respectively.

$$ R = \sum_{k=1}^{2N-1} \left( p_k E[U_k] + q_k E[D_k] + r_k E[H_k] \right), \quad p \approx e^{-R}. $$

The probability that a lineage escapes the sweep at a time when the number of $B$’s goes from $k$ to $k + 1$ is

$$ p_k \approx \frac{1}{k + 1} \cdot r \cdot \frac{2N - k}{2N}. $$

Calculate $q_k$, $r_k$ similarly. Calculate $E[U_k]$, $E[D_k]$, $E[H_k]$ using properties of conditioned random walks.

$$ p \approx \exp \left( -\frac{r}{s} \log(2N) \right) = 2N^{-r/s}. $$

- Probability of two recombinations is $O(1/(\log N)^2)$.
- Probability of coalescence and recombination is $O(1/(\log N))$.
- If $A_1, \ldots, A_n$ are the events that $n$ lineages escape the sweep, then $A_1, \ldots, A_n$ are approximately independent for large $N$. 

4
A simple approximation

Define a random partition $\Theta_p$ of $\{1, \ldots, n\}$ as follows. Flip $n$ independent coins that come up heads with probability $p$. One block of $\Theta_p$ is $\{i : \text{the } i\text{th coin is heads}\}$. The other blocks are singletons.

**Theorem 1:** Let $a = r \log(2N)/s$. Let $p = e^{-a}$. Suppose $r \leq A/(\log N)$ for some constant $A$. Then there exists a positive constant $C$ such that

$$|P(\Theta = \pi) - P(\Theta_p = \pi)| \leq \frac{C}{\log N}$$

for all $N$ and all partitions $\pi$ of $\{1, \ldots, n\}$.

For simulations: keep track of the fraction of lineages that escape the sweep. Also, we have the following possibilities for two lineages:
Simulation results

Choose \( r \) so that \( 1 - e^{-a} = 0.4 \), where \( a = r \log(2N)/s \).

<table>
<thead>
<tr>
<th>( N = 10,000; \ s = 0.03 )</th>
<th>( b )</th>
<th>( B-b )</th>
<th>( BB )</th>
<th>( bb )</th>
<th>( b-b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>simulations</td>
<td>.295</td>
<td>.303</td>
<td>.553</td>
<td>.067</td>
<td>.077</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>.400</td>
<td>.480</td>
<td>.360</td>
<td>.000</td>
<td>.160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( N = 100,000; \ s = 0.03 )</th>
<th>( b )</th>
<th>( B-b )</th>
<th>( BB )</th>
<th>( bb )</th>
<th>( b-b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>simulations</td>
<td>.318</td>
<td>.352</td>
<td>.505</td>
<td>.046</td>
<td>.096</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>.400</td>
<td>.480</td>
<td>.360</td>
<td>.000</td>
<td>.160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( N = 1,000,000; \ s = 0.01 )</th>
<th>( b )</th>
<th>( B-b )</th>
<th>( BB )</th>
<th>( bb )</th>
<th>( b-b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>simulations</td>
<td>.308</td>
<td>.355</td>
<td>.515</td>
<td>.039</td>
<td>.091</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>.400</td>
<td>.480</td>
<td>.360</td>
<td>.000</td>
<td>.160</td>
</tr>
</tbody>
</table>

The approximation based on Theorem 1 is poor. The error is \( O(1/\log N) \), which is not negligible if \( N = 1,000,000 \).

Dominant source of error (Barton, 1998): a recombination soon after the beneficial mutation may cause several lineages that have already coalesced to be descended from the same individual in the \( b \) population. Then \( \Theta \) has more than one large block.
The beginning of a selective sweep

The recombinations that cause additional large blocks in $\Theta$ are those that occur when the number of $B$'s is small.

When the $B$-population is small, it is approximately a continuous-time branching process in which each individual dies at rate $1 - s$ and gives birth at rate 1.

The number of lineages with an infinite line of descent is a branching process with no deaths and births at rate $s$.

Define $0 = \tau_1 < \tau_2 < \ldots$ such that $\tau_k$ is the first time at which there are $k$ individuals with an infinite line of descent.

If there is recombination along a lineage with an infinite line of descent between times $\tau_k$ and $\tau_{k+1}$, descendants of that lineage will have a different ancestor at the beginning of the sweep than descendants of the other $k - 1$ lineages.

What fraction of the population is descended from this lineage?
Polya Urns and Branching Processes

Start with one "marked" lineage and $k - 1$ "unmarked" lineages. Mark individuals descended from the marked lineage. When there are $x$ marked individuals and $y$ unmarked,

$$P(\text{next individual is marked}) = \frac{x}{x + y}.$$  

Polya urn: start with $a$ white balls and $b$ black balls. Repeatedly draw a ball at random, then return it to the urn along with another ball of the same color. When there are $x$ white balls and $y$ black balls, $P(\text{next ball is white}) = x/(x + y)$.

Equivalent description: let $U$ have a Beta$(a, b)$ distribution. Conditional on $U$, each ball is independently white with probability $U$, black with probability $1 - U$.

Limiting fraction of marked individuals: Beta$(1, k - 1)$ distribution.

Alternative approach: Let $(X(t), t \geq 0)$ be a Yule process, so $X(0) = 1$ and each individual gives birth at rate 1. Then

$$\lim_{t \to \infty} e^{-t} X(t) = W \quad \text{a.s.},$$

where $W$ has an Exponential$(1)$ distribution.

If $W_1, \ldots, W_k$ are i.i.d. Exponential$(1)$, the limiting fraction of marked individuals is

$$\frac{W_k}{W_1 + \cdots + W_k},$$

which has the Beta$(1, k - 1)$ distribution.
A second approximation

Stick-breaking (paintbox) construction (Kingman, 1978):

Let $M = \lfloor 2N s \rfloor$. For $k = M, M - 1, M - 2, \ldots, 3, 2$, we break off a fraction $W_k$ of the interval that is left.

$W_k$ corresponds to the fraction of lineages that escape the sweep between times $\tau_k$ and $\tau_{k+1}$.

Expected number of recombinations between $\tau_k$ and $\tau_{k+1}$ is $r/s$. Assume the number is 0 or 1.

With probability $r/s$, $W_k$ has the Beta$(1, k - 1)$ distribution.
With probability $1 - r/s$, $W_k = 0$.

Let $U_1, U_2, \ldots, U_n$ be i.i.d. with the uniform distribution on $[0, 1]$. Let $\Pi$ be the random partition of $\{1, \ldots, n\}$ such that $i$ and $j$ are in the same block if and only if $U_i$ and $U_j$ are in the same subinterval.

**Theorem 2.** If $r \leq A/\log(2N)$, then there exists a constant $C$ such that for all $N$ and all partitions $\pi$ of $\{1, \ldots, n\}$, we have

$$|P(\Theta = \pi) - P(\Pi = \pi)| \leq \frac{C}{(\log N)^2}.$$
Simulation results

Choose \( r \) so that \( 1 - e^{-a} = 0.4 \), where \( a = r \log(2N)/s \).

\[
\begin{array}{l|ccccc}
N = 10,000; s = 0.03 & b & B-b & BB & bb & b-b \\
\hline
\text{simulations} & .295 & .303 & .553 & .067 & .077 \\
\text{Theorem 2} & .301 & .318 & .540 & .059 & .082 \\
\end{array}
\]

\[
\begin{array}{l|ccccc}
N = 100,000; s = 0.03 & b & B-b & BB & bb & b-b \\
\hline
\text{simulations} & .318 & .352 & .505 & .046 & .096 \\
\text{Theorem 2} & .321 & .358 & .501 & .044 & .098 \\
\end{array}
\]

\[
\begin{array}{l|ccccc}
N = 1,000,000; s = 0.01 & b & B-b & BB & bb & b-b \\
\hline
\text{simulations} & .308 & .355 & .515 & .039 & .091 \\
\text{Theorem 2} & .308 & .358 & .513 & .038 & .091 \\
\end{array}
\]

Remarks

1. Stick-breaking approximation works much better than coin tossing approximation.

2. Theorems 1 and 2 hold for "strong selection" when the selective advantage \( s \) is \( O(1) \).

3. One can also consider "weak selection" when \( s \) is \( O(1/N) \). There is diffusion limit, studied by Krone-Neuhauser (1997), Donnelly-Kurtz (1999), Barton-Etheridge-Sturm (2004).

4. Etheridge-Pfaffelhuber-Wakolbinger (2005) show that same approximations work in the diffusion limit, if we set \( s = \alpha/N \) and then let \( \alpha \to \infty \).
Recurrent selective sweeps

Sample \( n \) individuals at time 0. Let \( \Psi_N(t) \) be the partition of \( \{1, \ldots, n\} \) such that \( i \) and \( j \) are in the same block iff the \( i \)th and \( j \)th individuals in the sample have the same ancestor at time \(-t\).

Consider the process \( \Psi_N = (\Psi_N(Nt), t \geq 0) \), which is a coalescent process taking its values in the set of partitions of \( \{1, \ldots, n\} \).

For the ordinary Moran model (no selective sweeps), \( \Psi_N \) is Kingman's coalescent (each pair of blocks merges at rate 1).

\[ \{1, 2, 3, 4\} \]

\[ \{1, 2\}, \{3, 4\} \]

\[ \{1\}, \{2\}, \{3, 4\} \]

\[ \{1\}, \{2\}, \{3\}, \{4\} \]

The duration of a selective sweep is approximately \((2/s) \log(2N)\). With strong selection, all of the lineages that coalesce during a selective sweep do so almost instantaneously for large \( N \).

Gillespie (2000) proposed that selective sweeps happen at times of a Poisson process. If selective sweeps happen at rate \( O(N^{-1}) \), then \( \Psi_N \) converges to a coalescent with multiple collisions (Pitman (1999), Sagitov (1999)) in which many blocks can merge at once.

A better approximation can be obtained using a coalescent with simultaneous multiple collisions (Möhle-Sagitov (2001), Schweinsberg (2000)) in which many mergers can occur simultaneously.
Coalescents with multiple collisions

Markov processes, take values in partitions of \(\{1, \ldots, n\}\).

Let \(\pi\) be a partition of \(\{1, \ldots, n\}\) into blocks \(B_1, \ldots, B_j\). Let \(p \in (0, 1]\). A \(p\)-merger of \(\pi\) is obtained as follows:

- Let \(\xi_1, \ldots, \xi_j\) be i.i.d. Bernoulli\((p)\).
- Merge the blocks \(B_i\) such that \(\xi_i = 1\).

Coalescents can be described in terms of a finite measure \(\Lambda\) on \([0, 1]\). Write \(\Lambda = a\delta_0 + \Lambda_0\), where \(\Lambda_0(\{0\}) = 0\). Transitions in the \(\Lambda\)-coalescent are as follows:

- Each pair of blocks merges at rate \(a\).
- Construct a Poisson point process on \([0, \infty) \times (0, 1]\) with intensity \(dt \times p^{-2}\Lambda_0(dp)\). If \((t, p)\) is a point of this Poisson process, then a \(p\)-merger occurs at time \(t\).

When there are \(b\) blocks, let \(\lambda_{b,k}\) denote the rate of a transition in which \(k\) blocks merge into one. Then, for \(2 \leq k \leq b\),

\[
\lambda_{b,k} = \int_0^1 p^{k-2}(1 - p)^{b-k} \Lambda(dp).
\]

Examples of limiting processes:

- No selection: \(\Lambda = \delta_0\) (Kingman’s coalescent).
- If the mutations all occur at one site, then \(\Lambda = \delta_0 + \alpha p^2 \delta_p\).
- If mutations and recombinations occur uniformly along the chromosome, then \(\Lambda(dx) = \delta_0 + \beta x \, dx\).
- Other \(\Lambda\) could arise under different assumptions.
Applications

Assume mutations occur along each lineage at rate $\theta/2$.

Assume the genealogy is given by a $\Lambda$-coalescent and either:

- Case 1: $\Lambda = \delta_0 + \alpha p^2 \delta_p$
- Case 2: $\Lambda(dx) = \delta_0 + \beta x \, dx$

Let $\lambda_b$ be merger rate for the $\Lambda$-coalescent when there are $b$ blocks.

Let $G_n(b) = P($coalescent has exactly $b$ blocks at some time$)$.

1. Pairwise differences

- Let $\Delta_{i,j}$ be number of sites at which segments $i$ and $j$ differ.

- Let $\Delta_n = \binom{n}{2}^{-1} \sum_{i<j} \Delta_{i,j}$.

- $E[\Delta_n] = \theta \lambda_2^{-1}$.

2. Segregating sites

- Let $S_n$ be the number of segregating sites.

- $E[S_n] = \frac{\theta}{2} \sum_{b=2}^{n} b \lambda_b^{-1} G_n(b)$

- Kingman: $E[S_n] = \frac{\theta}{2} \sum_{b=2}^{n} b \binom{b}{2}^{-1} = \theta \sum_{b=2}^{n} \frac{1}{b-1} = \theta h_{n-1}$.

- Cases 1 and 2: $\lim_{n \to \infty} (E[S_n] - \theta h_{n-1}) = -\rho$. 
3. Number of singletons

- Let $J_n$ be number of mutations that affect exactly one lineage.
- Kingman: $E[J_n] = \theta$.
- Case 1: $E[J_n] = \theta - O((\log n)/n)$.
- Case 2: $E[J_n] = \theta - O((\log n)^2/n)$.

Tajima’s (1989) $D$-statistic:

$$D = \frac{\Delta_n - S_n/h_{n-1}}{\sqrt{a_n S_n + b_n S_n^2}}.$$

Multiple mergers reduce $\Delta_n$ by $O(1)$ and $S_n/h_{n-1}$ by $O(1/\log n)$, so $D$ will be negative, consistent with simulations of Braverman-Hudson-Kaplan-Langley-Stephan (1995) and Simonsen-Churchill-Aquadro (1995).

Fu and Li’s $D$-statistic (1993):

$$D = \frac{S_n - h_{n-1} J_n}{\sqrt{c_n S_n + d_n S_n^2}}.$$

Expected value of numerator goes to $-\rho$ as $n \to \infty$.

Standard deviation of numerator is $O(\log n)$ for Fu and Li’s $D$-statistic but $O(1)$ for Tajima’s $D$-statistic, so Tajima’s $D$-statistic should be more powerful for detecting selective sweeps.
Site Frequency Spectrum

Let $M_k$ be the number of mutations that affect $k$ lineages.

The sequence $(M_1, M_2, \ldots, M_{n-1})$ is the site frequency spectrum. $\Delta_n$, $S_n$, and $J_n$ are functions of the site frequency spectrum.

The full site frequency spectrum is needed for Fay and Wu’s (2000)

$$H = \Delta_n - \sum_{k=1}^{n-1} \frac{2k^2 M_k}{n(n-1)}.$$

Kingman’s coalescent: $E[M_k] = \frac{\theta}{k}$ for all $k$.

A single selective sweep increases the number of high-frequency and low-frequency mutants (Fay-Wu, 2000; Kim-Stephan, 2002).

Recurrent selective sweeps lead to an excess of low-frequency mutants but not high-frequency mutants (Kim, 2006).

Analytical results for cases 1 and 2 have not yet been obtained.