Asymptotics for the site frequency spectrum associated with the genealogy of a birth and death process

by Jason Schweinsberg
University of California San Diego

Joint work with Kit Curtius, Brian Johnson, Yubo Shuai

Outline of talk

1. The birth and death process
2. The site frequency spectrum
3. The coalescent point process
4. Main results
5. Applications to cancer data

A birth and death process

To model tumor growth, we consider a birth and death process:

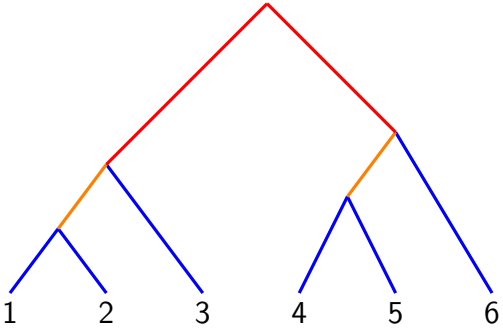
- We begin with one individual (cell) at time zero.
- Each individual independently gives birth at rate λ.
- Each individual independently dies at rate μ.
- We assume $\lambda > \mu$ and let $r = \lambda - \mu$ be the exponential growth rate.

We observe genetic data from a sample of size n taken from the population at time T.

Question: Can we infer the growth rate of a tumor from genetic data?

Branch lengths in the coalescent tree

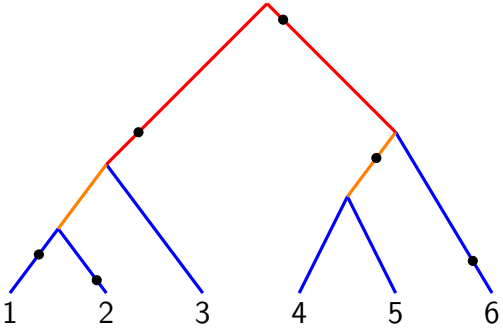
Consider the genealogical tree of a sample of n individuals.



Let L_n^{ex} = total length of external branches (blue)
Let L_n^{in} = total length of internal branches (red and orange)
Let L_n^k = total length of branches supporting k leaves
($k = 1$: blue; $k = 2$: orange; $k = 3$: red)

The site frequency spectrum

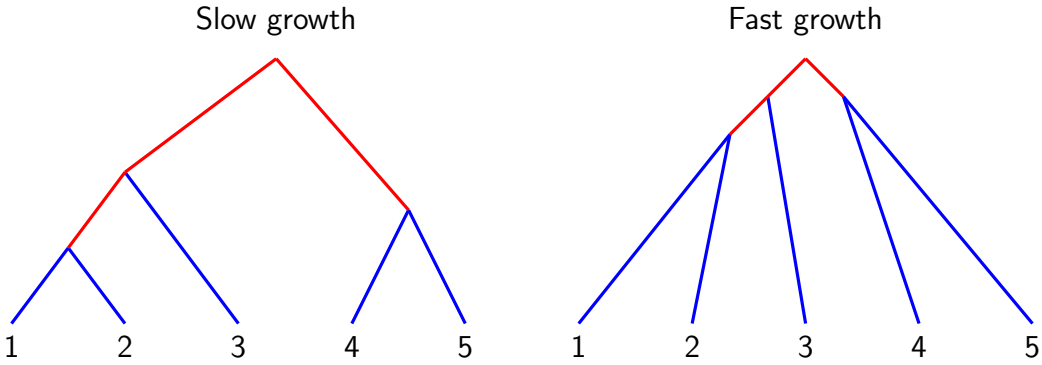
Suppose mutations appear at rate ν along each lineage.



Let M_n^k = number of mutations inherited by k sampled individuals.
The numbers $(M_n^1, \ldots, M_n^{n-1})$ are called the site frequency spectrum.
Example: $M_n^1 = 3, M_n^2 = 1, M_n^3 = 2$.
Note: conditional distribution of M_n^k given L_n^k is Poisson(νL_n^k).

A star-shaped genealogy

A faster growth rate leads to a tree with longer external branches (blue) and shorter internal branches (red).



Expected value of the site frequency spectrum

Theorem (Durrett, 2013): Fix $2 \leq k \leq n$. Then as $T \to \infty$, we have

$$E[M_n^k] \sim \frac{n\nu}{r} \cdot \frac{1}{k(k-1)}.$$

Similar calculations appeared in Bozic, Gerold, and Nowak (2016), and Williams, Werner, Barnes, Graham, and Sottoriva (2016).

Gunnarsson, Leder, and Foo (2021) calculated the exact expected site frequency spectrum when the sample consists of the entire population.

Calculating the expected site frequency spectrum

Consider a Yule process in which each individual gives birth at rate r.

Expected number of mutations while there are j individuals is $(1/jr) \cdot j\nu = \nu/r$.

The fraction of the population that inherits such a mutation has the $\mathrm{Beta}(1, j-1)$ distribution with density $(j-1)(1-x)^{j-2}$ on $[0,1]$.

The density of mutations with inherited by a fraction x of the population is

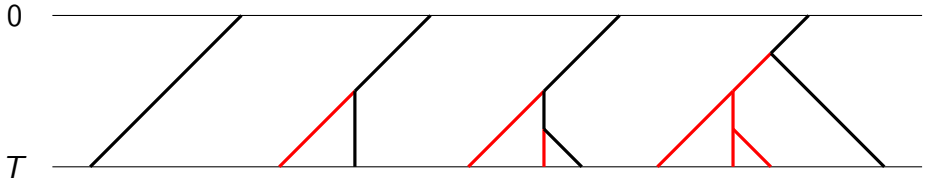
$$\sum_{j=2}^{\infty} \frac{\nu}{r}(j-1)(1-x)^{j-2} = \frac{\nu}{rx^2}.$$

In a sample of size n, for $2 \leq k \leq n-1$, the expected number of mutations affecting k individuals is therefore

$$\int_0^1 \frac{\nu}{rx^2} \cdot \binom{n}{k} x^k (1-x)^{n-k}\, dx = \frac{n\nu}{r} \cdot \frac{1}{k(k-1)}.$$

The Yule Process and the Geometric distribution

Let $(X(t), t \geq 0)$ be a Yule process started from $X(0) = 1$ in which each individual gives birth at rate r. Then $X(T) \sim \text{Geometric}(e^{-rT})$.

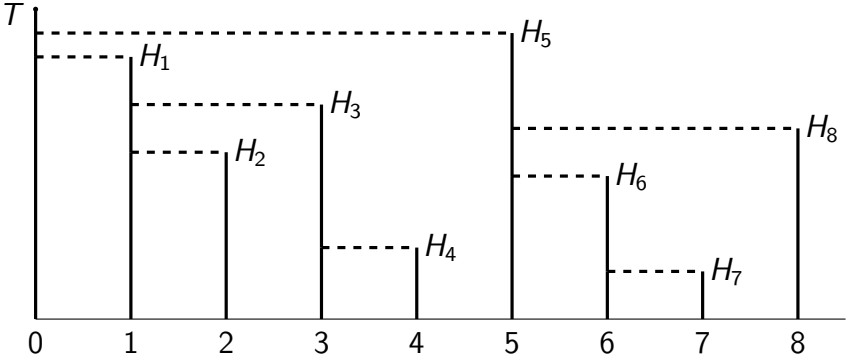


We can construct a Yule process as follows:

1. Begin with one individual at time zero, draw a line to time T.
2. Search for branchpoints, starting at time T and working backwards.
3. When a branchpoint is found, draw a line from the branchpoint to time T.
4. Repeat until there is no branchpoint.
5. At each step, the probability that there is no branchpoint is e^{-rT}.

The coalescent point process (CPP)

Goes back to Popovic (2004), Aldous and Popovic (2005).



Let $H_1, H_2, \ldots$ be i.i.d. Exponential(r).

Draw a vertical line on the left of height T. Draw vertical lines of heights $H_1, H_2, \ldots$, then draw a horizontal dashed line to the left, stopping when it hits a vertical line. Stop when we reach some $H_i > T$.

Generalizations and Consequences

Consider a branching process $(X(t), t \geq 0)$ in which:

- Individuals give birth at time t to one offspring at rate $\lambda(t)$.
- An individual born at time t has probability $q(t)$ of having a descendant alive at time T.

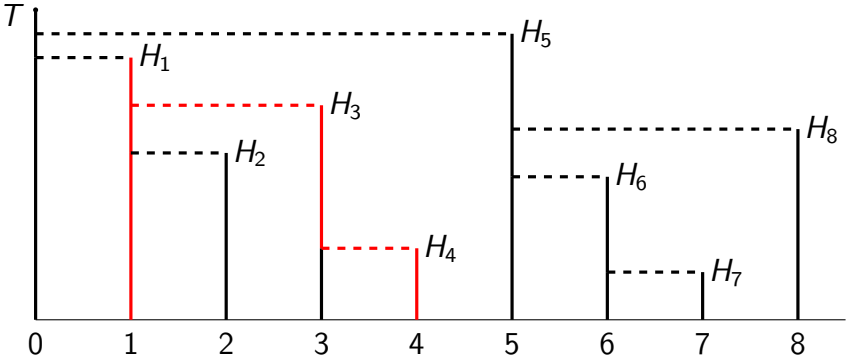
Note that the death rate (but not the birth rate) may be age dependent.

Lambert and Stadler (2013): The reduced tree (subtree consisting of individuals with a descendant alive at time T), conditioned on survival until time T, is a CPP with

$$P(H_i > t) = \exp\left(- \int_{T-t}^{T} \lambda(s) q(s)\, ds \right).$$

- We can construct the tree conditional on $X(T) = n$ by choosing $H_1^*, \ldots, H_{n-1}^*$ from the conditional distribution of H_i given $H_i < T$.
- Conditional on $X(T) = n$, the $n-1$ coalescence times are i.i.d.

Sampling each individual independently with probability y



If we sample i and j but not $i+1, \ldots j-1$, replace H_j by $\max\{H_{i+1}, \ldots, H_j\}$.

Stadler (2009), Lambert and Stadler (2013): If each individual is sampled with probability y, the genealogical tree is a CPP with H_i^y having the distribution of $\max\{H_1, \ldots, H_G\}$, where $G \sim \text{Geometric}(y)$.

A Lemma involving the Geometric distribution

Lemma (Lambert, 2018): Consider the following two sampling schemes:

- Let $N \sim \text{Geometric}(p)$. Conditional on $N \geq n$, choose a random subset S of n elements of $\{1, \ldots, N\}$. For $i \in \{1, \ldots, N\}$, let $I_i = \mathbb{1}_{\{i \in S\}}$.
- Obtain the sample in two steps:
  1. Choose Y to have density

$$f_n(y) = \frac{npy^{n-1}}{(1-(1-p)(1-y))^{n+1}}, \qquad 0 < y < 1.$$

  2. Let $M \sim \text{Geometric}(p)$, and let $J_1, J_2, \ldots, J_M$ be i.i.d. Bernoulli(y). Condition on $J_1 + \cdots + J_M = n$.

Then the distributions of $(N, (I_i)_{i=1}^N)$ and $(M, (J_i)_{i=1}^M)$ are the same, with the indicated conditioning.

Taking a sample of fixed size n

Consider a birth and death process $(X(t), t \geq 0)$ started with one individual at time 0. Each individual has birth rate λ and death rate μ, with $r = \lambda - \mu$.

Lambert (2018): The genealogical tree of a sample of size n at time T, conditioned on $X(T) \geq n$, is given by the following CPP:

1. Choose Y to have density on $(0, 1)$ given by

$$f_n(y) = \frac{npy^{n-1}}{(1-(1-p)(1-y))^{n+1}}, \qquad p = \frac{re^{-rT}}{\lambda(1-e^{-rT}) + re^{-rT}}.$$

2. Conditional on $Y = y$, let $H_1^y, \ldots, H_{n-1}^y$ be i.i.d. with density on $(0, T)$:

$$f_y(t) = \frac{y\lambda + (r - y\lambda)e^{-rT}}{y\lambda(1-e^{-rT})} \cdot \frac{y\lambda r^2 e^{-rt}}{(y\lambda + (r - y\lambda)e^{-rt})^2}.$$

Harris, Johnston, and Roberts (2020) obtained the same formula for the joint distribution of coalescence times using spines and a change of measure.

The large time and large sample limit

Let $T \to \infty$, and then let $n \to \infty$. We can approximate the genealogical tree of a sample of size n at time T, conditioned on $X(T) \geq n$, by the following CPP:

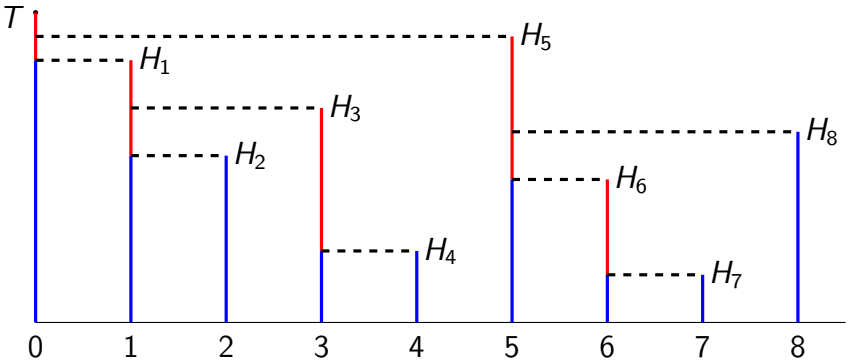
1. Let $W \sim \text{Exponential}(1)$.
2. Let $U_1, U_2, \ldots, U_{n-1}$ be i.i.d. logistic, with density on $\mathbb{R}$ given by
$$f(u) = \frac{e^u}{(1+e^u)^2}.$$
3. Let
$$H_i = T - \frac{1}{r}\Big(\log(1/W) + \log n + U_i\Big).$$

We can bound the mean difference between original H_i and this approximation.

We expect $X(t) \approx We^{rt}$ for large t, so the time it takes for the population to reach size n is approximately $\frac{1}{r}(\log(1/W) + \log n)$.

Ignatieva, Hein, and Jenkins (2020) used a random time change to show that the coalescence times are well approximated by i.i.d. logistic random variables.

Internal and External branch lengths from the CPP



For $1 \leq i \leq n-2$, we have $L_{i,n}^{in} = (H_i - H_{i+1})^+$ and $L_{i,n}^{ex} = \min\{H_i, H_{i+1}\}$.

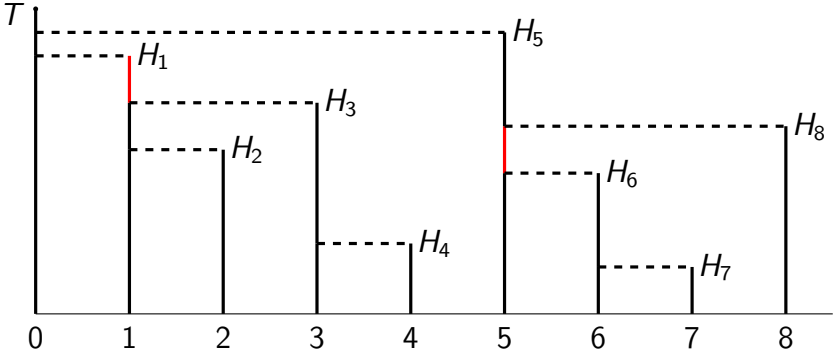
We have $L_n^{in} = \sum_{i=0}^{n-2} L_{i,n}^{in}$ and $L_n^{ex} = \sum_{i=0}^{n-1} L_{i,n}^{ex}$.

Site Frequency Spectrum from the CPP

Expected site frequency spectrum worked out by Lambert (2009).

Further asymptotics by Champagnat and Lambert (2012, 2013), Champagnat and Henry (2016), Delaporte, Achaz, and Lambert (2016).

Cancer applications: Dinh, Jaksik, Kimmel, Lambert, and Tavaré (2020).



$L^k_{i,n} = (\min\{H_i, H_{i+k}\} - \max\{H_{i+1}, \ldots, H_{i+k-1}\})^+$ for $1 \leq i \leq n-k-1$.

Approximate Moments for Internal Lengths

Recall that

$$H_i = T - \frac{1}{r}\Big(\log(1/W) + \log n + U_i\Big).$$

For $1 \le i \le n-2$,

$$L_{i,n}^{in} = (H_i - H_{i+1})^+ = \frac{1}{r}(U_{i+1} - U_i)^+.$$

We have $E[(U_{i+1} - U_i)^+] = 1$, so $E[L_n^{in}] \approx n/r$.

This agrees with Durrett's (2013) formula because

$$\frac{n}{r}\sum_{k=2}^{\infty}\frac{1}{k(k-1)} = \frac{n}{r}.$$

One can also estimate $\mathrm{Var}(L_{i,n}^{in})$ and $\mathrm{Cov}(L_{i,n}^{in}, L_{i+1,n}^{in})$ by making calculations involving two or three logistic random variables. This leads to $\mathrm{Var}(L_n^{in}) \approx n/r^2$.

Asymptotic normality follows from the m-dependent CLT.

A Limit Theorem for Internal Lengths

Consider a sequence of processes and write λ_n, μ_n, r_n, T_n.

Theorem: Suppose

$$\lim_{n\to\infty} ne^{-r_n T_n} = 0. \tag{1}$$

Then as $n \to \infty$,

$$\frac{r_n}{n} L_n^{in} \to_p 1.$$

If the stronger condition

$$\lim_{n\to\infty} n^{3/2}(\log n)e^{-r_n T_n} = 0 \tag{2}$$

holds, then as $n \to \infty$,

$$\frac{r_n}{\sqrt{n}}\left(L_n^{in} - \frac{n}{r_n}\right) \Rightarrow N(0,1).$$

A Limit Theorem for External Lengths

Recall that

$$H_i = T_n - \frac{1}{r_n}\Big(\log(1/W) + \log n + U_i \Big).$$

We have $L^{ex}_{i,n} = \min\{H_i, H_{i+1}\}$ and $E[\max\{U_i, U_{i+1}\}] = 1$.

Theorem: Suppose (1) holds. Let W have an Exponential(1) distribution. Then as $n \to \infty$,

$$\frac{r_n}{n} L^{ex}_n - \big(r_n T_n - \log n - 1 \big) \Rightarrow \log W.$$

Note: The internal and external branch lengths are asymptotically independent.

Approximate mean of site frequency spectrum

Recall that for $1 \leq i \leq n-k-1$,

$$\begin{aligned} L_{i,n}^k &= (\min\{H_i, H_{i+k}\} - \max\{H_{i+1}, \ldots, H_{i+k-1}\})^+ \\ &= \frac{1}{r}(\min\{U_{i+1}, \ldots, U_{i+k-1}\} - \max\{U_i, U_{i+k}\})^+. \end{aligned}$$

The expectation of this expression is

$$\begin{aligned} &\frac{1}{r}\int_{-\infty}^{\infty} P(U_i, U_{i+k} < x < U_{i+1}, \ldots, U_{i+k-1})\, dx \\ &\quad = \frac{1}{r}\int_{-\infty}^{\infty} \left(\frac{e^x}{1+e^x}\right)^2 \left(\frac{1}{1+e^x}\right)^{k-1} dx = \frac{1}{rk(k-1)}. \end{aligned}$$

which agrees with Durrett's (2013) formula.

The covariance computation is more complicated.
Asymptotic normality follows from the m-dependent CLT.

Asymptotics for the site frequency spectrum

Theorem: Suppose (1) holds. Fix $k \geq 2$. Then as $n \to \infty$,

$$\frac{r_n}{n} L_n^k \to_p \frac{1}{k(k-1)}.$$

Theorem: Suppose (2) holds. Fix $K \geq 2$. For $k = 2, 3, \ldots, K$, let

$$\mu_k = \frac{n}{r_n k(k-1)}.$$

As $n \to \infty$,

$$\frac{r_n}{\sqrt{n}} \left(L_n^2 - \mu_2, \ldots, L_n^K - \mu_K \right) \Rightarrow N(0, \mathbf{V}).$$

A complicated but explicit formula can be obtained for the covariances $V_{k,\ell}$.

The critical case

Suppose $\lambda = \mu = 1$. Let $N(t)$ denote the size of the population at time t.

Theorem: Suppose

$$\lim_{n\to\infty} \frac{n}{T_n} = 0.$$

Fix $K \geq 2$. For $k = 2, 3, \ldots, K$, let $\mu_k = 1/k$. As $n \to \infty$,

$$\sqrt{\frac{n}{\log n}}\left(\frac{L_n^1}{N(T_n)} - \mu_1, \ldots, \frac{L_n^K}{N(T_n)} - \mu_K\right) \Rightarrow N(0, \mathbf{I}).$$

Dahmer and Kersting (2015) proved a very similar result for Kingman's coalescent:

- Critical branching processes converge to the Feller diffusion.
- Genealogy of Feller diffusion is a time change of Kingman's coalescent (Perkins, 1992; Donnelly and Kurtz, 1999).

Site frequency spectrum in real data sets

We have data on 42 clones (20 malignant) from four papers on blood cancer: Fabre et al (2022), Williams et al (2022), Van Egeren et al (2021), Mitchell et al (2022).

We estimated L_n^k / L_n^{in} for the 42 data sets, compared to $1/(k(k-1))$.

k	Observed	Predicted
2	.472	.500
3	.173	.167
4	.081	.083
5	.063	.050
6	.046	.033
7	.034	.024
8	.020	.018
9	.015	.014
10+	.102	.110

Methods for estimating the growth rate

1. **Internal Lengths**: Recall that $E[L_n^{in}] \approx n/r$

$$\hat{r} = \frac{n}{L_n^{in}}, \qquad 95\% \text{ CI: } \left[\hat{r}\left(1 - \frac{z^*}{\sqrt{n}}\right), \hat{r}\left(1 + \frac{z^*}{\sqrt{n}}\right)\right].$$

2. **Maximum Likelihood**: Coalescence times can be approximated by

$$H_i = a + bU_i, \qquad a = T - \frac{1}{r}\Big(\log(1/W) + \log n\Big), \qquad b = \frac{1}{r}.$$

   We can estimate the scale parameter b of a logistic distribution by maximum likelihood and obtain asymptotic confidence intervals.

3. **Phylofit**: Method used by Williams et al (2022) based on MCMC. Assumes that population size evolves deterministically, lineages merge at rate $1/N(t)$.

4. **Birth-Death MCMC**: Method used by Stadler (2009). Assumes each individual in a birth and death process is sampled with probability y.

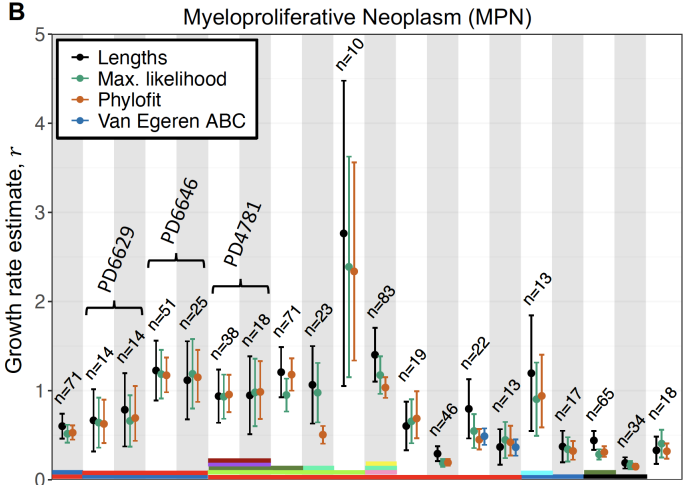
Simulation Results

$n = 100,\ T = 40,\ r = 0.5$, 500 runs

	Mean	RMSE	CI Coverage
Internal Lengths	0.508	0.102	0.948
Maximum Likelihood	0.521	0.093	0.948
Phylofit	0.514	0.101	0.798
Birth-Death MCMC	0.508	0.083	0.948

- Birth-death MCMC performs slightly better than our methods, but takes much longer to run.
- Phylofit confidence intervals are too narrow (observed for similar methods by Boskova, Bonhoeffer, and Stadler, 2014).
- All methods overestimate r when n is small.
- Our methods perform can perform poorly when r is small, but this is not a problem if $L_n^{ex}/L_n^{in} > 3$, which was true for all 42 clones.

Application to blood cancer data



The Internal Lengths and Maximum Likelihood methods perform comparably to Phylofit on data from 20 malignant clones.