The distances between tree changes under SMC for any n and under SMC' for $n = 3$

Shai Carmi

December 30, 2020

1 Introduction

The goal of this short note is to report results for the distance between tree changes under Markovian approximations of the coalescent with recombination.

2 SMC

Consider the Sequentially Markov Coalescent (SMC) model [1]. Under this model, every recombination leads to a tree change, because the lineage that broke can coalesce only with other branches of the current tree, not including itself. After coalescence, the original lineage is removed from the tree as we move along the sequence. Thus, the distance to the next tree change is simply the distance to the next recombination. Define the total tree length as T, where time is scaled by the effective population size N (in haploids), and we assume there are n samples. Given T, the sequence length to the next recombination (in Morgan) is exponential with rate NT.

Following our paper [2], define the stationary distribution of T at tree changes as $\pi(T)$, and define the transition probabilities between successive total tree lengths as $q(T|S)$, where S and T are the total tree lengths on both sides of a tree change. We can use the reversibility of SMC to write a detailed balance equation for the Markov chain that tracks T at tree changes,

$$\pi(T)q(S|T) = \pi(S)q(T|S). \tag{1}$$

(Reversibility can be seen by the fact that SMC, as well as SMC' (below), can be written as two-loci models backwards in time [3].) On the other hand, we can also consider detailed balance for the Markov chain that tracks T along equally spaced genomic positions,

$$P_c(T) \cdot NT \cdot q(S|T) = P_c(S) \cdot NS \cdot q(T|S), \tag{2}$$

where $P_c(T)$ is the distribution of T at a random site (according to the standard coalescent), and NT is the rate at which tree changes occur (here, due to any recombination events). Combining Eqs. (1) and (2), it is easy to see that

$$\pi(T) \propto NTP_c(T). \tag{3}$$

For the coalescent with a constant population size ([4], Eq. 3.34),

$$P_c(T) = \frac{n-1}{2} e^{-T/2} \left(1 - e^{-T/2}\right)^{n-2}. \tag{4}$$

After normalization,

$$\pi(T) = \frac{NT\frac{n-1}{2} e^{-T/2} \left(1 - e^{-T/2}\right)^{n-2}}{2N\sum_{k=1}^{n-1}\frac{1}{k}} = \frac{(n-1)Te^{-T/2} \left(1 - e^{-T/2}\right)^{n-2}}{4\sum_{k=1}^{n-1}\frac{1}{k}}. \tag{5}$$

Denote by ℓ the distance between tree changes. Given T, $\ell \,|\, T \sim \mathrm{Exp}(NT)$. The unconditional distribution of ℓ, $\psi(\ell)$, can be obtained by integrating over the total tree length (T),

$$\begin{aligned}
\psi(\ell) &= \int_0^\infty \pi(T) \cdot NTe^{-NT\ell} dT \\
&= \frac{2N(n-1)!\Gamma(1+2N\ell)\left[\left(H_{2N\ell} - H_{n-1+2N\ell}\right)^2 - \psi^{(1)}(n+2N\ell) + \psi^{(1)}(1+2N\ell)\right]}{\Gamma(n+2N\ell)H_{n-1}},
\end{aligned} \tag{6}$$

where Γ is the Gamma function, $\psi^{(1)}$ is the polygamma function of order 1, and H is the harmonic number.

For $n = 2$, Eq. (6) reduces to

$$\psi(\ell) = \frac{4N}{(1+2N\ell)^3}, \tag{7}$$

as we and others have previously derived [2]. For $n = 3$,

$$\psi(\ell) = \frac{2N(7+6N\ell(3+2N\ell))}{3(1+N\ell)^3(1+2N\ell)^3}. \tag{8}$$

The mean distance between tree changes is

$$\begin{aligned}
\langle \ell \rangle &= \int_0^\infty \ell d\ell \int_0^\infty \pi(T) \cdot NTe^{-NT\ell} dT \\
&= \int_0^\infty \pi(T) dT \int_0^\infty \ell \cdot NTe^{-NT\ell} d\ell \\
&= \int_0^\infty \frac{\pi(T)}{NT} dT \\
&= \left[2N \sum_{k=1}^{n-1} \frac{1}{k}\right]^{-1}.
\end{aligned} \tag{9}$$

In the last line we used Eq. (5) for $\pi(T)$ and the fact that $P_c(t)$ must be normalized.

The average number of tree changes is, where L is the total chromosome length (in Morgan),

$$\langle n_c \rangle = \frac{L}{\langle \ell \rangle} = 2NL \sum_{k=1}^{n-1} \frac{1}{k}. \tag{10}$$

This is simply N times the average size of the tree ([4], Eq. 3.23). In fact, the same result holds for the ancestral recombination graph; see corollary 2.2 in [5]. For $n = 2$, $\langle n_c \rangle = 2NL$, and for $n = 3$, $\langle n_c \rangle = 3NL$.

2.1 SMC'

The SMC' model is similar to SMC, but recombination events do not necessarily lead to tree changes [6]. Instead, the lineage that broke can coalesce with any branch in the current tree, including itself. But following coalescence, any lineage that does not contain any ancestral material is removed from the tree, similarly to SMC.

To generalize the above results to SMC', then instead of tracking just the total tree size T, we need to track the times of each coalescence event: t_n, t_{n-1}, ..., t_2, where t_i is the time of coalescence when there were i lineages, and we define $t_{n+1} \equiv 0$. We have $T = \sum_{k=2}^{n} k(t_k - t_{k+1}) = t_2 + \sum_{k=2}^{n} t_k$. The tree will change only if both a recombination event happens (at rate NT) and also the coalescence of the lineage that broke is such that the tree has changed. Thus, the rate of tree changes (as a process along the genome) is $\lambda(t_2, \ldots, t_n) \equiv NT \cdot P(\text{tree change} \,|\, t_2, \ldots, t_n)$. Following the same logic as for SMC (Eqs. (1) and (2)), the stationary distribution of the coalescence times at tree changes is

$$\pi(t_2, \ldots, t_n) \propto P_c(t_2, \ldots, t_n) \lambda(t_2, \ldots, t_n), \tag{11}$$

where the distribution of coalescence times at random sites is

$$P_c(t_2, \ldots, t_n) = \Pi_{k=2}^{n} \frac{k(k-1)}{2} e^{-\frac{k(k-1)}{2}(t_k - t_{k+1})}. \tag{12}$$

To compute $P(\text{tree change} \,|\, t_2, \ldots, t_n)$, let us derive the complementary probability $P(\text{tree unchanged} \,|\, t_2, \ldots, t_n)$. (In the following, we suppress the conditioning on $t_2, \ldots, t_n$, for brevity.) We condition this probability on when the breakpoint occurred, assuming that any position along the tree is equally likely to recombine. The breakpoint will occur at a time when the number of remaining lineages is k with probability $\frac{k(t_k - t_{k+1})}{T}$. The precise position of the breakpoint along the branch is then uniform with probability $1/(t_k - t_{k+1})$.

In this note, we consider only the case of $n = 3$, which has the advantage of a single tree topology, shown and annotated in Figure 1.

For $n = 3$, $T = 2t_2 + t_3$, and

$$P_c(t_2, t_3) = 3e^{-3t_3} e^{-(t_2 - t_3)} = 3e^{-(2t_3 + t_2)}. \tag{13}$$

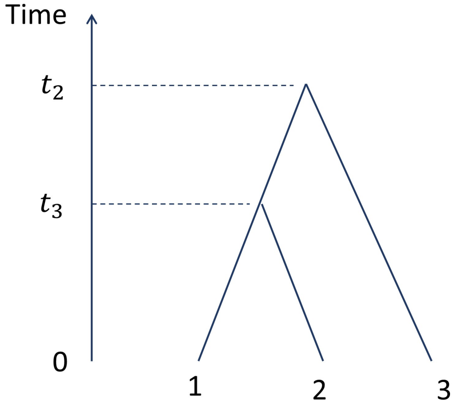

Figure 1: The tree topology for $n = 3$.

Let us now compute the probability of no tree change. Between $[0, t_3]$, there are three lineages, two of which coalesce at t_3 and another that does not. The probability that recombination happens at the third lineage and that it does not change the tree is

$$\begin{aligned} P([0,t_3];3) &= \frac{1}{2t_2+t_3}\int_0^{t_3}\left[\int_{t_r}^{t_3} e^{-3(t_c-t_r)}dt_c + e^{-3(t_3-t_r)}\int_{t_3}^{t_2} e^{-2(t_c-t_3)}dt_c\right]dt_r \\ &= \frac{1}{2t_2+t_3}\left[\frac{t_3}{3}+\frac{1}{18}\left(1-e^{-3t_3}\right)-\frac{1}{6}\left(e^{-2(t_2-t_3)}-e^{-(2t_2+t_3)}\right)\right] \\ &= \frac{1}{3(2t_2+t_3)}\left[t_3+\frac{1}{6}\left(1-e^{-3t_3}\right)-\frac{1}{2}\left(e^{-2(t_2-t_3)}-e^{-(2t_2+t_3)}\right)\right]. \end{aligned} \tag{14}$$

Next, the probability that recombination happens in any of the two lineages that coalesced at t_3 and does not change the tree is

$$\begin{aligned} P([0,t_3];1,2) &= \frac{2}{2t_2+t_3}\int_0^{t_3}\left[\int_{t_r}^{t_3} e^{-3(t_c-t_r)}dt_c\right]dt_r \\ &= \frac{2}{2t_2+t_3}\left[\frac{t_3}{3}-\frac{1}{9}\left(1-e^{-3t_3}\right)\right] \\ &= \frac{2}{3(2t_2+t_3)}\left[t_3-\frac{1}{3}\left(1-e^{-3t_3}\right)\right]. \end{aligned} \tag{15}$$

Finally, recombination can happen in one of the lineages whenever two lineages remain,

$$P([t_3,t_2]) = \frac{2}{2t_2+t_3}\int_{t_3}^{t_2}\left[\int_{t_r}^{t_2} e^{-2(t_c-t_r)}dt_c\right]dt_r$$
$$= \frac{1}{2t_2+t_3}\left[(t_2-t_3)-\frac{1}{2}\left(1-e^{-2(t_2-t_3)}\right)\right]. \tag{16}$$

Adding the three terms,

$$P(\text{tree unchanged}) = P([0,t_3];3)+P([0,t_3];1,2)+P([t_3,t_2])$$
$$= \frac{6t_2-4+e^{-(2t_2+t_3)}+2e^{-2(t_2-t_3)}+e^{-3t_3}}{6(2t_2+t_3)}. \tag{17}$$

For the rate of tree changes, $\lambda(t_2,t_3) = NT\left(1-P(\text{tree unchanged})\right)$, giving

$$\lambda(t_2,t_3) = N\left[\frac{2}{3}-\frac{1}{6}e^{-2t_2-3t_3}\left(e^{2t_2}+e^{2t_3}+2e^{5t_3}\right)+t_2+t_3\right]. \tag{18}$$

The stationary distribution of the coalescence times at tree changes is, using Eq. (11)

$$\pi(t_2,t_3) = \frac{P_c(t_2,t_3)\lambda(t_2,t_3)}{\int_0^\infty\left[\int_{t_3}^\infty P_c(t_2,t_3)\lambda(t_2,t_3)dt_2\right]dt_3}. \tag{19}$$

Conditional on (t_2,t_3), the distribution of the distance to the next tree change (in Morgan) is $\ell\,|\,t_2,t_3 \sim \text{Exp}(\lambda(t_2,t_3))$. The unconditional distribution of the distances between tree changes is thus

$$\psi(\ell) = \int_0^\infty\int_{t_3}^\infty \pi(T)\cdot\lambda(t_2,t_3)e^{-\lambda(t_2,t_3)\ell}dt_2dt_3$$
$$= \frac{\int_0^\infty\int_{t_3}^\infty P_c(t_2,t_3)\lambda^2(t_2,t_3)e^{-\lambda(t_2,t_3)\ell}dt_2dt_3}{\int_0^\infty\left[\int_{t_3}^\infty P_c(t_2,t_3)\lambda(t_2,t_3)dt_2\right]dt_3}. \tag{20}$$

Using a similar technique to the SMC case, the mean distance is

$$\langle\ell\rangle = \left[\int_0^\infty\left[\int_{t_3}^\infty P_c(t_2,t_3)\lambda(t_2,t_3)dt_2\right]dt_3\right]^{-1}, \tag{21}$$

and the mean number of tree changes is

$$\langle n_c\rangle = \frac{L}{\langle\ell\rangle} = L\int_0^\infty\left[\int_{t_3}^\infty P_c(t_2,t_3)\lambda(t_2,t_3)dt_2\right]dt_3. \tag{22}$$

Substituting Eq. (13) for $P_c(t_2,t_3)$ and Eq. (18) for $\lambda(t_2,t_3)$, we obtain

$$\langle n_c\rangle = \frac{19NL}{9}. \tag{23}$$

Recall that the number of tree changes for SMC was $3NL$ (Eq. (10)), which is the same as the number of recombinations. Therefore, we conclude that under SMC', exactly 19/27 of recombination events result in a tree change.

3 Acknowledgments

We thank Asger Hobolth for suggesting the use of the detailed balance equation to find the stationary distribution.

References

[1] G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B*, 360:1387–1393, 2005.

[2] S. Carmi, P. R. Wilton, J. Wakeley, and I. Pe'er. A renewal theory approach to ibd sharing. *Theor. Popul. Biol.*, 97:35, 2014.

[3] P. R. Wilton, S. Carmi, and A. Hobolth. The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200:343, 2015.

[4] J. Wakeley. *Coalescent Theory: An Introduction.* Roberts & Company Publishers, Greenwood Village, Colorado, USA, 2009.

[5] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution (IMA Volumes in Mathematics and its Applications)*, volume 87, pages 257–270. Springer-Verlag, Berlin, 1997.

[6] P. Marjoram and J. D. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7:16, 2006.