PROBABILITY OF TREE CHANGES IN THE ANCESTRAL RECOMBINATION GRAPH

ELENA AXINN

Abstract. We study the Ancestral Recombination Graph (ARG) and the Sequentially Markov Coalescent' (SMC'). We first recall the algorithms of the ARG and SMC', providing concise definitions of their respective methodologies. Building upon this foundation, conducted a literature review to establish a classification system for different recombination events, drawing from multiple scholarly papers. Next, we focus on the SMC' structure with two leaves. We computed the probabilities associated with each classification of recombination events. These formulas represent the core findings of this research endeavor. It is important to note that the fluctuation in tree lengths adheres to a Markov process, which serves as the key mathematical property of interest. This research contributes to the advancement of our understanding of evolutionary processes and their mathematical underpinnings, paving the way for further exploration and applications in related fields.

1. Introduction and Motivation

Mathematical biology largely refers to the area of mathematics that applies its knowledge to a wide array of biological topics. My project would be considered a probabilistic approach to studying population genetics, which is the study of the evolutionary history and relationships among or within groups of organisms.

Mathematicians study many things in population genetics, but each study will consider the genetic relationship among a group of individuals. Typically this is applied to groups of genetic interest, like different strains of viruses or prehistoric fossils. Such studies quantify the genetic "closeness" of two individuals by focusing on their Most Recent Common Ancestor (MRCA). This definition is right in the name – the closest genetic ancestor to both of the individuals. Typically this is a good measure for how related they are. The process of reaching the MRCA is called coalescence [Kin82]. Coalescence is defined in the biological glossary.

There are many methods by which coalescence can be modeled, but the two that are most relevant for my project are the Wright-Fisher model and the Kingman Coalescent, both of which are described below.

The Wright-Fisher model is a very simple and idealized model of genetic reproduction [Fis30]. It can be described with the following steps:

(1) Choose a population size denoted N. This will be kept constant throughout the simulation, so each generation (iteration of the process) will have size N.
(2) Assign allele types to each member of the original generation. Assign reproductive probabilities to each member of the original generation.

(3) Create a new generation of the same size N. Assign each member of the new generation a parent dependent on the probabilities assigned to the previous generation. Individuals will inherit their parents' exact allele type with probability 1.
(4) Repeat until there are as many generations as desired.

Once this process is complete, there will be N individuals, each with an allele type inherited from their parent. To find the MRCA of two individuals, simply trace their lineages backward in time until they become one lineage. This process can be modeled as a tree where two branches become one at this coalescence point.

However, because this is a stochastic (random) process, each iteration of the Wright-Fisher model will produce a different tree even when every initial condition is held constant. For that reason, it is necessary to consider the average of all of these trees: The Kingman Coalescent.

With a full knowledge of the Wright-Fisher model, it is easy to understand the Kingman Coalescent in an intuitive way: it is the average of the trees generated by the Wright-Fisher model for any two individuals in the last generation [Kin82]. This can be thought of as eliminating some of the randomness in the Wright-Fisher model. It provides biologists with more relevant information by reducing the possibility of the tree being a probabilistic outlier.

These models made incredible strides in probabilistic population genetics. However, by the very nature of models, they can be improved. There are several ways that these coalescent models could be improved – like allowing for variation in population size, incorporating spatial tactics, or utilizing mutations. My project focuses on one of many ways to improve the biological accuracy of coalescent models: integrating recombination. Here we focus on understanding how the coalescent tree changes at a recombination event, which can be useful for developing sampling formula [Cra16, WFKS23] in the future.

2. Background on Recombination

Amalgamated from several sources, this is the definition of the Ancestral Recombination Graph (ARG) [YD21] [MC05].

Definition 1. ***ARG*** *A process starting in the present and looking backward in time in which the ancestral lineages relating to the sampled chromosomes are traced until coalescence or recombination. Note that chromosomes are defined in the biological glossary. It is also the mathematical structure which fully describes the joint distribution of coalescent trees along the genome, providing all of the information about the genealogical history of a sample, including the locations of recombination events.*

This can be thought of much like a family tree, but represented as a mathematical graph with nodes representing individuals and edges representing the shared genetic material retained in reproduction. Biologically relevant ARGs can have thousands or millions of individuals.

In the ARG, coalescence is typically done in accordance with the Kingman Coalescent model [Kin82]. The primary difference is that it also incorporates recombination events as defined in the biological glossary.

Recombination is done according to one of many different algorithms that vary greatly in speed and accuracy.

It is worth giving some thought to the biological significance of recombination. This process happens very often in the natural world – an offspring will inherit not one parents' allele type or the other, but a mixture of the two. On a genetic level, this means that the two parent alleles were torn apart and sewn back together in a new order to create the child's allele type. This creates genetic richness in ancestral lineages.

The ARG is an interesting mathematical object in addition to being very biologically relevant. These are some of the interesting mathematical properties of the ARG.

(1) Graph structure: The ARG is a Directed Acyclic Graph, meaning the edges have direction, there are no cycles, and it is composed of nodes and edges
(2) Coalescent property: Any 2 lineages will merge into a common ancestor
(3) Recombination: Shuffling/exchange of genetic material is dictated by a probabilistic model
(4) Data-driven: Becomes complex with the extensive genetic data used in frontier computational applications
(5) Mutations: Can include mutations, which add to complexity but also to biological accuracy.

The Sequentially Markov Coalescent' (SMC') is one algorithm that supplies recombination events to a coalescent tree. This is the most accurate algorithm that approximates the ARG with a reasonable run time. That is to say that while other algorithms may be slightly more accurate, they take much longer to run, resulting in the extreme popularity of the SMC'.

This process looks both left to right across a genome and backwards in time on the coalescent tree. Note that the genetic material is represented by the unit interval, so the beginning of the genetic material is represented by 0 and the end is represented by 1. The SMC' algorithm is as follows:

(1) Set $x = 0$ as a distance left to right on the genetic material and generate a coalescent tree (by the Kingman process) denoted $T(x)$ with numeric length $L(x)$.
(2) Generate $y \sim Exp(\frac{\rho}{2}L(x))$, the left to right distance along the chromosome until the next recombination event.
(3) Choose a point g on $T(x)$ uniformly.
(4) Add a recombination event to the graph at g. This recombination occurs at chromosomal location y. The left emerging branch of the recombination event follows the original path of the lineage, the right emerging branch coalesces with the other lineages with an exponential distribution at a rate of 1 per lineage. The lineage that the right emerging branch coalesces with will determine the type of the recombination, with types classified by section 1.2.
(5) Delete the part of the left emerging branch that lies between the recombination event and the branch's coalescence with another branch to revert the graph back to a tree.
(6) Set $\hat{x} = x + y$, with $T(\hat{x})$ and $L(\hat{x})$ constructed based on this new $\hat{x}$.
(7) If $\hat{x} < 1$, return to step 2, replacing every x with $\hat{x}$. If $\hat{x} = 1$, stop.

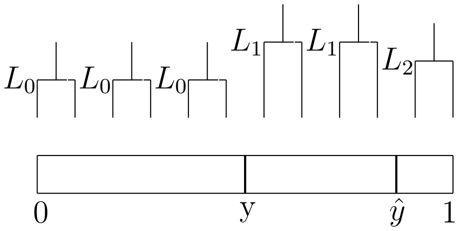


FIGURE 1. Result of the SMC' Algorithm

The SMC' is closely related to the SMC, but with a slight variation regarding the order of steps 4 and 5. The SMC' algorithm is slightly more accurate, so it is favored over the SMC [MW06].

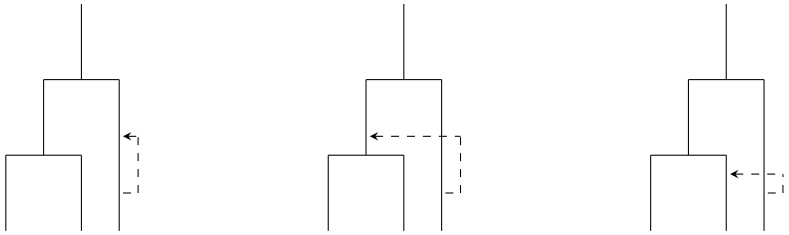
It is clear that the SMC' algorithm is an extremely accurate approximation of the ARG. One must wonder, though, why computational biologists choose not to simulate the entire ARG. The original object has two main limitations: The state-space is huge and the data can be uninformative. The ARG itself does not bound recombination events in any way, resulting in a huge number of recombinations in any given simulations. These simulations can take days to run successfully. Despite this computational inconvenience, many computational biologists would be forced to continue to use this method if it guaranteed the best data. However, much of the data that the ARG produces is irrelevant to the original sample of individuals because it can quickly become so genetically distant. Part of the beauty of the SMC' algorithm is that it maintains a focus on the most genetically relevant material by deleting the leftmost branch. This helps the recombination to happen only with the most genetically relevant material, producing more genetically relevant material rather than irrelevant offshoots. These are the problems that originally motivated the creation of the SMC algorithm.

3. Recombination Types

There are many approaches to defining recombination events, with different classifications depending on the level of specificity required for the problem. In our work, we use the following classification developed by Hein et al [JH04].

(1) These two recombining sequences coalesce with each other before coalescing with any other genetic material. Consequently, they have identical genealogies. That is why this is known as an invisible recombination – it does not affect the genetic data in any way and cannot be meaningfully represented by any known algorithm.
(2) These two recombining sequences coalesce with only one other sequence before coalescing with each other. This elongates the branches of the tree, but does not change the topology.
(3) These two recombining sequences coalesce with two or more sequences before coalescing with each other. This does change the topology of the tree. These recombination events are considered the most genetically relevant.

As mentioned, ome biologists take different approaches to classifying recombination events, both more broadly and more specifically. Some results focus more generally on recombination



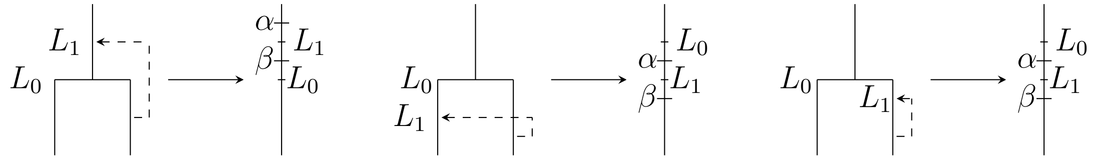
(A) Type 1: No Change

(B) Type 2: Length Change

(C) Type 3: Topology Change

events that result in no change to the tree in comparison to those that do. In the case of the waiting distance paper, they outline four types of recombination events, specifying whether a length-changing event shortens or lengthens the tree [YD21]. The approach to classifying events as outlined in this paper is certainly not more genetically accurate, but simply highlights the most relevant differences for our study.

With that established, we begin with the most simple example of the recombination graph possible: The n=2 case. The point at the "beginning" of the recombination event (denoted g by the original algorithm) is placed uniformly on the tree. The point at the "end" of the recombination event can then be on any of the 3 branches of the graph. The branch that it combines with will dictate whether the resulting tree is lengthened, shortened, or has no change. This distinction is as follows:



(A) Type 2: Lengthening

(B) Type 2: Shortening

(C) Type 1: No Change

Note that these diagrams depict the mapping of the 2 dimensional tree graph to the 1 dimensional vertical interval. This is intended to highlight that the probability of these changes is dictated by the vertical position of L_1 with respect to L_0. In the case of the shortening and no change events, they take place in the same vertical area of the graph, with the only difference being their horizontal location. The recombination event happens on each of these lower branches with uniform probability $\frac{1}{2}$, so these probabilities can be considered the same with that factor included.

There are two more variables that are not defined by these figures, but are important for the following theorems. The first is $\lambda(r)$, which represents the rate of recombination. This happens with every branch at a rate of 1, so in the vertical area with two branches (before L_0) $\lambda(r) = 2$ and in the vertical area with one branch (after L_0) $\lambda(r) = 1$. The second is A. A is a random variable representing a distance backward in time on the figure at which the recombination event begins. On these figures, it would be placed where the arrow begins, whereas L_1 is placed where the arrow ends.

4. The SMC' n=2 case

4.1. Probability of the next tree length.

Theorem 1. *The conditional probability that L_1 falls within the range (α, β), given A and L_0, is*

$$\mathbb{P}(L_1 \in (\alpha, \beta)|A, L_0) = e^{-\int_A^\alpha \lambda(r)dr}(1 - e^{-\int_\alpha^\beta \lambda(r)dr}).$$

Proof.

$$\mathbb{P}(B \in (\alpha, \beta)|A) = \mathbb{P}(N[A, \alpha] = 0 \cap N[\alpha, \beta] = 1)$$

By definition of recombination, this simply states that the recombination hit occurs 0 times from A to α, and then once from α to β. It is not necessary to include the probability that there are no recombination hits after β, as this is not relevant to the calculation of the next tree height: we assume that there is only one recombination hit total.

$\mathbb{P}(N[A, \alpha] = 0 \cap N[\alpha, \beta]) = \mathbb{P}(N[A, \alpha] = 0)\mathbb{P}(N[\alpha, \beta] = 1|N[A, \alpha] = 0)$

By the definition of intersection, and its relationship to conditional probabilities.

$\mathbb{P}(N[A, \alpha] = 0)\mathbb{P}(N[\alpha, \beta] = 1|N[A, \alpha] = 0) = \mathbb{P}(N[A, a] = 0)\mathbb{P}(N[\alpha, \beta] = 1)$

By the conditional independence of a Poisson process.

$\mathbb{P}(N[A, \alpha] = 0)\mathbb{P}(N[\alpha, \beta] = 1) = \mathbb{P}(N[A, \alpha] = 0)(1 - \mathbb{P}(N[\alpha, \beta] = 0)$

By the assumption that there is only one recombination hit total, similar to the reasoning for not including the probability of having no recombination hits after b.

$\mathbb{P}(N[A, \alpha] = 0)(1 - \mathbb{P}(N[\alpha, \beta] = 0) = e^{-\int_\alpha^A \lambda(r)dr}(1 - e^{-\int_\alpha^\beta \lambda(r)dr})$

By the definition of the probability of no hits in a Poisson process. □

Most of these variables are defined in the diagrams above, but two are not: A and $\lambda(r)$. A refers to the "starting" point of the recombination event, or where the arrow begins its formation in the diagrams. $\lambda(r)$ refers to the recombination rate, which is 2 in the vertical location before L_0 (1 on each branch, then combined) and 1 in the vertical location after L_0.

In order to make this formula more specific, particularly by replacing the $\lambda(r)$ with a numerical value, it is necessary to break the formula into different cases: lengthening, shortening, and no change. These numbers are the result of the piecewise recombination parameter, which is 2 while there are 2 branches (before L_0) and 1 while there is 1 branch (after L_0).

Corollary 1. *The probability of lengthening is*

$$\mathbb{P}(L_1 \in (\alpha, \beta)|A, L_0) = e^{-\int_A^{L_0} 2dr + \int_{L_0}^\alpha 1dr}(1 - e^{-\int_\alpha^\beta 1dr})$$

For the following 2 theorems, keep in mind that the only difference between the shortening case and the no change case is their horizontal position on the tree – there is no difference vertically. Thus, when computing the probability of each event within a certain range, it is necessary to multiply by $\frac{1}{2}$ to represent each horizontal option on the tree. This assumes that the distribution between each horizontal position on the tree is uniform.

Corollary 2. *The probability of shortening is* $\frac{1}{2}\mathbb{P}(L_1 \in (\alpha, \beta)|A, L_0) = \frac{1}{2}e^{-\int_A^\alpha 1dr}(1 - e^{-\int_\alpha^\beta 1dr})$

Corollary 3. *The probability of no change is* $\frac{1}{2}\mathbb{P}(L_1 \in (\alpha, \beta)|A, L_0) = \frac{1}{2}e^{-\int_A^\alpha 1dr}(1 - e^{-\int_\alpha^\beta 1dr})$

This means that this same probability can be expressed piecewise and more generally as follows.

Theorem 2. $\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \begin{cases} \frac{1}{2}(1 - e^{-2l_1+2a}) & \text{if } l_1 < l_0 \\ 1 - e^{2a-l_0-l_1} & \text{if } l_1 > l_0 \end{cases}$

Proof. Theorem 1 is written with $\lambda(r)$ included because $\lambda(r)$ changes at l_0. To include specific numbers for $\lambda(r)$, we must specify whether this recombination event is happening before or after l_0. This necessitates a piecewise formula, though this proof utilizes many of the same concepts as the proof for Theorem 1.

First consider when $l_1 < l_0$. This is conceptualizes as the Poisson process having 1 hit in the area from a to l_1. Because there is only 1 hitting event total in this biological example, the probability of 1 event in that area can be written as 1 minus the probability of no hits in that area. Note that this vertical range has two horizontal areas, each with uniform hitting probability of $\frac{1}{2}$. Thus $\frac{1}{2}$ must be multiplied in front of the formula. Combining these facts produces the following initial formula.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \frac{1}{2}\mathbb{P}(1 - N[a, l_1] = 0)$ if $l_1 < l_0$

By the definition of the probability of no hits in a Poisson process, we get the following more specific formula.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \frac{1}{2}(1 - e^{-\int_a^{l_1} \lambda(r)dr})$ if $l_1 < l_0$

In this area, $\lambda(r) = 2$, so the formula can be rewritten as follows.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \frac{1}{2}(1 - e^{-\int_a^{l_1} 2dr})$ if $l_1 < l_0$

Use the constant rule of integration to simplify as follows.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \frac{1}{2}(1 - e^{-2l_1 - a})$ if $l_1 < l_0$

This concludes the proof of the equation given for the first interval. Now, approach the second.

In the next interval, $l_1 > l_0$. This means there is one hit either from a to l_0 or from l_0 to l_1. Again because there is only one total hitting event in this recombination process, this can be expressed as 1 minus the probability that there are no hits both from a to l_0 and from l_0 to l_1. It is tempting to multiply this expression by the probability of having no hits from a to l_0, but this is not necessary as we are considering $L_1 \leq l_1$, not $L_1 = l_1$. That is to say, it is fine if the hitting event happens before l_0, so long as $l_1 > l_0$ in this interval. Combining these facts produces the following formula.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = 1 - \mathbb{P}(N[a, l_0] = 0)\mathbb{P}(N[l_0, l_1])$ if $l_0 < l_1$

By the definition of the probability of no hits in a Poisson process, we get the following more specific formula.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = 1 - e^{-\int_a^{l_0} \lambda(r)dr} e^{-\int_{l_0}^{l_1} \lambda(r)dr}$ if $l_0 < l_1$

$\lambda(r) = 2$ from a to l_0 and $\lambda(r) = 1$ from l_0 to l_1, so insert these values appropriately.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = 1 - e^{-\int_a^{l_0} 2dr} e^{-\int_{l_0}^{l_1} 1dr}$ if $l_0 < l_1$

Use the constant rule of integration to simplify.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = 1 - e^{-2(l_0-a)}e^{-(l_1-l_0)}$ if $l_0 < l_1$

Use the multiplication rule of exponents to add these exponents as follows.

$1 - e^{2a-l_0-l_1}$ if $l_0 < l_1$

This concludes the proof of the equation given for the second interval. Combine these intervals into one piecewise equation.

$$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0, A = a) = \begin{cases} \frac{1}{2}(1 - e^{-2l_1+2a}) & \text{if } l_1 < l_0 \\ 1 - e^{2a-l_0-l_1} & \text{if } l_1 > l_0 \end{cases} \qquad \square$$

In order to generalize this probability, it is necessary to remove the conditioning on A, the original breakpoint of the recombination event. In accordance with conditional probability rules, it is necessary to multiply the CDF (Theorem 2) by $\frac{1}{l_0}$ (the range that A can be found in) as well as integrate from 0 to l_0 with respect to A.

Theorem 3. $\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \begin{cases} \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1) & \textit{if } l_1 < l_0 \\ -\frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1) & \textit{if } l_1 > l_0 \end{cases}$

Proof. In order to generalize over all $A = a$ values, it is necessary to integrate with respect to a over the piecewise range of A (which changes based on interval) and then to divide that expression by the total distance of that interval (which is $l_0 - 0 = l_0$). Because the original equation (Theorem 2) is expressed piecewise, this process will be done in two pieces according to the different intervals. Begin with $l_1 < l_0$.

Start by dividing by l_0 and integrating with respect to a from 0 to l_1, the range of A for this interval. This produces the following.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \frac{1}{2l_0}\int_0^{l_1}(1 - e^{-2l_1-a})$ if $l_1 < l_0$

Now compute this integration. First, apply linearity to break up the integral.

$\int_0^{l_1}(1 - e^{-2l_1-a}) = \int_0^{l_1} 1da - e^{-2l_1}\int_0^{l_1} e^{2a}da$ if $l_1 < l_0$

Now use the constant rule to solve this first integral.

$\int_0^{l_1} 1da = a|_0^{l_1}$ if $l_1 < l_0$

Next, use u-substitution with $u = 2a$ and $du = 2da$ to solve the second integral.

$\int_0^{l_1} e^{2a}da = \frac{e^{2a}}{2}|_0^{l_1}$ if $l_1 < l_0$

Apply these evaluated integrals to the original broken up integral.

$\int_0^{l_1} 1da - e^{-2l_1}\int_0^{l_1} e^{2a}da = a|_0^{l_1} - e^{-2l_1}\frac{e^{2a}}{2}|_0^{l_1}$ if $l_1 < l_0$

Using multiplication and addition to simplify that result produces the following equation.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \frac{1}{2l_0}\frac{2a-e^{2a-2l_1}}{2}|_0^{l_1}$ if $l_1 < l_0$ if $l_1 < l_0$

Evaluating from 0 to l_1 along with some simple multiplication produces the final equation.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1)$ if $l_1 < l_0$

This completes the proof for the first interval. Now begin the second interval.

In this case, the possible range for A is from 0 all the way up to l_0. This changes the bound of integration. However, the equation is still divided by l_0 as this is still the total range of A. Those facts produce the following equation.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \frac{1}{l_0}\int_0^{l_0} 1 - e^{2a-l_0-l_1} dl_1$ if $l_1 > l_0$

Now compute this integration. First, apply linearity to break up the integral.

$\int_0^{l_0} 1 - e^{2a-l_0-l_1} dl_1 = \int_0^{l_0} 1da - e^{-l_0-l_1}\int_0^{l_0} e^{2a}da$ if $l_1 > l_0$

Use the constant rule to solve the first integral.

$\int_0^{l_0} 1da = a|_0^{l_0}$ if $l_1 > l_0$

Use u-substitution with $u = 2a$ and $du = 2da$ to solve the second integral.

$\int_0^{l_0} e^{2a}da = \frac{e^{2a}}{2}|_0^{l_0}$ if $l_1 > l_0$

Apply these evaluated integrals to the original broken up integral.

$\int_0^{l_0} 1da - e^{-l_0-l_1}\int_0^{l_0} e^{2a}da = a|_0^{l_0} e^{-l_0-l_1}\frac{e^{2a}}{2}|_0^{l_0}$ if $l_1 > l_0$

Using multiplication and addition to simplify that result produces the following equation.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \frac{1}{l_0}\frac{2a - e^{2a-l_0-l_1}}{2}|_0^{l_0}$ if $l_1 > l_0$

Evaluating from 0 to l_1 along with some simple multiplication produces the final equation.

$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = -\frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1)$ if $l_1 > l_0$

This concludes the proof of the equation given for the second interval. Combine these intervals into one piecewise equation.

$$\mathbb{P}(L_1 \leq l_1 | L_0 = l_0) = \begin{cases} \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1) & \text{if } l_1 < l_0 \\ -\frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1) & \text{if } l_1 > l_0 \end{cases} \qquad \square$$

Note that the Probability Density Function (PDF) is not necessary to calculate here because it does not contribute to the expectation of the eventual tree length. In addition, at the point l_0, the Cumulative Distribution Function (CDF) experiences a jump that would need to be modeled with the Dirac delta function. Both of those reasons contribute to the fact that it is not outlined in this paper.

4.2. **Expectation of eventual length.** The expectation of eventual tree length is the mathematical object that our project sought – this perfectly describes the long-time tree behavior of the n=2 case, since the tree can only experience length changes (as opposed to topology changes). This section describes the expectation of final tree length.

Theorem 4. $\mathbb{E}[L_1 | L_0 = l_0] = \frac{e^{-2l_0}((6l_0^2+2l_0-1)e^{2l_0}+1)+4-4e^{-2l_0}}{8l_0}$

Proof. To calculate expectation, use this well-known fact from probability in general.

$\mathbb{E}[X] = \int_{-\infty}^{\infty} 1 - \mathbb{P}(X \leq x)dx$

Which, in our case, becomes the following theorem.

$\mathbb{E}[L_1 | L_0 = l_0] = \int_0^{\infty} 1 - \mathbb{P}(L_1 \leq l_1 | L_0 = l_0)dl_1$

Then, we can take this theorem a step further by integrating our actual cumulative distribution function. Keep in mind that this is piecewise, so it must be integrated separately and then added together to produce the integral on the full range of the function. Use Theorem 3, which has removed the conditioning on A.

$\mathbb{E}[L_1|L_0 = l_0] = \int_0^{l_0} 1 - \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1)dl_1 + \int_{l_0}^{\infty} 1 + \frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1)dl_1$

Evaluate these integrals separately and then add the results. Start with the first integral.

$\int_0^{l_0} 1 - \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1)dl_1 = -\frac{1}{4l_0}\int_0^{l_0} e^{-2l_1}dl_1 - \frac{1}{2l_0}\int_0^{l_0} l_1dl_1 + (\frac{1}{4l_0} + 1)\int_0^{l_0} 1dl_1$

By applying linearity. Now evaluate each integral separately, applying the answers to the equation above. Begin with the first integral, using the method of u-substitution with $u = -2l_0 \to du = -2dn$.

$\int_0^{l_0} e^{-2l_1}dl_1 = -\frac{e^{-2l_1}}{2}|_0^{l_0}$

Next, use the power rule to evaluate the second integral.

$\int_0^{l_0} l_1dl_1 = \frac{l_1^2}{2}|_0^{l_0}$

Finally, use the constant rule on the third integral.

$\int_0^{l_0} 1dl_1 = l_1|_0^{l_0}$

Plug these values into the previous equation for the following equation.

$-\frac{1}{4l_0}\int_0^{l_0} e^{-2l_1}dl_1 - \frac{1}{2l_0}\int_0^{l_0} l_1dl_1 + (\frac{1}{4l_0} + 1)\int_0^{l_0} 1dl_1 = (\frac{1}{4l_0}\frac{e^{-2l_1}}{2} - \frac{1}{2l_0}\frac{l_1^2}{2} + (\frac{1}{4l_0} + 1)l_1)|_0^{l_0}$

Simplify this formula through multiplication and addition for the following.

$(\frac{1}{4l_0}\frac{e^{-2l_1}}{2} - \frac{1}{2l_0}\frac{l_1^2}{2} + (\frac{1}{4l_0} + 1)l_1)|_0^{l_0}=(\frac{e^{-2l_1}-2nl_1^2+(8l_0+2)l_1}{8l_0})|_0^{l_0}$

Evaluate this expression from 0 to l_0. Then the final evaluation of this first integral is as follows.

$\int_0^{l_0} 1 - \frac{1}{4l_0}(e^{-2l_1} + 2l_1 - 1)dl_1 = \frac{e^{-2l_0}+6l_0^2+2l_0}{8l_0}$

Now approach the evaluation of the second integral. First distribute the first two terms into the third term to reduce the complexity of the equation. Elementary multiplication and addition produces the following result.

$\int_{l_0}^{\infty} \frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1)dl_1 = \int_{l_0}^{\infty} \frac{1}{2l_0}(e^{l_0-l_1} - e^{-l_0-l_1})dl_1$

Then apply linearity to simplify.

$\int_{l_0}^{\infty} \frac{1}{2l_0}(e^{l_0-l_1} - e^{-l_0-l_1}) = \frac{1}{2l_0}(e^{l_0} - e^{-l_0})\int_{l_0}^{\infty} e^{-l_1}dl_1$

Apply u-substitution with u=-n and du=-dn to the final term in the integral.

$\int_{l_0}^{\infty} e^{-l_1} = -e^{-l_1}|_{l_0}^{\infty}$

$\frac{1}{2l_0}(e^{l_0} - e^{-l_0}\int_{l_0}^{\infty} e^{-l_1}dl_1 = \frac{1}{2l_0}(e^{l_0} - e^{-l_0}) - e^{-l_1}|_{l_0}^{\infty}$

Simplifying using multiplication and addition produces the following result.

$\frac{e^{-l_0-l_1}-e^{l_0-l_1}}{2l_0}|_{l_0}^{\infty}$

Evaluating from 0 to ∞ produces the following final result for the second integral.

$\int_{l_0}^{\infty} \frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1} + e^{2l_0} - 1)dl_1 = \frac{1-e^{2l_0}}{2l_0}$

These pieces are added together for the following expression of the expectation.

$\int_0^{l_0} 1-\frac{1}{4l_0}(e^{-2l_1}+2l_1-1)dl_1+\int_{l_0}^{\infty} 1+\frac{1}{2l_0}(e^{-l_0-l_1})(-2l_0e^{l_0+l_1}+e^{2l_0}-1)dl_1 = \frac{e^{-2l_0}+6l_0^2+2l_0}{8l_0}+\frac{1-e^{2l_0}}{2l_0}$

This expression need only be simplified through multiplication and addition for the final expectation result.

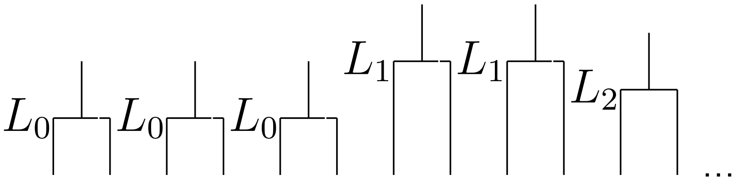
$\mathbb{E}[L_1|L_0 = l_0] = \frac{e^{-2l_0}((6l_0^2+2l_0-1)e^{2l_0}+1)+4-4e^{-2l_0}}{8l_0}$ □

The following table is provided to build intuition on the expectation of L_1 for different values of l_0.

l_0	$\mathbb{E}[L_1\vert L_0 = l_0]$
1	10.594
2	30.945
3	62.993

5. Discussion on Markov Chains

Though we have primarily focused on the changes themselves, the most mathematically relevant properties lie in the relationships between different changes. The changing of a tree over time is a Markov Process: only the current structure of a tree has any affect on the probabilities of how the tree may change. This is clear from the variables involved in the probability calculations – they only reflect the current height of the tree, current place on the genome, and so forth. This is what makes the coalescent with recombination so mathematically interesting – the recombination events maintain the Markov Property.



Both mathematicians and biologists are interested in the long-term behavior of this tree length. This is why computing the probabilities of the respective tree changes is so relevant – knowing how the tree will change is the same thing as knowing what its eventual length will be.

6. Connections to The Distribution of Waiting Distances in ARGs

The paper [YD21] was the inspiration for my work on the n=2 case. It produces several very relevant results. It defines waiting distance as the following.

Definition 2. ***Waiting distance*** *The distance until the next recombination event along the chromosome, exponentially distributed.*

Though the paper is centered around waiting distances, in seeking these theorems, the authors ended up finding several theorems regarding the probability of different tree changes that closely mirror the calculations I did for the n=2 case. In particular, they found the probability for general n that a recombination event does not produce a topology change. As per the recombination classifications in the n=2 case, this means the recombination event produces no change, a shortening change, or a lengthening change. The following equation can be drawn as an initial step to comparing that paper's results to the results in this paper.

$\mathbb{P}$(Topology does not change)= $\mathbb{P}$(Lengthening + Shortening + No Change)

First, it is important to modify the theorems in this paper to be easily compared to those in the other paper. Observe corollaries 1.1, 1.2, and 1.3, but choose the α and β values to be their entire respective ranges. The new corollaries are as follows.

Lengthening: $\mathbb{P}(l_1 \in (l_0, \infty)|A = a, L_0 = l_0) = e^{-\int_A^{l_0} 2dr - \int_{l_0}^{l_0} 1dr}(1 - e^{-\int_{l_0}^{\infty} 1dr} = e^{2A-2l_0}$
Shortening: $\mathbb{P}(l_1 \in (A, l_0)|A = a, L_0 = l_0) = \frac{1}{2}e^{-\int_A^A 1dr}(1 - e^{-\int_A^{l_0} 1dr}) = \frac{1}{2}(1 - e^{A-l_0})$
No change: $\mathbb{P}(l_1 \in (A, l_0)|A = a, L_0 = l_0) = \frac{1}{2}e^{-\int_A^A 1dr}(1 - e^{-\int_A^{l_0} 1dr}) = \frac{1}{2}(1 - e^{A-l_0})$

Now it is necessary to remove the conditioning on A.

Lengthening: $\mathbb{P}(l_1 \in (l_0, \infty)|L_0 = l_0) = \frac{1}{l_0}\int_0^{l_0} e^{2A-2l_0}dA = \frac{1}{2l_0}(1 - e^{-2l_0})$
Shortening: $\mathbb{P}(l_1 \in (A, l_0)|L_0 = l_0) = \frac{1}{2l_0}\int_0^{l_0}(1 - e^{A-l_0})dA = \frac{1}{2l_0}(e^{-l_0} + l_0 - 1)$
No change: $\mathbb{P}(l_1 \in (A, l_0)|L_0 = l_0) = \frac{1}{2l_0}\int_0^{l_0}(1 - e^{A-l_0})dA = \frac{1}{2l_0}(e^{-l_0} + l_0 - 1)$

Now add these results together for the final result to compare to the waiting distance paper.

$\mathbb{P}$(Lengthening + Shortening + No Change)= $\frac{1}{2l_0}(1 - e^{-2l_0}) + \frac{1}{2l_0}(e^{-l_0} + l_0 - 1) + \frac{1}{2l_0}(e^{-l_0} + l_0 - 1) = \frac{2e^{-l_0} - e^{-2l_0} + 2l_0 - 1}{2l_0}$

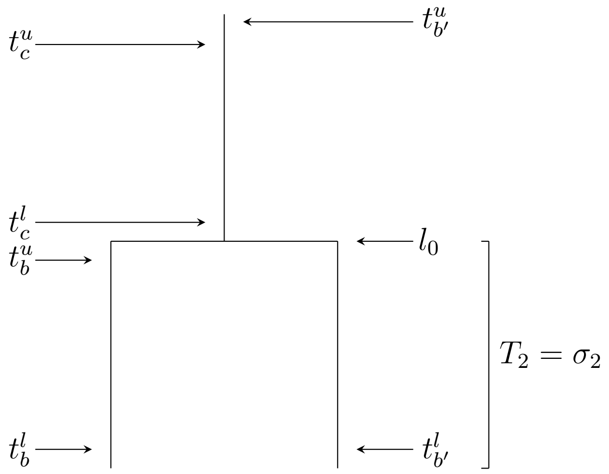
Now, evaluate the key theorem from the waiting time paper for the n=2 case. Begin with the statement of the theorem itself.

$\mathbb{P}$(Topology does not change)= $\frac{1}{L(T)}[\sum_{i=A_n(t_{b'}^u)+1}^{A_n(t_b^l)} \frac{1}{i}[T_i + (e^{i\sigma_i} - e^{i\sigma_{i-1}})(\sum_{j=A_n(t_{b'}^u)}^{A_n(t_{b'}^l)} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}] + \sum_{j=A_n(t_c^u+1)}^{i} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{jT_j}]] + \sum_{i=A_n(t_{b'}^u)+1}^{A_n(t_{b'}^l)} \frac{1}{i}[2T_i + (e^{i\sigma_i} - e^{i}\sigma_{i+1})(2\sum_{j=A_n(t_c^u)}^{i} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}] - \sum_{j=A_n(t_c^u+1)}^{A_n(t_c^l} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}]] - \frac{1}{2}e^{-2\sigma_2} - e^{-\sigma_1}]$

The next step to evaluating this theorem when n=2 is to take note of the variables and symbols found in this equation. Also take note of the meaning in the context of the paper and the evaluation for n=2.

Symbol	Meaning	Evaluation for n=2
b	branch at the start of recombination	no value
b'	branch at the end of recombination	no value
c	the parental branch of b and b'	no value
i	summing variable such that $i \in (2, .., n)$	2
j	summing variable	no value
k	summing variable	no value
$L(T)$	total length of the tree	$2l_0$
n	sample size	2
T_i	length of epoch with i lineages	l_0
σ_i	sum of T_i from i to n	l_0
$A_n(t_b^u)$	number of lineages at t_b^u	2
$A_n(t_b^l)$	number of lineages at t_b^l	2
$A_n(t_{b'}^u)$	number of lineages at $t_{b'}^u$	1
$A_n(t_{b'}^l)$	number of lineages at $t_{b'}^l$	2
$A_n(t_c^u)$	number of lineages at t_c^u	1
$A_n(t_c^l)$	number of lineages at t_c^l	1

To explain the evaluations for n=2, the following diagram is provided with labelings of the stated variables.



The next step toward solving this equation is to plug in the summation bounds. This is done based on the symbol and meaning chart above. Specifically, plug in the values denoting the number of lineages in each area.

$\mathbb{P}$(Topology does not change)$= \frac{1}{L(T)}[\sum_{i=2}^{2} \frac{1}{i}[T_i + (e^{i\sigma_i} - e^{i\sigma_{i-1}})(\sum_{j=3}^{2} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}] + \sum_{j=2}^{i} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{jT_j}]] + \sum_{i=2}^{2} \frac{1}{i}[2T_i + (e^{i\sigma_i} - e^{i}\sigma_{i+1})(2\sum_{j=1}^{i} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}] - \sum_{j=2}^{1} exp(-i\sigma_i - \sum_{k=j+1}^{i-1} KT_k)\frac{1}{j}[1 - e^{-jT_j}]] - \frac{1}{2}e^{-2\sigma_2} - e^{-\sigma_1}]$

Now, because i=2, remove both summations with respect to i and replace every i with 2.

$\mathbb{P}$(Topology does not change)= $\frac{1}{L(T)}[\frac{1}{2}[T_2+(e^{2\sigma_2}-e^{2\sigma_1})(\sum_{j=3}^{2} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}]+\sum_{j=2}^{2} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{jT_j}]]+\frac{1}{2}[2T_2+(e^{2\sigma_2}-e^{2}\sigma_3)(2\sum_{j=1}^{2} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}]-\sum_{j=2}^{1} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}]]-\frac{1}{2}e^{-2\sigma_2}-e^{-\sigma_1}]$

An empty sum is every sum $\sum_p^q$ such that $p > q$. In this case, $\sum_p^q = 0$. Replace every empty sum in this formula with 0.

$\mathbb{P}$(Topology does not change)= $\frac{1}{L(T)}[\frac{1}{2}[T_2+(e^{2\sigma_2}-e^{2\sigma_1})(\sum_{j=2}^{2} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{jT_j}])]+\frac{1}{2}[2T_2+(e^{2\sigma_2}-e^{2}\sigma_3)(2\sum_{j=1}^{2} exp(-2\sigma_2-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}])]-\frac{1}{2}e^{-2\sigma_2}-e^{-\sigma_1}]$

Now replace $T_2 = \sigma_2 = l_0$ and $\sigma_1 = \sigma_3 = 0$ and $L(T) = 2l_0$ as denoted by the chart and diagram.

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(\sum_{j=2}^{2} exp(-2l_0-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{jT_j}])]+\frac{1}{2}[2l_0+(e^{2l_0}-1)(2\sum_{j=1}^{2} exp(-2l_0-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}])]-\frac{1}{2}e^{-2l_0}-1]$

Now sum over j.

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(exp(-2l_0-\sum_{k=3}^{1} KT_k)\frac{1}{2}[1-e^{2l_0}])]+\frac{1}{2}[2l_0+(e^{2l_0}-1)(2\sum_{j=1}^{2} exp(-2l_0-\sum_{k=j+1}^{1} KT_k)\frac{1}{j}[1-e^{-jT_j}])]-\frac{1}{2}e^{-2l_0}-1]$

Now sum over k, taking special note that the sum $\sum_{k=j+1}^{1}$ becomes an empty sum in the case of j=1 or j=2.

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(exp(-2l_0)\frac{1}{2}[1-e^{2l_0}])]+\frac{1}{2}[2l_0+(e^{2l_0}-1)(2\sum_{j=1}^{2} exp(-2l_0)\frac{1}{j}[1-e^{-jT_j}])]-\frac{1}{2}e^{-2l_0}-1]$

Repeat the process of summing over j for the final sum.

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(exp(-2l_0)\frac{1}{2}[1-e^{2l_0}])]+\frac{1}{2}[exp(-2l_0)\frac{1}{2}(1-e^{-2l_0})]-\frac{1}{2}e^{-2l_0}-1]$

Now replace $exp(x) = e^x$

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(e^{-2l_0}\frac{1}{2}[1-e^{2l_0}])]+\frac{1}{2}[e^{-2l_0}\frac{1}{2}(1-e^{-2l_0})]-\frac{1}{2}e^{-2l_0}-1]$

Simplify by multiplying and adding.

$\mathbb{P}$(Topology does not change)= $\frac{1}{2l_0}[\frac{1}{2}[l_0+(e^{2l_0}-1)(\frac{1}{2}e^{-2l_0}-\frac{1}{2}e^{-}4l_0)]+\frac{1}{2}[2l_0+(e^{2l_0}-1)(e^{-2l_0}-e^{-4l_0})]-\frac{1}{2}e^{-2l_0}-1]$

Simplify further by multiplying and adding further.

$\mathbb{P}$(Topology does not change)= $\frac{1}{6l_0}[2l_0+1-2e^{-2l_0}+e^{-4l_0}+4l_0+2-4e^{-2l_0}+2e^{e^{-}4l_0}-2e^{-2l_0}-4]$

Combining like terms produces the final equation as follows.

$\mathbb{P}$(Topology does not change)= $\frac{3e^{-4l_0}-8e^{-2l_0}+6l_0-1}{6l_0}$

The two final results from each paper are below.

$\mathbb{P}$(Topology does not change)= $\frac{3e^{-4l_0}-8e^{-2l_0}+6l_0-1}{6l_0}$

$\mathbb{P}$(Lengthening + Shortening + No Change)= $\frac{-e^{-2l_0}+2e^{-l_0}+2l_0-1}{2l_0}$

Though these are obviously two slightly different fractions, they contain all of the same exponents and follow the same behavior after $l_0 = 2$. They both rapidly approach the value 1.

7. Biological Glossary

- Coalescence: “To coalesce means to grow together, to join, or to fuse. When two copies of a gene are descended from a common ancestor which gave rise to them in some past generation, looking back we say that they coalesce in that specific generation. Seen forward in time, coalescent events are simply DNA replication events, and are only of special interest due to their place in the history of a particular sample. Kingman showed that the joining up of lineages into common ancestors is described by a particular mathematical process" [Wak09].
- Recombination: “In general recombination, genetic exchange takes place between a pair of homologous DNA sequences. These are usually located on two copies of the same chromosome. The details of the intimate interplay between replication and recombination are still incompletely understood" [BA02].
- Chromosome: Any of the rod-shaped or threadlike DNA-containing structures of cellular organisms that contain all or most of the genes of the organism (Merriam-Webster dictionary).
- MRCA: Most Recent Common Ancestor of two individuals. This is the common ancestor at which two copies of a gene coalesce as described in the coalescence definition [Wak09].
- Generation: An iteration of reproduction. Describes a group of offspring that are at the same stage of descent from a common ancestor (Merriam-Webster dictionary).
- Allele: "Alternative forms of the same gene" [BA02].

8. Variable Glossary

- x: A left to right distance on the genetic material.
- $T(x)$: The coalescent tree for x. A function of x.
- $L(x)$: The length of $T(x)$. A function of x.
- y: A left to right distance on the genetic material.
- ρ: The recombination rate for the sample.
- g: A point on $T(x)$.
- $\hat{x}$: A left to right distance on the genetic material.
- A: The random variable representing the distance backward in time on the tree until the recombination event begins.
- a: The realization of one A distance.
- $\lambda(r)$: The recombination rate for a recombination event.
- L_0: The random variable representing the length of the first iteration of a tree.
- l_0: The realization of one L_0 length.
- L_1: The random variable representing the length of the next iteration of a tree.
- l_1: The realization of one L_1 length.
- α: The upper bound on the range of possible L_1 lengths.
- β: The lower bound on the range of possible L_1 lengths.
- $N[n_1, n_2]$: The number of recombination events within the interval $[n_1, n_2]$

9. Acknowledgements

Huge thanks to Dr. Wai-Tong (Louis) Fan for the incredible mentoring. His direction and attention to detail were critical in the completion of this project. Also thanks to Daniel Rickert and Justin Liu for being dedicated collaborators. Their guidance has been invaluable to me. Thanks to Dr. Matthew Hahn, Sarthak Mishra, and Dimitrios Diamantidis for helpful discussions at the beginning of the project. Thanks also to Indiana University for organizational support, especially Dr. Dylan Thurston. Final thanks to NSF grant DMS-2051032 for financial support.

References

[BA02] Julian Lewis Bruce Alberts, Alexander Johnson, *Molecular biology of the cell*, 4 ed., Garland Science, 2002.

[Cra16] Harry Crane, *The ubiquitous ewens sampling formula*, Statistical Science **31** (2016), no. 1, 1–19.

[Fis30] Ronald A. Fisher, *The genetical theory of natural selection*, Transactions of the Royal Society of Edinburgh (1930).

[JH04] Carsten Wiuf Jotun Hein, Mikkel H Schierup, *Gene genealogies, variation and evolution: A primer in coalescent theory*, 148–153.

[Kin82] JFC Kingman, *Origins of the coalescent*, Genetics (1982).

[MC05] Gilean A. T. McVean and Niall J. Cardin, *Approximating the coalescent with recombination*, Philosophical Transactions of the Royal Society (2005).

[MW06] Paul Majoram and Jeff D. Wall, *Fast "coalescent" simulation*, BMC Genetics (2006).

[Wak09] John Wakeley, *Coalescent theory: An introduction*, Roberts and Company Publishers, 2009.

[WFKS23] John Wakeley, Wai-Tong Fan, Evan Koch, and Shamil Sunyaev, *Recurrent mutation in the ancestry of a rare variant*, Genetics **224** (2023), no. 3, iyad049.

[YD21] Rasmus Nielsen Yun Deng, Yun Sung, *The distribution of waiting distances in ancestral recombination graphs*, Theoretical population Biology (2021).

Department of Mathematics, Indiana University, Bloomington, IN 47401
Email address: elena.axinn@tufts.edu