UNIVERSITÄT ZU KÖLN

Coalescent Theory and Yule Trees in time and space

Inaugural-Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät der Universität zu Köln

vorgelegt von *Johannes M. Wirtz*
aus Willich

Köln, 2019

Berichterstatter: (Gutachter) Prof. Dr. Thomas Wiehe

Prof. Dr. Joachim Krug

Tag der mündlichen Prüfung: 09.01.2019

"The Blues is easy to play but hard to feel."

—Jimi Hendrix

Abstract

Mathematically, Coalescent Theory describes genealogies within a population in the form of (binary) trees. The original Coalescent Model is based on population models that are evolving neutrally. With respect to graph isomorphy, the tree-structures it provides can be equivalently described in a discrete setting by the Yule Process. As a population evolves (in time), the genealogy of the population is subject to change, and so is the tree structure associated with it. A similar statement holds true if the population is assumed to be recombining; then, in space, i.e. along the genome, the genealogy of a sample may be subject to change in a similar way.
The two main focuses of this thesis are the description of the processes that shape the genealogy in time and in space, making use of the relation between Coalescent and Yule Process. As for the process in time, the presented approach differs from existing ones mainly in that the population considered is strictly finite. The results we obtain are of mainly theoretical nature. In case of the process along the genome, we focus on mathematical properties of Linkage Disequilibrium, a quantity that is relevant in the analysis of population-genetical data. Similarities and differences between the two are discussed, and a possibility of performing similar analyses when the assumption of neutrality is abandoned is pointed out.

Zusammenfassung

Die Koaleszenztheorie beschreibt Genealogien innerhalb einer Population durch (binäre) Bäune. Die ihr zugrundeliegenden Populationsmodelle beruhen auf der Annahme neutraler Evolution. In Bezug auf Graphisomorphie können die Baumstrukturen, die sie generiert, in diskreter Form durch den Yule-Prozess äquivalent beschrieben werden. Wenn sich eine Population (in Zeit) entwickelt, ändert sich auch die Genealogie der Population, ebenso wie die damit verbundene Baumstruktur. Ähnliches gilt, wenn Rekombination betrachtet wird: Entlang des Genoms (was als eine räumliche Komponente angesehen werden kann), kann sich die Genealogie einer kleinen Auswahl an Individuen ("Sample") auf ähnliche Weise ändern.
Die beiden Hauptschwerpunkte dieser Arbeit sind die Beschreibung der Prozesse, die die Genealogie in Zeit und Raum gestalten, unter Ausnutzung der Beziehung zwischen Koaleszenztheorie und Yule-Prozess. Was den zeitlichen Prozess anbelangt, unterscheidet sich der behandelte Ansatz hauptsächlich darin von bestehenden, dass sich auf die Betrachtung endlicher Populationen beschränkt wird. Die hergeleiteten Ergebnisse sind hauptsächlich theoretischer Natur. Bei dem Prozess entlang des Genoms liegt das Augenmerk auf den mathematischen Eigenschaften einer Größe, die bekannt ist unter dem Namen "Linkage Disequilibrium", und die bei der Analyse populationsgenetischer Daten relevant ist. Ähnlichkeiten und Unterschiede zwischen den beiden Prozessen werden diskutiert, und es wird eine Möglichkeit aufgezeigt, wie ähnliche Analysen durchgeführt werden können, wenn die Annahme der Neutralität fallen gelassen wird.

Acknowledgements

I thank my supervisor, Thomas Wiehe, for giving me the opportunity to do Ph.D-studies in his lab, and for allowing me to bump my nose into each and every coalescent-theoretical wall (tree trunk?) in existence. I thank Joachim Krug, Michael Nothnagel and Peter Heger for agreeing to be part of the examination committee. I thank Filippo, who brought me into the field, and Fabian, who I hope will keep me company. I thank Martina, Kamel, Christopher and the other present and past lab members. I thank my parents for their inexhaustible patience, and my sister, wo, hopefully, will soon be able to refer to me as "Mein Bruder, der Doktor".[1] I thank Lucas, Jan, Vincent, Marvin and the rest of the old mathematics-studying crowd. I thank Sandy, the ukulele girl, Uli, the man-child, Tillsche and die Axt.
And I thank the Mauselwi, who was with me whenever I couldn't tell green field from cold steel rail.

Final remark, 27.01.2019: You are reading the final version of this thesis. I would like to thank the supervisors again for their commentary, especially Joachim Krug, who pointed out a number of nasty typing errors, which greatly inproved the soundness and "flow" of the mathematical parts. Thanks again also to my Mother and the Mauselwi for pointing out spelling errors in the Epilogue.

[1] See "The 'burbs" (*Meine teuflischen Nachbarn*), 1989, starring Tom Hanks and Carrie Fisher.

Contents

List of Symbols and Abbreviations

Frequently used symbols and abbreviations, listed roughly in order of appearance.

$\mathcal{T}_n$	The class (set) of *Yule Trees* on n leaves, $n \in \mathbb{N}$.
$\mathcal{L}_n$	The class (set) of *labelled trees* on n leaves, $n \in \mathbb{N}$.
$\mathcal{G}_n$	The class (set) of *coalescent trees* on n leaves, $n \in \mathbb{N}$.
$\|T\|$	The *size* of a Yule Tree T, equivalent to the number of leaves.
T_S	The subtree of a Yule Tree T *induced* by the restriction of the set of leaves to S.
$\cdot \uparrow \cdot$	(For Yule Trees) Right argument obtained by *random grafting* in left argument.
$\cdot \rightharpoonup \cdot$	(For Yule Trees, boolean) Right argument obtainable by *EMG*-transformation of left argument.
$\cdot \leftarrow \cdot$	(In algorithms) (Re-)Assignment of value of left argument to that of right argument.
$\Pr(A)$	The *probability* of an event A in some specified probability space.
P_X	The *probability distribution* of a random variable X; for $x \in \mathbb{R}$, $P_X(x)$ is defined as $\Pr(X \leq x)$.
$\mathbb{E}(X)$	The *expectation* of a random variable X, defined in terms of a Lebesgue integral by $\int_\Omega X \mathrm{d}P_X$, given convergence of this integral.
$Var(X)$	The *variance* of a random variable X, defined in terms of a Lebesgue integral by $\int_\Omega (X - \mathbb{E}(X))^2 \mathrm{d}P_X$, given convergence of this integral.
e	Euler's number, ≈ 2.71.
$\exp(\lambda)$	Exponential distribution with parameter λ.
a_n	n'th harmonic number, $a_n = \sum_{i=1}^{n} 1/i \approx \log(n)$.
$f(a)$	*Frequency* of the allele a ($f_P(a)$ in a specific population P).
u	(In population models) *Mutation probability* per individual per generation.
θ	(In population models) Population-scaled *mutation rate*.
ρ	(In population models) Population-scaled *recombination rate*.
s	(In population models) *Selection coefficient* of a specific allele (although effect is additive).
MRCA	The *most recent common ancestor*, depending on context of a set of individuals or of the entire population.
SNP	*Single-nucleotide Polymorphism*, usually considered bi-allelic.
bp	*Base Pair*; complementary nucleotides in a strand of DNA.
Kb,Mb	*Kilo (Mega) Base Pairs*; 10^3 (10^6) single *bp*.
LD	*Linkage Disequilibrium*, as a concept.
$D_{\alpha,\beta}$	Original measure of *Linkage Disequilibrium* at two loci α, β.
$r^2_{\alpha,\beta}$	Normalisation of D in the form of a Pearson-Correlation.
$r^2_{S,U}$	Version of $r^2_{\alpha,\beta}$, calculated from topological assignment.

Chapter 1

Introduction

In Biology, the term evolution refers to the ongoing process of animate matter changing its appearance and characteristics while being generated, and in turn, giving rise to new animate matter, which may be subject to similar gradual change. This process can be considered at different levels; for instance, when classifying organisms into species or genera according to certain traits, e.g., the way they reproduce, a species may be thought of as a seperately evolving entity in time. The term Macroevolution refers to the evolutionary process of the entirety of organisms and species, and possible interdependencies that may enmerge between them. On the other hand, the process that a single or few species undergo on relatively short time scale and without regards to the evolutionary process as a whole, is called Microevolution.
Population genetics is the theory of the state of, and change in, genetic composition observable in a "population", which may be though of as a subset, possibly the entirety, of members of a species. As such, the purpose of this field is to provide a mathematical understanding of Microevolution. In the 20th century, many researchers with both mathematical and biological background became invested in developing mathematical models of biological processes; because of that, by now a rich mathematical theory of evolution, and in particular, of population genetics exists. Stochastic processes [Eth11] are used to provide models of the change in genetic composition of a population, particle models describe the theory of its distribution in physical and genotype space (e.g. the *Parabolic Anderson Model* [Kön+09]), results from game theory and calculus predict equilibrium situations between competing individuals, populations and species [BCH18; McA+18], and information and computer science have found use in understanding host-parasite and host-pathogen interactions [NT15]. In the more recent past, the question has been raised whether evolution can even, to some degree, be predicted by combining computational means with modern technology and the possibilities it offers with regard to, e.g., sampling of genetic material.
From a modern perspective, it is hard to believe that evolution is a quite young scientific concept, and was probably nonexistent before the early 19th century, when apparent similarities between dinosaur bones and those of existing reptiles became a subject of study (Gideon Mantell, Richard Owen). Charles Darwin is usually considered the founding father of the theory of evolution in nature. In his book "On the Origin of Species" [Dar59], he published many of the conclusions he had drawn from his travels and investigations, some of which could be considered revolutionary in retrospect, such as that evolution is an intrinsic mechanism to life, and a major determining force of evolution is "natural selection", a somewhat vague term which would later often be paraphrased by "survival of the fittest". The book received a lot of attention already back then, along with much criticism; after all, none of his theories could be soundly "proven" like in other scientific fields. Notably, Darwin himself mentioned that he had no knowledge about how organisms inherit their

traits from their parents, and how evolution would be facilitated by the failure of heredity. It would take 50 years until the scientific community became fully aware of the significance of Gregor Mendel's findings, and another 50 years to get to a unified understanding of heredity and evolution (Huxley: "The modern synthesis" [Hux42])
The only illustration in the "Origin of species" is a diagram, in which the generation of species is represented by a system of "splitting" lines forward in time, which in turn necessarily have to merge into a single line backward in time. The appearance of this drawing has been likened by many people to a "tree", and it has been hypothesized that this single drawing is a major reason that up until today researchers of evolution use such tree-structures as a representation of evolutionary processes. From a macroevolutionary point of view, a line in such a tree may be considered representative of a species, and the splitting pattern of the tree dictates how species are generated out of each other, and which species are ancestral to others. With the advent of modern theory of population genetics, using trees also to represent the evolutionary history of organisms *within* a population, even within a small sample, became a widespread approach, and Coalescent Theory [Kin82; Wak] provided a convincing way of modeling such tree structures. Furthermore, it was discovered that this could be extended across the entire genome, taking the mechanism of recombination into account [Hud83]. Also, as a population changes its composition over time in certain theoretical settings, so does the tree representing its history [PWW09].
Mathematically, the tree-structures encountered in a coalescent-theoretical setting can be described in a combinatorial way. It turns out that a process described by G. Udny Yule [Yul25] in the early 20th century, makes it possible to consider those trees in a discrete setting, which is called the Yule Model. One of the advantages it provides is that the time component of the evolutionary process may be almost completely disregarded and replaced by integer labellings and subdivisions of the trees into layers, at little cost. On top of that, the collection of objects to study becomes finite and enumerable.
In this thesis, we will reiterate Yule's construction and point out several important properties of the tree-structures obtained in the Yule Model (Chapter 2). We will reformulate the argument of David Aldous [Ald00] to show that the Yule Process and Coalescent Theory are indeed related (Chapter 3). After this, we will investigate how "neutral evolution" shapes the discretely-represented genealogy of a population in time (Chapter 4) and how it shapes that of a sample in space, where space is to be interpreted as "along the genome" in a recombining species (Chapter 5). The consideration of the process in time can be used to recover and extend some previous results about genealogical traits of large populations, and uncover some new approaches of estimating others. The consideration of the spatial process, on the other hand, was inspired by problems encountered in the analysis of genomic data. The estimation of the underlying tree structure(s) from such data, incidentally, offers a way of measuring haplotype correlation across the genome similarly to known methods, but with theoretical properties that can prove "favourable" in the application.
Most of the results we obtain regarding the spatial process have been published [WRW18]. Some of the experimental results have also found entry into [Jab+18], currently under review. There exists another preprint ([WW18], under review at the time of completion of this thesis) comprising many of the results about the genealogy in time. Here, we will discuss the mathematics involved in a less comprimized and more intuitive way, and point out some possible extensions. In both cases, one

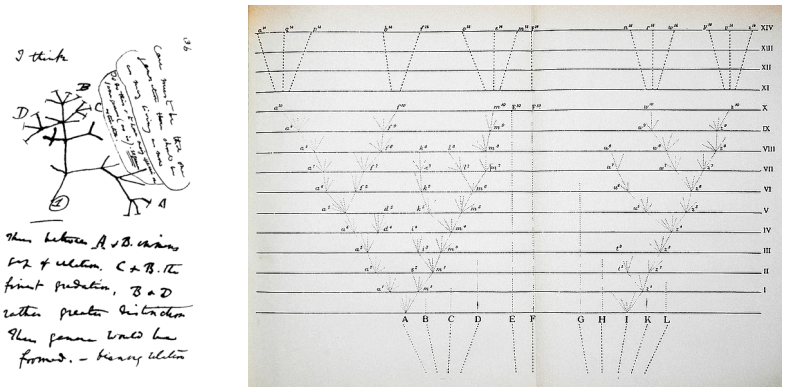


FIGURE 1.1: The "first evolutionary tree" from one of Darwin's notebooks, and the "tree" in [Dar59]

encounters a Markov Chain on discrete tree structures, whose transition probabilities are determined by combinatorial operations performed on the tree. While those operations differ between time and space, there are also unifying features about the respective Chains. We will take this opportunity to discuss the similarities and differences of the two processes in the last chapter.
In the outlook, we will briefly consider a possibility of transfering these processes into a framework that includes the Darwinian mechanism of natural selection. Other opportunities of future research will also be discussed then.

Chapter 2

Yule Trees and related constructions

2.1 The Yule Speciation Model

In 1925, at a time when evolution (due to Darwin) and heredity (Mendel) were still very young concepts, and large parts of the mathematics to describe them had not yet been developed, George Udny Yule published an article called "A mathematical theory of evolution" [Yul25], in which he introduced a speciation model that would later become a cornerstone of theoretical biology. The model featured a set of "genera" and "species" belonging to those genera. Existing species could give rise to new species by throwing "specific mutations", where the new species would be assorted to the same genus as the "parent species". In the same way, new genera could be generated by existing genera throwing "generic mutations", which would result in a new genus containing a single species initially. Both types of mutations were assumed to happen at certain "rates" $s > g > 0$, and, importantly, independently of the size of a genus, i.e., the number of species contained in this genus. To simplify matters, extinction of species and genera was not incorporated. Yule's work may be interpreted as an early example of the use of branching processes in combination with evolutionary rates in theoretical biology.
The motivation of his work was to find a mathematical explanation for the diversity of species in nature. One example "genus" that received particular attention was that of flowering plants, which was already then known to be evolutionarily "young" (having originated around 10^8 years in the past by an old estimate), but also to feature a huge number (160.000) of different species.[1] Yule argued that the generation of species and genera had to be exponential; such that, if there was one species at time zero and two at time one, it would be expected that four, eight, sixteen etc. species would be encountered at times three, four, five etc. On the other hand, it seemed most reasonable to assume that the generation of new species and genera happened independently per genus and species; so some of them would over time throw zero or very few mutations, while others would give rise to many new genera or species by chance, introducing a natural skew in the distribution of species within a genus. He proceeded to calculate the expected frequencies $f_1, f_2, \ldots, f_n, \ldots$ of monotypic, ditypic, n-typic genera after an infinite time, which turned out to depend on the

[1] In current phylogenetic terminology, flowering plants are usually labelled a "taxon", which alone contains over 10^4 genera, and the number of species assorted to the taxon of flowering plants has almost doubled since.

parameter $\rho = g/s$:

$$
\begin{aligned}
f_1 &= \frac{1}{1+\rho} \\
f_2 &= \frac{1}{1+\rho}\frac{\rho}{1+2\rho} \\
&\vdots \\
f_n &= \frac{1}{1+\rho}\frac{\rho}{1+2\rho}\cdots\frac{(n-1)\rho}{1+n\rho} \\
&\vdots
\end{aligned}
$$

This became known as the *Yule-Simon Distribution*, after Yule and Herbert Simon, who later picked up this approach. It is also due to Simon that the speciation model is called *Yule Model*. Nowadays, it is also often labelled *preferential attachment*, due to the fact that in a short interval of time, a genus of many species is more likely to be affected by a specific mutation than a genus that contains only few. It should be stressed, though, that inside a genus, each species is equally likely to throw a specific mutation, and each genus is equally likely to throw a generic one.
Within a genus, it is rather intuitive to associate the process with a tree-like structure "growing" in time. Each species may be represented by a line (a "lineage") running in some direction representing time, and when a species experiences a specific mutation, the line splits into two. Then, at time zero, there is only one line, and at time t, there is a number $l(t)$ of lines, following a distribution depending on t. One might also consider conditioned versions of this process, where at time t some number k of species is assumed to exist, or consider discrete time and exactly one specific mutation in each time step.
It turns out that the Yule Model, or (to stress the stochastic aspect) Yule Process or variants of it are engrained and can be recovered in many modern approaches of evolutionary biology. In particular, the tree structures such processes generate are powerful tools for the analysis of genealogical traits within populations.
In this chapter, we will obtain a formal description of those tree structures by combinatorial means, enumerate them and point out some important properties of their distribution. We will also see that the Yule Process is not the only way to generate these structures, which will become important in Chapter 4. In the last two sections, we will consider a class of slightly extended tree structures, but with similar properties to those obtained under the Yule Model.

2.2 Trees generated under the Yule Process

Perhaps the most basic version of the Yule Process is without generic mutations and with only one genus containing one species at the beginning, in discrete time with one specific mutation per time step, and stopped as soon as a certain number $n \in \mathbb{N}$ of species exists. This version (see also [SM01]) has been used, implicitly or explicitly, many times throughout the literature of theoretical biology. Procedure 1 is an algorithmic representation of this process:

PROCEDURE 1: Discrete Yule Process

1: Start with a tree consisting of one single leaf node ι.
2: **while** Tree has $k < n$ leaves **do**
3: Choose one leaf ι uniformly, label it by the current total number of leaves, turn it into an internal node ν with label k and append two new leaves to it.
4: **end while**
Output: Tree with n leaves

Figure 2.1 outlines some possible runs of the discrete Yule Process for small n. Let the output of such a procedure be denoted by T. T can be interpreted as a *tree* , i.e., a connected acyclic graph with labeled vertices called "internal nodes", unlabelled vertices called "leaves" and edges, called branches, which are generated by the process whenever leaves are appended to an internal node. We assume that appending is graphically carried out in downward direction and in such a way that T is a plane graph.
The final number $n \in \mathbb{N}$ of leaves corresponds to the number of iterations of the procedure plus one, and will be referred to as the *size* of T. Since the object generated by the process is a tree not only at termination, but also after each iteration, we use $\iota = T^{(1)}, \cdots, T^{(n)} = T$ to denote the trees at intermediate stages.
Any tree T of size n has n leaves (nodes of degree 0 or 1) $\iota_1, \ldots, \iota_n$, and $n-1$ *internal nodes* $\nu_1, \ldots, \nu_{n-1}$, which are nodes of degree 2 or 3. We identify the index k of ν_k with the label of ν_k. If $n \geq 2$, the internal node ν_1 is of degree 2 and is called *root* of T, while all other internal nodes are of degree 3. T furthermore has exactly $2n-2$ branches. One may think of the branches as directed from top to bottom; in this case, all internal nodes are of out-degree 2 and leaves are of out-degree 0. Because of this, we may refer to T as a *rooted binary tree*. Other than that, directedness of the branches is not of too much importance.
For any leaf $\iota \in \{\iota_1, \ldots, \iota_n\}$, when moving downward on the unique path from ν_1 toward ι, the sequence of labels of internal nodes on this path is increasing; hence such trees are also called *binary increasing trees*. Suppose further that all n leaves of T are drawn on the same vertical "height" 0, and all internal nodes ν_k on height $n-k$. Then, the leaves $\iota_1, \ldots, \iota_n$ are implicitly ordered horizontally and can be identified with their (integer) position.
Furthermore, under this assumption T divides the plane into n layers $l_1, \ldots, l_n$, where

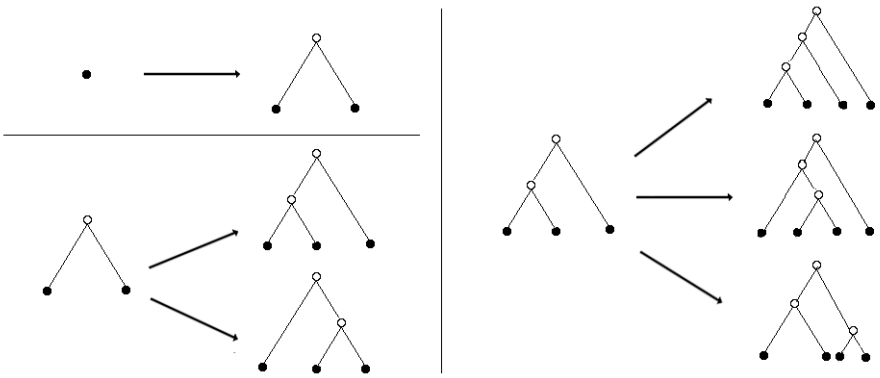


FIGURE 2.1: Some possible iterations of the Yule tree-generating procedure

layer $l_k, k = 2, \dots, n-1$ is vertically restricted by the heights of v_{k-1} and v_k. Layer 1 extends upwards to infinity from the root's height, and layer n from height 1 to 0. If $k \geq 2$, the k'th layer of T is the layer which is crossed by precisely k branches. This notion can be extended to layer 1 by assuming that it contains an *imaginary branch* extending from the root upwards.
Also, it turns out extremely convenient to think of a branch β as a composite of *branch segments*, where a segment only extends over one layer. Then T contains $1 + 2 + \dots + n = \frac{n(n+1)}{2}$ such segments (counting the imaginary branch as a single segment). We denote them by $b_1, \dots, b_{\frac{n(n+1)}{2}}$ from top to bottom and left to right (see Figure 2.2).
Having described the objects generated by the discrete Yule Process, we define:

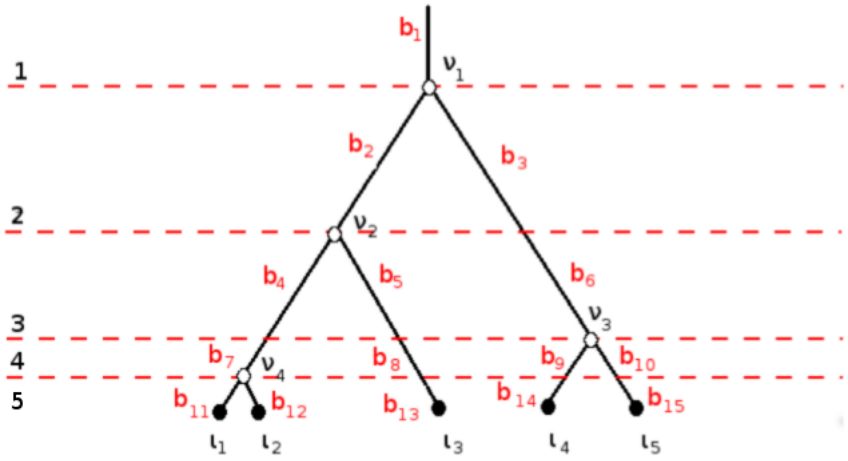


FIGURE 2.2: A Yule tree of size 5 with all layer, branch segment and node labellings depicted

Definition 1.

1. An object T generated by the discrete Yule Process is called a *Yule Tree*.
2. For two Yule Trees T, T' of size n, we write $T = T'$ if and only if for all pairs of internal nodes v_k, v'_k, in T, T' of the same label i, the indices (i_1, i_2) of the branch segments b_{i_1}, b_{i_2} below v_i in T are equal to the indices (i'_1, i'_2) of the branch segments $b'_{i'_1}, b'_{i'_2}$ below v'_i in T'.
3. $\mathcal{T}_n$ denotes the set of all possible Yule Trees of size n, i.e., the set of equivalence classes of n-sized Yule Trees with respect to the relation "=".

In order to carry out the following calculations, we define some additional notation:

Definition 2.

1. Let the function $\sigma_T(i)$ denote the leaf ι of $T^{(i)}$ chosen in the i'th iteration of the Yule Process generating a Yule Tree T of size n.
2. Let the function $l(b)$ denote the layer over which a segment b extends
3. Let the function $k(v)$ denote the label of an internal node v in a Yule Tree T.

4. Let the function $h(\iota)$ yield the horizontal position of a given leaf ι of a Yule Tree T (in terms of an integer $1, \dots, n$).

For instance, we easily see by induction on n:

Proposition 1. *For two n-sized Yule Trees T, T', $T = T'$ holds if and only if for $i = 2, \dots, n$, we have $h(\sigma_T(i)) = h(\sigma_{T'}(i))$.*

In other words, one Yule Tree equals another if in each iteration of procedure 1, the same leaf with respect to horizontal position is chosen. Not only is this much more intuitive compared to the exact but somewhat clumsy Definition 1, it also facilitates the enumeration of the elements of $\mathcal{T}_n$: After iteration $0 \leq k \leq n-2$, there exist $k+1$ leaves and therefore $k+1$ possibilities in the next iteration. It follows that there are $(n-1)!$ possibilities of generating a tree T of size n. Because leaves are chosen uniformly in each iteration, each tree is generated with probability $\frac{1}{(n-1)!}$.
There are as many Yule Trees of size n as there are permutations of size $n-1$; in fact, by traversing trees and observing labels of internal nodes, it is possible to construct a bijection between Yule Trees and permutations. This result also agrees with a result of enumerative combinatorics, which states that there are $(n-1)!$ binary increasing trees of size n [FS09].
In a similar manner, we easily obtain one of the results of [SM01]:

Proposition 2. *Let ω denote the number of leaves appended somewhere below the left branch of the root in a random Yule Tree T of size n. We have $\Pr(\omega = m) = \frac{1}{n-1}$ for $m \in \{1, \dots, n-1\}$*

Proof. If there are m leaves found on the left side below the root of T, there must have been $m-1$ iterations of the Yule Process that have targeted some leaf on the left side of T, and therefore $m-1$ internal nodes on the left side of T; this means $n-m-1$ internal nodes are found on the right side.
The number of possiblilities of assigning labels to those nodes on the left is $\binom{n-2}{m-1}$, since the root is always labelled 1. The number of possibilities of choosing leaves on the left side during the entire process generating T is $(m-1)!$, and for the right side, this number is $(n-m-1)!$. Hence, the total number of possibilities to generate a Yule Tree with m leaves on the left side is

$$\binom{n-2}{m-1}(m-1)!(n-m-1)! = (n-2)!$$

and since all possibilities are equally likely, we have

$$\Pr(\omega = m) = \frac{(n-2)!}{(n-1)!} = \frac{1}{n-1}$$

□

The "uniformity" of the number of leaves on the left and right sides of a Yule Tree, while being a rather simple principle, proves useful in the analysis of several stochastic processes of mathematical population biology; we will encounter a couple of such instances in the following chapters.

2.3 Induced Subtrees

Let S denote a set of leaves of some Yule tree T of size n. Connecting all leaves of S according to the branching pattern of T generates another tree T_S on $|S|$ leaves,

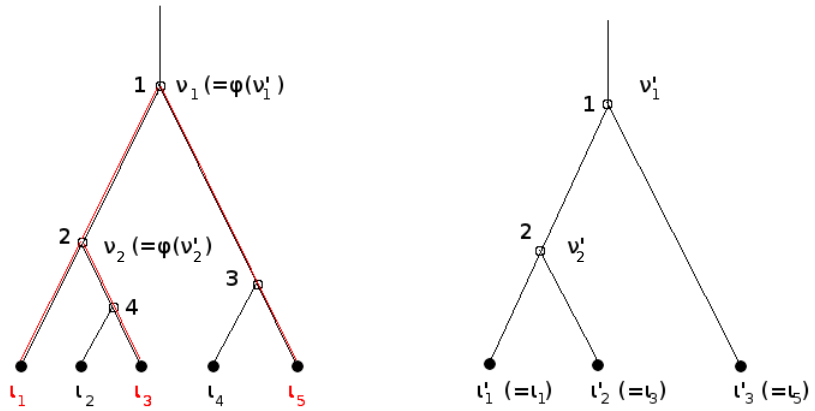


FIGURE 2.3: A Yule tree of size 5 and the induced subtree of leaves $\iota_1, \iota_3, \iota_5$.

where $|S| - 1$ internal nodes of T are preserved. If we label the internal nodes of T_S by $1, \dots, |S| - 1$ such that their relations with respect to height are preserved from T, and subdivide the tree into layers as described in section 2.2, we can identify T_S as an object of $\mathcal{T}_{|S|}$. Each leaf ι' in T_S corresponds to ("equals") some leaf $\iota \in \{\iota_1, \dots, \iota_n\}$ of T, and the horizontal order of leaves in T_S is in accordance with that in T. Similarly, each internal node ν' in T_S is representative of some internal node ν in T, with $k(\nu') \leq k(\nu)$.

Definition 3. For any n-sized Yule Tree T and $\emptyset \neq S \subseteq \{\iota_1, \dots, \iota_n\}$:

1. The object T_S is called the (S-)*induced subtree* of T.
2. For an internal node $\nu' \in \{\nu'_1, \dots, \nu'_{|S|-1}\}$ of T_S, let $\phi(\nu')$ denote the internal node of T that is represented by ν' in T_S.
3. For all $j = 1, \dots, |S| - 1$, let $\tau(j) \in \{1, \dots, n-1\}$ denote the label of $\phi(\nu'_j)$ in T

See Figure 2.3 for an example. If $S = \{\iota\}$ for some single leaf ι of T, T_S equals the tree of size 1 consisting just of ι, and $T_{\{\iota_1,\dots,\iota_n\}} = T$.
In the following Lemma, we explore the relationship between the distributions of Yule Trees and random induced Subtrees of Yule Trees. It turns out that under the correct assumptions, they are actually in agreement with each other.

Lemma 1 (Sample-Subtree Invariance of Yule trees)**.** *Let T be a random tree of size n generated by the Yule Process, and $S \subseteq \{\iota_1, \dots, \iota_n\}$, $|S| = k$ a* random *subset of leaves. Then*

$$\forall \tilde{T} \in \mathcal{T}_k : \Pr(T_S = \tilde{T}) = \frac{1}{(k-1)!} \tag{2.1}$$

Proof. We show that we can treat T_S as a tree generated by the Yule Process. Since this is obviously true for $|S| = 1$ (or $S = 2$), we apply induction on k.
Let $S = \{\iota'_1, \dots, \iota'_k\}$. Tracing back the iterations $l = n, \dots, \tau(|S| - 1)$ of the process generating T, for each $\iota'_j \in S$ there is a unique leaf $\iota_j^{(l)}$ of $T^{(l)}$ such that either $\iota'_j = \iota_j^{(l)}$ or ι'_j is appended below $\iota_j^{(l)}$ by one or more Yule iterations. In $T^{(\tau(|S|-1)-1)}$, a leaf $\iota^* = \sigma_T(\tau(|S| - 1))$ is turned into $\phi(\nu_{|S|-1})$ in iteration $\tau(|S| - 1)$ and two of the leafs $\iota_m^{(\tau(|S|-1))}, \iota_{m+1}^{(\tau(|S|-1))}$ that are the correspondents of ι'_m, ι'_{m+1} in $T^{(\tau(|S|-1))}$ are appended

below.
Consider the set $S' = \{\iota_1^{\tau(|S|-1)}, \ldots, \iota_{m-1}^{\tau(|S|-1)}, \iota^*, \iota_{m+2}^{\tau(|S|-1)}, \ldots, \iota_k^{\tau(|S|-1)}\}$. Because of the established correspondence of internal nodes between T_S and $T_{S'}^{(\tau(|S|-1)-1)}$, T_S is created out of $T_{S'}^{(\tau(|S|-1)-1)}$ by turning ι^* into an internal node and appending two new leaves. If ι^* is chosen uniformly from S', then this simply corresponds to one Yule iteration. We verify this, writing $\Pr(\sigma_{T_S}(|S'|) = \iota^*)$ for the probability that $\iota = \iota^*$ for $\iota \in S'$:

$$\begin{aligned}\Pr(\sigma_{T_S}(|S'|) = \iota^*) &= \Pr\left(\sigma_T(\tau(|S|-1)) = \iota^* | \sigma_T(\tau(|S|-1)) \in S'\right) \\ &= \frac{1/\tau(|S|-1)}{|S'|/\tau(|S|-1)} \\ &= \frac{1}{|S'|}\end{aligned}$$

In addition, the fact that i^* is chosen uniformly from S' implies that S' can be treated as a set of size $k-1$ that is randomly chosen from the leaves of $T^{(\tau(|S|-1)-1)}$. By induction hypothesis, the induced subtree $T_{S'}^{(\tau(|S|-1)-1)}$ is then a random Yule tree of size $k-1$ and generated by $k-2$ iterations of the Yule Process. Since the last step from $T_{S'}^{(\tau(|S|-1)-1)}$ to T_S can be interpreted as a $k-1$'th iteration, we conclude that the process generating T_S is a Yule Process of $|S|-1 = k-1$ iterations. □

The assumption that S is random can be weakened to some extent; however, if we fix the indices of $\iota \in S$, this statement is not true in general any more.
The equivalence between n-sized Yule Trees and random n-sized induced Subtrees of Yule Tress of size $m \geq n$ is a form of what one might call "stochastic self-similarity", in that a random substructure of a random object is generated by the same stochastic process as the object itself. This constitutes an important feature of Yule Trees and distinguishes them from other combinatorial tree classes; for instance, *Catalan Trees* (see [FS09]) do not have this property.

2.4 The Random Grafting Operation

The discrete Yule Process is the natural, but not the only way of generating the uniform distribution on $\mathcal{T}_n$. Suppose $T \in \mathcal{T}_n$ is a Yule tree of size n. Instead of applying an iteration of the Yule process, T can also be transformed into a tree of size $n+1$ by *random grafting* (2) a new branch leading to a leaf into T.

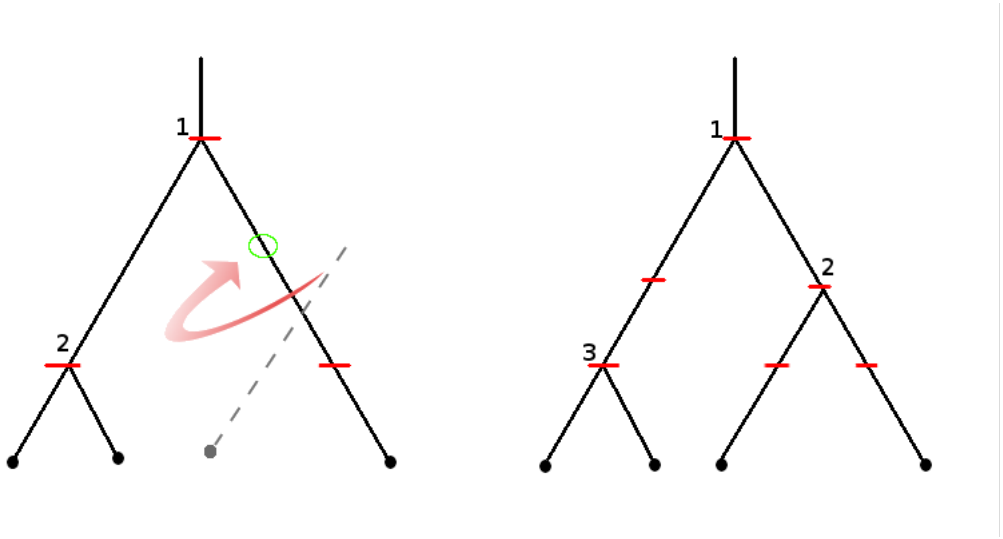


FIGURE 2.4: The regrafting operation 2 performed on the branch segment with the "o" mark, transforming the 3-sized tree on the left into a tree of size 4.

PROCEDURE 2: Random Grafting Operation

Input: Yule Tree T of size n

1: Choose a branch segment b uniformly from all $\frac{n(n+1)}{2}$ possible segments and an "orientation" $\chi \in \{left, right\}$ uniformly ▷ including the imaginary branch
2: Split all branch segments $b', l(b') = l(b)$ into two separate branch segments ▷ forming an additional layer
3: Between the two pieces $b^{(1)}, b^{(2)}$ resulting from splitting b, place a new internal node v with label $l(b)$.
4: Increase the labels of all internal nodes in layers $k > l(b)$ by one;
5: At v, append a new branch β consisting of $n - l(b) + 1$ segments and ending in a new leaf ι, to the left or right depending on χ;
6: $\hat{T} \leftarrow T$

Output: Tree $\hat{T}$ with $n + 1$ leaves

Note that the choice of branch orientation χ determines the horizontal position $h(\iota)$ of the new leaf in $\hat{T}$. A possible realization of procedure 2 is depicted in Figure 2.4. Applying procedure 2, we obtain an object $\hat{T} \in \mathcal{T}_{n+1}$. We write $T \uparrow \hat{T}$ if $\hat{T}$ was constructed from T by random grafting. In total, there are $k(k+1)$ possibilities (b, χ) of performing a grafting in T of equal probability, and unique with respect to which leaf and internal node of $\hat{T}$ they generate. However, different grafting operations on T may generate the same object $\hat{T}$.

The relation between grafting operation and the original Yule Process is described by the following Lemma:

Lemma 2 (Piecewise Recovery by Grafting). *Let T be a random tree of size n, $S = \{\iota'_1, \ldots, \iota'_{k+1}\} \subseteq \{\iota_1, \ldots, \iota_n\}$ a set of leaves chosen uniformly without replacement, and $\iota' \in S$ chosen uniformly. Then*

$$\forall T' \in \mathcal{T}_k, T'' \in \mathcal{T}_{k+1} : \Pr(T_S = T'' | T_{S \setminus \iota'} = T') = \Pr(T' \uparrow T'') \tag{2.2}$$

Proof. Let $l \in \mathbb{N}_0$ denote the number of graftings that can be performed on T' to generate T'', thus $\Pr(T' \uparrow T'') = \frac{l}{k(k+1)}$. On the other hand,

$$\Pr(T_S = T'' | T_{S\setminus\iota'} = T') = \frac{\Pr(T_S = T'', T_{S\setminus\iota'} = T')}{\Pr(T_{S\setminus\iota'} = T')}$$

and by Lemma 1, $\Pr(T_{S\setminus\iota'} = T') = 1/(k-1)!$. Let $m \in \mathbb{N}_0$ denote the number of leafs $\iota' \in S$ such that $T_{S\setminus\iota'} = T'$. Since each tree $\tilde{T} \in \mathcal{T}_{k+1}$ is equally likely to be the induced subtree T_S and $\iota' \in S$ is chosen uniformly, we have

$$\Pr(T_S = T'', T_{S\setminus\iota'} = T') = \frac{m}{k!(k+1)}$$

and thus $\Pr(T_S = T'' | T_{S\setminus\iota'} = T') = \frac{m}{k(k+1)}$.
Let $\iota' \in S$ such that $T_{S\setminus\iota'} = T'$, and ν' the internal node ι' is appended to. There exists exactly one tuple (b, χ) such that, performing the associated grafting operation in T', we obtain T'', the leaf generated by the operation occupies the position of ι' in T'', and the internal node generated by it carries the label of ν'. Conversely, each tuple (b, χ) such that the associated grafting operation on T' yields T'' generates a unique leaf ι^* with respect to horizontal position and an internal node ν^*. Then, there exists a unique $\iota' \in S$ that occupies the position of ι^* in T_S, and since $T'' = T_S$, the induced subtree $T_{S\setminus\iota'}$ of T_S equals T'. Therefore, $m = l$ holds, which ends the proof. □

We immediately conclude

Corollary 1. *The distributions of n-sized Yule trees generated under the Yule Process and generated by random grafting are equal, therefore*

$$\Pr(T | T \text{ generated by random grafting}) = \frac{1}{(n-1)!}$$

Proof. This follows by induction on n, making use of Lemma 2. □

2.5 Labelled Trees

In this section, we will consider a slightly different class of tree objects, but which can be generated in a similar way making use of a principle similar to random grafting. The procedure to generate a tree object of this class on n leaves is similar to random grafting in Yule Trees (Procedure 3).

PROCEDURE 3: Random Grafting in Labelled Trees

1: Start with a tree consisting of one single leaf node ι_1 and a branch attached on top of ι_1.
2: **while** Tree has $l < n$ leaves **do**
3: Choose a branch segment b uniformly from all $\frac{l(l+1)}{2}$ possible segments uniformly;
4: Split all branch segments $b', l(b') = l(b)$ into two separate branch segments; ▷ forming an additional layer
5: Between the two pieces $b^{(1)}, b^{(2)}$ resulting from splitting b, place a new internal node ν with label $l(b)$.
6: Increase the labels of all internal nodes in layers $k > l(b)$ by one;
7: At ν, append a new branch β consisting of $n - l(b) + 1$ segments and ending in a new leaf ι_{l+1};
8: **end while**
Output: Tree with $n + 1$ leaves

Tree objects generated by procedure 3 differ from Yule Trees mainly in two aspects: First, branches do not feature an orientation, as we observe it under the Yule Model. As a consequence, exchanging of subtrees or repositioning of leaves below some internal node does not alter the tree in the sense of procedure 3. Secondly, the leaves are labelled too, in such a way that their labels may be interpreted as names or other kinds of identifiers; they are thus distinguishable, so two trees might be congruent with respect to branching pattern and internal nodes, but will still be treated as different objects if the labels of leaves do not match. We may imagine the tree drawn in a way that the leaf labels are ordered, e.g. from left to right, but possibly at the cost of planarity. Therefore, we usually do not consider it as embedded in the two-dimensional plane.

Definition 4.

1. A tree L on n leaves generated according to procedure 3 is called a *Labelled Tree*.
2. The set $\mathcal{L}_n$ is the set of all labelled trees L of size n.

Labelled trees may be subdivided into layers, and their branches may be interpreted as composites of branch segments, similarly to Yule Trees, but it is important to keep in mind that we may not assign indices to branch segments as easily as in Yule Trees because of the missing orientation. Notably, there exists a single branch segment on top of each tree by construction, playing the role of the imaginary branch in a Yule Tree.
There exists one tree of size 1. As we can see from the algorithm, there are $\frac{l(l+1)}{2}$ possibilities of turning a tree of size l into a tree of size $l + 1$; and similarly to the considerations in section 2.2 we can convince ourselves that each possible sequence of branch segment choices in the iteration generates a unique tree object. Iterating, we obtain

$$|\mathcal{L}_n| = n!(n-1)!2^{n-1} \tag{2.3}$$

as the total number of labelled trees of size n (see also [Mur84]). The presence of the term $(n-1)!$ is an indication that there is some connection between $\mathcal{L}_n$ and $\mathcal{T}_n$. In fact, with respect to topology, both classes are equivalent, meaning that any tree topology is contained at equal proportion in $\mathcal{L}_n$ and $\mathcal{T}_n$. At close inspection,

this seems rather obvious; however, a formal proof of this requires much technical detail and the consideration of equivalence classes of binary trees with respect to graph isomorphy; for simplicity's sake, we sketch this in the following by constructing a random Yule Tree out of a uniformly chosen labelled tree without altering the branching pattern, and showing that the probability distribution on $\mathcal{T}_n$ under this randomized mapping is also uniform.
Let L denote a labelled tree of size n. Remove the labelling on the leaves, and for all internal nodes $v \in \{v_1, \dots, v_k\}$ choose one of the two branches $b_v \in \{b_v^1, b_v^2\}$ appended to it with equal probability. Let the chosen branch b_v point to the left and the other one to the right. With this random transformation, L is turned into a Yule Tree, which we denote by $\chi(L)$ (The reason to use χ is to indicate that this function essentially assigns a random branch orientation to L). Then, we state

Lemma 3. *For any Yule Tree $T \in \mathcal{T}_n$, we have*

$$\Pr(\chi(L) = T) = \frac{1}{(n-1)!}$$

Proof (Sketch). Reiterate the sequence of random graftings (see procedure 3) that were used to generate L. Generate a second object L', where the same graftings are performed, only that the leaf present in the beginning is unlabelled, no labels are given to the inserted leaves and the orientation on the branches induced by $\chi(L)$ are imposed on L' right away. It is obvious that L' can be interpreted as a Yule Tree of size n, and moreover, $\chi(L) = L'$.
But if no labels are assigned to the leaves and a branch orientation at each grafting is chosen uniformly, each step in generating L' is a random grafting operation (procedure 2) on a Yule Tree. We know from Corollary 1 that Yule Trees generated by successive random graftings are uniform. Thus $\Pr(\chi(L) = T) = \frac{1}{(n-1)!}$.
□

One conclusion we may draw from this right away is that a slightly modified version of Proposition 2 also holds for the class of labelled trees: If we randomly pick one branch b extending from the root of a random labelled tree L of size n, the number of leaves we find below b is $m \in \{1, \dots, n-1\}$ with uniform probability.
To avoid confusion, we close this section remarking that several different names have been used throughout the literature for labelled trees, e.g. "dendrograms" [Mur84], "totally-ordered phylogenetic trees" [Son06], or also simply "phylogenetic trees" [Ald01]. In [WRW18], we used the term "coalescent tree topologies" to describe them; the reason will become clear in the next chapter.

2.6 Pruning and Regrafting

In many ways, Yule Trees are the more refined combinatorial class to consider; they sport a bijection to permutations and feature an implicit planar embedding due to the implicit branch orientation while being of overall smaller number than labelled trees. The reason why considering labelled trees is still useful often is that some operations on labelled trees can be realized in a more meaningful way than on Yule Trees. One example is the Prune-Regraft operation.

PROCEDURE 4: Subtree Pruning and Regrafting

Input: Labelled tree L of size n, branch segment b

1: Choose a branch segment b' uniformly from all $\frac{l(b)(l(b)+1)}{2}$ branch segments in layers of height less than or equal to $l(b)$ uniformly;
2: Remove the internal node v from which the branch containing b originates;
3: Decrease the label of all internal nodes $v', k(v') > k(v)$ by one;
4: Split all branch segments $b'', l(b'') = l(b')$ into two separate branch segments
5: Between the two pieces $b^{(1)}, b^{(2)}$ resulting from splitting b', place a new internal node v^* with label $k(v^*) = \max\{k(v') : k(v) < l(b)\} + 1$.
6: Re-attach the branch containing b at v^*;
7: Increase the label of all internal nodes $v' \neq v^*, k(v') \geq k(v^*)$ by one;
8: Readjust the segmentation of branches to reduce the number of branch segments to $\frac{n(n+1)}{2}$ again;
9: $\hat{L} \leftarrow L$

Output: Labelled tree $\hat{L}$

In simple terms, in Procedure 4 the subtree below b is cut off and reattached at some other point of height less than or equal to $l(b)$. Because of that, the procedure is also called *Prune-Regraft Operation* for short. This operation has a biological interpretation in the context of recombination [EW06], to be discussed in Section 3.3 and Chapter 5. Also, it is linked to a well-known problem of computer science [Son06]. The reason to define this operation on labelled trees instead of Yule Trees is that while leaves may be considered as horizontally ordered in a Yule Tree, they are still effectively unlabelled and one might be unable to tell which leaves have been affected by such an operation by looking at the tree before and after. In a labelled tree, on the other hand, the leaf labels provide unique identifiers for the leaves, making it possible to determine which leaves were moved, and where.

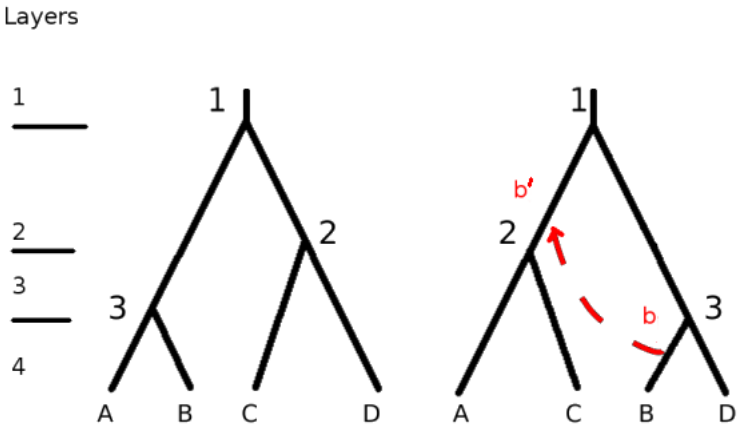


FIGURE 2.5: A labelled tree (*left*) of size 4, and a prune-regraft operation performed on the same tree (*right*). The subtree below the branch segment b is cut off and re-attached at the branch segment b'.

Chapter 3

Theory of large populations

3.1 Fundamental Models of Evolution

Population models are developed with the aim of understanding evolution in a mathematical way. Obviously, a model needs to be kept as simple as possible to provide a degree of mathematical manageability, and no model is an exact depiction of the ramifications and processes in reality. While there exists a broad range of approches of modeling a population, in most of the cases, evolution is modeled via a time-dependent stochastic, often Markovian process. One big concern is the size of a population; there are models assuming a "continuum" of individuals (such as the Hardy-Weinberg Model [Har08; Wei08][1]), in others individuals form a set of finite or countable entities. A follow-up question to this is whether the population size is variable - in nature, this seems almost inevitable because it is hard to imagine that reproduction and death always keep each other at an exact equilibrium. Many finite-population models, however, make this assumption with the hope that if populations size does not vary too much in reality, the theoretical results will still be valid in an approximate sense.
A good way of modeling a finite population of fixed size N is an *ordered multiset*, where individuals are represented by the elements this set contains. In such a set, an individual has an assigned position, and the same element may be contained multiple times, which can be interpreted as the genotype of certain individuals being equal. Over time, the composition of this set is changed gradually, according to a specified mechanism that reflects reproduction and death of individuals.
The elements representing individuals can be almost arbitrary objects $x_1, \dots, x_N$, however, it usually suffices to think of them as types, words (resembling genomic composition in terms of nucleotides) or differently colored atoms. Reproduction will be realized by generating exact copies of individuals. If an individual x is represented by the same element as another individual y, we will write $x = y$. Note that "=" thus extends to an equivalence relation on P. To avoid confusion between cases where x, y are copies and where they actually denote the same individual in a population, we will write $x \equiv y$ to denote the latter case. $x \equiv y$ implies $x = y$.

Definition 5. A population P of size N is an ordered multiset $P = \{x_1, \dots, x_N\}$ of elements $x_1, \dots, x_n$.

If there are two (k) elements a, b of which all x_i in P are copies, we call P a two-allelic (k- or multi-allelic) population. Over time, the number of differing types in a population may increase, since evolution is also driven by random changes to genomic material, in addition to reproduction and heredity. It is therefore desirable

[1] W. Weinberg described this model independently of G. H. Hardy, the latter being also famous for his work with S. Ramanujan.

that a population model is capable of incorporating a mechanism of mutation. The more drastic changes that can affect an organism's genotype, such as copying, reversals and alteration of genomic *3d*-structure, are hard to represent accurately and in simple terms. Luckily, small changes like single-nucleotide differences between parent and offspring, which may happen due to errors and inaccurate DNA repair, or environmental factors (e.g., radiation), can be realized much more easily. Usually, we assume that mutations only happen at the birth of an individual with some probability $0 \leq u \leq 1$, and that it carries its genotype for the entirety of its life span. If some individual x produces an offspring individual x' which is affected by a mutation, we assume that this automatically entails $x' \neq y$ for all other individuals $y \not\equiv x'$. Importantly, we also assume that no further mutation on some offspring individual x'' of x' may ever yield $x'' = y$. Mutations can therefore not be reverted, and each mutation creates its own unique genotype. This is known as the assumption of *infinite sites* [Gil98]. In reality, while it is possible that one mutation targets the same nucleotide position of another mutation, the probability of observing this is extremely low, as mutations are rare, while the total number of nucleotide positions in an individual's DNA may be enormous.[2] In human sampling data, one indeed observes that the vast majority of variable nucleotide positions usually feature only two different nucleotides [HE10] and most of the variation at the nucleotide level is made up of such *bi-allelic Single-Nucleotide Polymorphisms* (to which we usually refer as *SNP*s for short).
We now consider a very simplistic way of creating a new population out of an existing one:

PROCEDURE 5: Random Multinomial Sampling

Input: Population P of size N
1: **for** $i = 1, \ldots, N$ **do**
2: Choose one element x of P uniformly and create a copy $x'_i = x$
3: **end for**
4: $P' \leftarrow \{x'_1, \ldots, x'_n\}$
Output: New population P' of size N

If P is two-allelic with types a, b, procedure 5 is also called *binomial sampling*. Indeed, let

$$f_P(a) = \frac{|\{x_i \in P | x_i = a\}|}{N}$$

denote the frequency of a-type individuals; then the probability $\Pr(f_{P'}(a) = k/N)$, $k = 0, \ldots, N$ is that of a binomial distribution:

$$\Pr(f_{P'}(a) = k/N) = \binom{N}{k} f_P(a)^k (1 - f_P(a))^{N-k}$$

In observing a sequence of populations created out of each other by random sampling, we recover a fundamental model of mathematical Population Genetics:

Definition 6. Let P_0 denote a population. The sequence $(P_i)_{i \in \mathbb{N}}$, where P_{i+1} is the result of random sampling in P_i for all i, is called *Wright-Fisher Process*, and the populations P_i evolve according to the Wright-Fisher Model.

[2] Around $1.2 \cdot 10^7$ *bp* in *S. cerevisiae* (baker's yeast), $3.1 \cdot 10^9$ *bp* in *H. sapiens*, $3.9 \cdot 10^9$ *bp* in *C. carcharias* (great white shark), $1.6 \cdot 10^{10}$ *bp* in *A. cepa* (domestic onion) and an astonishing $1.5 \cdot 10^{11}$ *bp* in *P. japonica* (canopy plant)[Bio18; GVL83; PFL10].

This model is named after Sewall Wright and Ronald A. Fisher, two influential figures of the field in the early twentieth century.
The Wright-Fisher Process is a discrete Markov Chain on populations. One time step is usually called a generation, probably because random sampling can lead to many different outcomes; in the extreme case, one individual may replace the entire rest of the population by its copies within one step. Because of that, we may be left with the feeling that the Wright-Fisher Model features a rather fast "speed of evolution" in general, and it might be desireable to consider another model, which allows us to observe more sublte changes to a population.

PROCEDURE 6: Moran Step

Input: Population P of size N
1: Choose one element x_k of P uniformly and create a copy $x'_k = x_k$
2: Choose one element x_l of P uniformly
3: $P' \leftarrow \{x_1, \dots, x_{l-1}, x'_k, x_{l+1}, \dots, x_N\}$
Output: New population P' of size N

By applying such a Moran Step, one individual x_k is duplicated, thereby reproducing, and one individual x_l is removed ("killed"). Note that $k = l$ is not excluded. The resulting Population Model is named after Patrick A. P. Moran, who was the first to explicitly describe it (see [Mor58]).

Definition 7. Let P_0 denote a population. The sequence $(P_i)_{i\in\mathbb{N}}$, where P_{i+1} is the result of a Moran Step applied to P_i for all i, is called *Moran Process*, and the populations P_i evolve according to the Moran Model.[3]

In the Moran Model, the intuitive assumption would be that what we have before, cautiously, labelled speed of evolution, is less than in the Wright-Fisher Model. A way to mathematically formulate this is to consider the heterozygosity of a population.

Definition 8. The heterozygosity $h(P)$ of a population is the probability that two uniformly chosen individuals $x, y \in P$ are of different genotype, i.e. $x \neq y$.

Consider a two-allelic population P with $0 < f_P(a) < 1, f_P(b) = 1 - f_P(a)$. The heterozygosity of P is $h(P) = 2f_P(a)\,(1 - f_P(a))$ The expected heterozygosity $h(P')$ in P', if we apply a step of the Wright-Fisher Model, can be calculated as follows,

$$\begin{aligned}
\mathbb{E}(h(P')) &= \mathbb{E}\left(2f_{P'}(a)\,(1 - f_{P'}(a))\right) \\
&= 2\left[\mathbb{E}\left(f_{P'}(a)\right) - \mathbb{E}\left(f_{P'}(a)^2\right)\right] \\
&= 2\left[f_P(a) - Var\left(f_{P'}(a)\right) - \mathbb{E}\left(f_{P'}(a)\right)^2\right] \\
&= 2\left[f_P(a) - \mathbb{E}\left(f_{P'}(a)\right)^2 - f_P(a)\,(1 - f_P(a))\,\frac{1}{N}\right] \\
&= h(P)\left(1 - \frac{1}{N}\right)
\end{aligned}$$

[3] Throughout the literature, it appears that the most frequently used version of the Moran Model is one where splitting and killing is initiated according to exponential clocks, such that the process runs in continuous instead of discrete time.

making use of the fact that $f_{P'}(a)N$ is binomially distributed.
Under the Moran model, we have instead

$$\begin{aligned}
\mathbb{E}(h(P')) &= \mathbb{E}\left(2f_{P'}(a)\left(1-f_{P'}(a)\right)\right) \\
&= 2\left[\mathbb{E}\left(f_{P'}(a)\right) - \mathbb{E}\left(f_{P'}(a)^2\right)\right] \\
&= 2\left[f_P(a) - \left(f_P(a)^4 + \left(1-f_P(a)\right)^2 f_P(a)^2\right)\right] \\
&\quad + 2\left[f_P(a)\left(1-f_P(a)\right)\left(2f_P(a)^2 + \frac{2}{N^2}\right)\right] \\
&= 2\left[f_P(a) - f_P(a)^2 - \left(f_P(a) - f_P(a)^2\right)\frac{2}{N^2}\right] \\
&= h(P)\left(1-\frac{2}{N^2}\right)
\end{aligned}$$

In both models, the heterozygosity is expected to decline by a constant factor, but slower by one order of magnitude with respect to N in the Moran case. They are unified, though, by the fact that the equilibrium heterozygosity is zero. Allele frequencies over time are subject to a random fluctuation induced by the respective reproduction mechanisms. This fluctuation, called *genetic drift*, is an intrinsic factor of most finite-size population models, and while it usually operates on large time scales, it is sufficient to have all but one of the allele types purged out of the population by chance alone. In fact, with probability one, within finite time the processes enter a stage where there is one individual x^* in one of the past populations, called *Most Recent Common Ancestor*, such that all individuals of the present are duplicates of x^*, and x^* is the youngest (hence, "most recent") individual with this property. This is a simple consequence of drift and the law of large numbers, but constitutes an essential determinant of the theory of finite-size populations and an important factor in the considerations of section 3.2 and chapter 4.
However, one might note that the different decline rate of $h(P)$ is amendable by considering $N/2$ Moran steps as one generation. Indeed, most of the difference between the two models vanishes with large N and appropriate rescaling. If we consider the stochastic process of the allele frequency $(f_i(a))_{i\in\mathbb{N}}, f_i(a) := f_{P_i}(a)$ under a Wright-Fisher Model, let $N \to \infty$ and consider time in units of $\frac{1}{N}$, the process converges to a continuous-time Markov Process $(f_t(a))_{t\in\mathbb{R}_0^+}$ called Wright-Fisher diffusion, which is characterized by the *SDE*

$$\mathrm{d}f_t(a) = \sqrt{f_t(a)(1-f_t(a))}\mathrm{d}B_t \tag{3.1}$$

where B_t is a standard Brownian motion. It can be shown that the allele-frequency process in a Moran Model also converges to the Wright-Fisher diffusion, if time is rescaled by $\frac{2}{N^2}$ instead of $\frac{1}{N}$. Thus, for our purposes we may treat the Moran and Wright-Fisher Models as almost equivalent if N becomes large. One might note that equation 3.1 does not feature a "drift term" independent of B_t; thus "genetic diffusion" would be the mathematically less ambiguous terminology. However, it has a long tradition in theoretical biology to label random fluctuations in allele frequencies as genetic drift.
The technical details of the above are much more intricate than it may sound here. Unfortunately, a detailed discussion of convergence of stochastic processes, Brownian motions and stochastic differential equations is beyond the scope of this work.

For a more rigorous treatment of this topic, we refer to [Eth11].

3.2 The Kingman Coalescent

By sampling individuals from large populations, decrypting their genetic material and comparing it, researchers aim to investigate the evolutionary background of a population. Particularly the process of decrypting, i.e., sequencing and assembling, used to be a costly and error-prone procedure, and to this day Biologists are keen on achieving maximum informative value while keeping sample-sizes as small as possible. One tool of mathematical population genetics that is very useful in this regard is the Coalescent Process. On the technical side, the Coalescent is a dual process of the Wright-Fisher diffusion (therefore, of the infinite-population limit of both Wright-Fisher and Moran Process), describing the evolutionary history of a sample taken from such a population backward in time. As such, it can be used to make predictions on the properties of such a sample.
In the following, we will consider a Wright-Fisher Population that has been evolving for a long time, which we will indicate by considering the sequence of genealogies $\{P_i\}_{-\infty<i\leq j}, j \in \mathbb{N}$ being the present. Also, the size of the population is $2N$, which is often done througout the literature with diploid organisms in mind; the population instead consists of a collection of haplotypes.

Definition 9. Let x denote an individual of P_j. The *lineage* of x is the sequence $Y(x) := (y_i(x))_{-\infty<i\leq j}$, such that $y_j(x) \equiv x$ and for all $-\infty < i \leq j$ $y_{i-1}(x)$ denotes the individual of generation $i-1$ which $y_i(x)$ is generated of as a copy.

The lineage of an individual denotes the sequence of its ancestors. Now, we consider the lineages $Y(x_1), Y(x_2)$ of two individuals $x_1, x_2 \in P_j, x_1 \not\equiv x_2$ and ask for the probability that those lineages become congruent ("coalesce") at a given point $i^* < j$, i.e. $y_i(x_1) \equiv y_i(x_2)$ for all $i \leq i^*$. We calculate these probabilities one by one backwards in time. The probability $p_1^{(2)}$ that x_1, x_2 are copies of the same individual of generation $j-1$, is the probability that the ancestor of x_2 coincides with that of x_1 by chance, so

$$p_1^{(2)} = \frac{1}{2N}$$

The probability of coalescing two generations in the past is the product of the probability of no coalescence one generation in the past, and coalescence in the following, which is simply

$$p_2^{(2)} = (1-p_1)p_1 = \left(1-\frac{1}{2N}\right)\frac{1}{2N}$$

Continuing in the same way for more generations, we find

$$p_l^{(2)} = \left(1-\frac{1}{2N}\right)^{l-1}\frac{1}{2N}$$

Therefore, the time X until coalescence of two lineages is geometrically distributed with parameter $\frac{1}{2N}$. As $N \to \infty$, the random variable $X/2N$ converges to an $\exp(1)$ distribution. From this, we learn that the eventual coalescence of two lineages can also be observed in the infinite-population limit, and the (rescaled) waiting time to a coalescence is distributed exponentially with parameter 1.
We can repeat this consideration for an arbitrary number k of lineages. It turns out that particularly important in this calculation is the number $\binom{k}{2}$ of possible pairs of

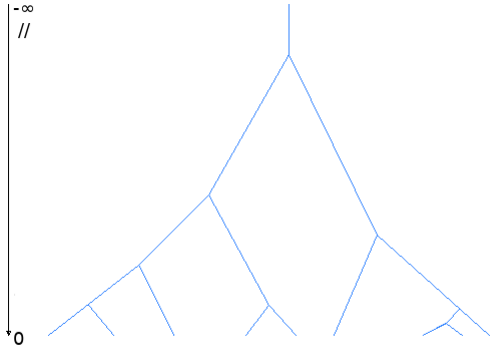


FIGURE 3.1: Example shape of a coalescent genealogy. Note that coalescent events become less frequent while moving backward in time.

lineages, because each pair is equally likely to coalesce first. For sufficiently large N, the probability $p_l^{(k)}$ is reasonably well estimated in the finite-population case by

$$p_l^{(k)} \approx \left(1 - \frac{k(k-1)}{4N}\right)^{l-1} \frac{k(k-1)}{4N} \tag{3.2}$$

and converges, in the infinite-population case, to an $\exp(\binom{k}{2})$ distribution.
It is noteworthy that in the Wright-Fisher Model, multiple lineages may coalesce in one step. However, the probability of this tends to zero sufficiently fast with growing $2N$, so that in the limit, a sample of size n backwards in time experiences $n-1$ distinct coalescent events almost surely. Equation 3.2 determines the distributions of waiting times to the next coalescent event backwards in time. As with the Yule Process in Chapter 2, a natural way of representing this process is to draw a tree-like structure G, where branches represent the lineages, and at each time of a coalescent event, a pair of branches merges into a single branch, until at the end, only one lineage extends into the past to $-\infty$. Such trees represent the genealogy of the sample and are called coalescent trees; we denote the collection of all coalescent trees of n-sized samples by $\mathcal{G}_n$.[4] The process itself is named Kingman's Coalescent, after John Kingman who presented a detailed derivation in 1982 [Kin82], or simply Coalescent Model. It is noteworthy that a coalescent tree of infinitely many lineages exists also for the entire Wright-Fisher Population in the limit of infinite population size because of the quadratic rates of coalescent events. This property is known as "coming down from infinity".[5]
This construction has a lot of useful implications in practice. We will point out some in the following.

Tree Height and Length The time until a set of k lineages collapses to a set of $k-1$ lineages is distributed $\exp(\binom{k}{2})$. The expectation of this random variable is $\frac{2}{k(k-1)}$.

[4] For now, this is best interpreted as an uncountable set. We do not impose a probability distribution on it. One possibility of doing that is to interpret a coalescent tree as a *metric measure space* $G \cong (X, \mu, d)$, see e.g. [DPP15].

[5] There exists a variety of similar processes labelled "Coalescent" (e.g. the β-Coalescent, [BBS07]), which allow multiple mergers and induce different coalescent rates. However, not all of them incorporate the "coming down from infinity"-property.

Given a sample of size n, we can calculate the expected time back to the *MRCA* of the sample by summing up all these expectations from $k = 2$ to n. This time corresponds to the "height" $h(G)$ of the associate Coalescent Tree G. We obtain

$$\mathbb{E}h(G) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2 - \frac{2}{n}$$

The expected total branch length $l(G)$ can be treated similarly. Because there are exactly k lineages during the waiting time for the coalescent event merging them into $k-1$, the derivation changes only slightly:

$$\mathbb{E}l(G) = \sum_{k=2}^{n} \frac{2}{k(k-1)} \cdot k = 2 \sum_{k=1}^{n-1} \frac{1}{k} = 2a_{n-1} \tag{3.3}$$

One might imagine the length of branches as a measure of informative value on the past. A tree of high length is supposed to carry more information on the past than a short tree. On average, by equation 3.3 tree length measured in Coalescent time is $2a_{n-1}$. Importantly, it increases with sample size, but only logarithmically. Tree height instead gets saturated at the value 2. If we interpret $h(G)$ as the amount of time in the past that our sample contains information on, we have to conclude that increasing sample size is not expected to yield a significant increase in this regard; on the other hand, already the genealogy of a sample of size 2 is expected to reach back into the past half as far as the genealogy of the entire population.

Mutations, θ and the frequency spectrum If we incorporate mutations in our population model (see Section 3.1), mutation events will happen with some probability $u > 0$ at each duplication. In order to consider the infinite-population limit, we need to assume $u \to 0$ as $2N \to \infty$ and moreover, that $Nu \to \theta/4$ for some $\theta > 0$; otherwise either zero or an infinite number of mutations occur almost surely in the evolutionary history of a finite sample. But under this assumption, the accumulation of mutations is a Poisson process on each individual lineage of intensity $\theta/2$. θ is called the *population-scaled mutation rate*.
Knowing this, it comes as no surprise that the process of mutation can be realized directly on a Coalescent genealogy. Each branch represents a piece of a lineage, so one simply has to add mutation events on the tree, which may be e.g. drawn as dots; they don't have to be distinguishable because we are assuming an infinite-sites model anyway, and each mutation creates its own distinct haplotype. The way to interpret this in sequencing data is that each *SNP* (recall that most variation is provided by *SNP*'s) results from such a mutation. In fact, one may attempt to reconstruct the genealogy of a sample by phylogenetic means using the *SNP*-pattern observed in this sample.
In the previous paragraph, we calculated the expected total branch length of a Coalescent Tree. If mutation is a Poisson process along all branches of the tree, it follows immediately for the number S of mutations observed on a Coalescent Tree G of size n

$$\mathbb{E}S = \mathbb{E}l(G)\theta = 2a_{n-1}\theta/2 = \theta a_{n-1}$$

For large $2N$, we have $\theta \approx 4Nu$ and the approximation

$$\mathbb{E}S \approx 4Na_{n-1}u$$

In real populations, the parameter θ serves as an important reference value. By our considerations, we directly obtain a maximum-likelihood estimator for θ by counting the number of mutations $\hat{S}$ observed in a sample,

$$\hat{\theta}_w = \hat{S}/a_{n-1}$$

which is known as *Watterson's Estimator*.
Exploiting the tree structure of a coalescent tree, it is possible to calculate the distribution of the "size" i of a mutation, which in this context denotes the number of individuals affected by the mutation. These individuals are precisely the ones found "below" the branch on which the mutation occurs. Let $\xi_i, i \leq n-1$ denote the number of mutations of size i in a coalescent tree of size n. Then

$$\mathbb{E}\xi_i = \theta\frac{1}{i} \tag{3.4}$$

This is known as the *expected site frequency spectrum*, derived very early on by Crow and Kimura [CK70], actually without referring to the Coalescent Model. Interestingly, the probabilities of mutations of some size i resemble the expected waiting times for coalescent events with $i+1$ lineages.

Topological Equivalence of Coalescent and Yule Model Say G is a Coalescent Tree of n individuals from the same population, and we wish to include another, $n+1$'th individual and ask how the tree G' of those $n+1$ individuals looks like. On the stochastic side, G' is a realisation of what is called *conditional coalescent*; the new lineage of the additional individual coalesces at rate $l(t)$, where $1 \leq l(t) \leq n$ is the number of existing lineages t units of time back in the past, with any of these lineages. Assuming that at some time t in the past, there are k lineages of the original sample left to coalesce, and the new lineage has not coalesced with any other lineage yet, we can calculate the distribution of the waiting time until the next coalescent event $\hat{t}_k$ by considering the waiting time t_k until the next coalescent in the original sample and the waiting time c until the new lineage coalesces with one of the original ones. Because t_k and c are independent of each other,

$$\Pr(\hat{t}_k \geq x) = \Pr(t_k \geq x)\Pr(c \geq x) = e^{-\frac{k(k-1)}{2}x}e^{-kx} = e^{\frac{k(k+1)}{2}x}$$

This proves that the waiting time until *any* coalescence is the same as the waiting time in a Coalescent of size $k+1$, owing to the fact that in the latter, there are $\frac{k(k+1)}{2}$ possible pairs of lineages that can coalesce. The Coalescent can thus be described equivalently by subsequently adding lineages, coalescing with the existing lineages at rate proportional to the number of these lineages.
This has another important implication: Suppose G is subdivided into layers $1, \ldots, n$ like a Yule Tree (see Chapter 2) with coalescent events playing the role of internal nodes. The branches of the tree can be subdivided again into $\frac{n(n+1)}{2}$ branch segments (*lineage segments*), all of which extend over one layer. Then, we find that regarding the probability of the new lineage coalescing with a specific lineage segment b of G,

Theorem 1. *Any lineage segment in G is chosen for coalescence with the new lineage with equal probability.*

Proof. Let p_l denote the probability of the new lineage coalescing in some layer l of G. Let also $\tilde{p}_l$ denote the probability of coalescence in layer l conditioned on no coalescence in any layer $l' \geq l$. The duration $dur(l)$ of layer l in G is exponentially

distributed with mean $\frac{2}{l(l-1)}$ There are l lineage segments in layer l, thus coalescence of the new lineage happens with rate l. Therefore,

$$\begin{aligned}
\tilde{p}_l &= \int_0^\infty \Pr(\text{Regrafting in layer } l | dur(l) = t) \mathrm{dPr}_{dur(l)}(t) \\
&= \int_0^\infty \frac{l(l-1)}{2} e^{-t\frac{l(l-1)}{2}} \int_0^t l e^{-ul} \mathrm{d}u\mathrm{d}t \\
&= \int_0^\infty \frac{l(l-1)}{2} \left(e^{-t\frac{l(l-1)}{2}} - e^{-t\frac{l(l+1)}{2}} \right) \mathrm{d}t \\
&= 1 - \frac{l-1}{l+1} \\
&= \frac{2}{l+1}
\end{aligned}$$

If $l = n$, we have $\tilde{p}_l = p_l$. For $l < n$, we obtain iteratively

$$\begin{aligned}
p_l &= (1-\tilde{p}_n) \cdot \cdots \cdot (1-\tilde{p}_{l+1}) \tilde{p}_l \\
&= \frac{n-1}{n+1} \cdot \cdots \cdot \frac{l}{l+2} \frac{2}{l+1} \\
&= \frac{2l}{n(n+1)}
\end{aligned}$$

The statement follows now because there are l segments in layer l and the new lineage coalesces with each of them at equal rate.

□

Theorem 1 essentially tells us that we may construct a Coalescent Tree iteratively by adding new lineages which coalesce with any existing lineage segment uniformly. Labelled trees (see Section 2.5) were generated in exactly the same way, if we disregard the exponentially distributed branch lengths of Coalescent Trees. This shows that the class of labelled trees is equivalent to $\mathcal{G}_n$ with respect to graph isomorphy, and since we have seen in the same section that labelled trees are equivalent to Yule Trees in a similar way, we may conclude that the Coalescent, with respect to graph-theoretical properties, is equivalent to the Yule Process. Therefore, in order to make predictions about combinatorial properties of Coalescent Trees, we may consider the finite classes of Yule Trees or labelled trees. Formally, Theorem 1 provides a surjective mapping between Coalescent Trees of size n and the set $\mathcal{L}_n$:

Definition 10. The function

$$E : \mathcal{G}_n \to \mathcal{L}_n$$

associating the labelled tree $L \in \mathcal{L}_n$ to a Coalescent Tree $G \in \mathcal{G}_n$ obtained by removing branch lengths, labelling internal nodes by integers $1, \ldots, n-1$ such that the time ordering of coalescent events is respected and dividing branches into branch segments in the sense of Chapter 2 in G is called the *canonical embedding* of $\mathcal{G}_n$ into $\mathcal{L}_n$.

If the argument G is generated according to Kingman's Coalescent, E induces the uniform distribution on $\mathcal{L}_n$. This observation has been made long ago, and used

quite extensively by David Aldous [AP96]. Steel et al. [SM01] exploited this equivalence to develop a number of expectations of tree statistics under the Coalescent. More recently, it seems to have been put to use only rarely, probably because the topology of a Coalescent Tree itself is usually not the object of study. For this work, it is of utmost importance, because many of the constructions in the following chapters would be impossible without it.
We end this paragraph with the following remark: When dealing with Coalescent Trees, Theorem 1 allows us to knock the number of objects to consider down to an integer (although of superexponential growth with n). In principle, the Coalescent can be simulated by generating a random permutation, translating it into a Yule Tree, randomizing the horizontal positions of leaves and endowing branch segments with exponentially distributed lengths.

3.3 Recombination

Recombination is the reciprocal exchange of hereditary material between haplotypes. Often, the term is used synonymously to the genetic crossover that may happen when cells of higher organisms undergo meiosis. In an early phase of meiosis, chromosomes are cut at some points (*double-strand breaks*), which is usually repaired by *DNA* repair mechanisms; however, it may happen that two pieces of formerly different chromosomes are reattached; which is called a *crossover event*. Per double-strand break, the probability of such an event is rather small, but due to the number of double-strand breaks, genetic crossover attains measurable quantities. In human, a popular rule-of-thumb is to expect one crossover per chromosome and meiosis.
One implication of this in theoretical biology is that we cannot rely on the Coalescent alone, because time is not the only mechanism determining relationships between members of a population. Some individual x may be most closely related to individual y on a stretch of loci on the chromosome, but, due to a crossover event way back in the past, it might be more closely related to (that is, coalesce earlier with) individual z when considering the loci past the one at which the crossover event occurred. The ancestral process therefore becomes a process of time and space, with space translating to distance on the chromosome.
It is possible to describe a dual process similar to the coalescent in a Wright-Fisher population of size $2N$ that incorporates crossovers. In such a population, each individual of the new generation is sampled uniformly as before, but undergoes a crossover with another individual with a certain, small probability c. An individual's chromosome can be represented by a finite set of loci, or, alternatively, by the unit interval $[0, 1]$, which reflects the infinite-site assumption that is also made to consider the process of mutation; we will assume the latter in the following. The location of a crossover is chosen somewhere on the chromosome, usually uniformly. It is implied by the construction that at each individual position, the genealogy of a sample is still represented by a coalescent tree. Moving along the chromosome, one may encounter different trees, due to crossover events in the genealogical history of the sample. If a crossover event is found at position $r \in [0, 1]$, the lineage affected by the crossover is split into two, one representing the portion of genetic material from the parent haplotype on the interval $[0, r)$ (without loss of generality, "to the left" of the crossover site) and the other one representing the genetic material on $[r, 1]$ (the "right" parent haplotype). Both of these lineages are subject to random coalescences afterwards; they might even coalesce with each other right away.

The entire ancestral process therefore features coalescing and splitting lineages. Considering $2N \to \infty$, we assume $4Nc \to \rho, \rho > 0$. The parameter ρ is called *recombination rate*. Under this assumption, it is possible to show that the history of a sample is determined by Coalescent events at the rates calculated in section 3.2, and split events at rate $\rho l(t)$, where $l(t)$ denotes the number of lineages at time t. This process is known as the Ancestral Recombination Graph (*ARG*, [Hud83], see Figure 3.2), owing to the graph structure that can be drawn as a visual representation similarly to a coalescent tree. It is attributed to Hudson and Griffiths, who discovered it almost simultaneously.
Because Coalescent rates are quadratic while split rates are linear, the process almost surely enters a state where there is only one lineage left. Looking further in the past, this lineage may split again, so the *ARG* does not feature the absorbing state the Coalescent has. However, one is usually interested in the genealogy up to the first point in time at which there is only one lineage left. The ancestor of the sample discovered at this point is called the *grand MRCA* of the sample.
On any physical range S (called *segment*) between neighbouring positions of crossovers, the position of each crossover event tells us which of the two split lineages is the one "valid" on this segment. Hence, the process restricted to the segment S consists solely of coalescences; and the genealogical history of the sample on S is given by a regular Coalescent Tree G_S. The almost-sure existence of the grand *MRCA* implies that there is a *MRCA* of G_S, which either coincides with the grand *MRCA* or is younger. In the latter case, all valid lineages on S coalesce before the final coalescence that establishes the grand *MRCA*.
Finally, we note that the ancestral process restricted to an interval $[a, b] \subseteq [0, 1]$ is itself an *ARG* with a recombination rate of $\rho' := \rho \cdot (b - a)$. Thus, recombination rate is proportional to physical distance in this model. In reality, it is often reasonable to assume that this is true in at least some approximate sense, even though there is knowledge about specific hot- and coldspots of recombination on chromosomes.
As we know from section 3.2, G_S can be represented by a labelled tree. Moreover, under the infinite-site assumption, there almost surely exists precisely one crossover event at the border between neighbouring segments S, S'. Consequentially, there is at most one crossover in the *ARG* at which one has to follow another lineage on S' than on S. Therefore, the transition between $G_S, G_{S'}$ for two neighbouring segments S, S' can be represented by a Prune-Regraft Operation (see procedure 4 in section 2.6). Knowing this, one may hope that the process defined by observing genealogies along the genome, with transitions between them facilitated by Prune-Regraft Operations, is equivalent to the *ARG*, but this turns out not to be true. Because all ancestral lines may coalesce with each other regardless of whether they result from crossover events and independently of the position of those events, genealogies may be reverted back to non-neighbouring states with a probability greater than zero, meaning one cannot disregard the sequence of previous genealogies while moving along the chromosome. Therefore, the sequence of genealogies along a chromosome generated by the *ARG* is not Markovian, which implies that the above construction does not represent the same process.
However, it has been argued that such a construction represents a reasonably accurate approximation of the *ARG*. This led to the development of the *Sequential Markov Coalescent* (*SMC*) [MC05], which is indeed realized by starting with a random Coalescent Tree and generating successive trees by pruning and regrafting. In the *SMC*, the pruning site is chosen uniformly on the current tree, and the resulting "free" lineage merges with some other lineage of the current tree according to the conditional Coalescent. If one chooses to represent the genealogy by a labelled tree, it can be

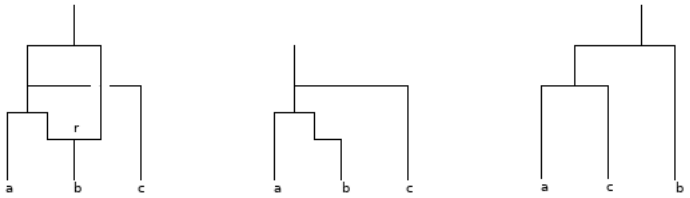


FIGURE 3.2: *Left:* Example of an *ARG* of three individuals and one recombination event in their genealogical history, at some place $r \in (0,1)$ on the chromosome. *Middle, right*: The corresponding coalescent genealogies on chromosome segments $[0,r)$ and $[r,1]$. Note that only in the second case, the *MRCA* coincides with the grand *MRCA*.

shown that the equivalent way of choosing the pruning site is to choose the layer in which to prune proportional to $\frac{1}{k-1}$, where k is the height of the layer, and then choosing the pruned branch segment uniformly from this layer. By Theorem 1, the regrafting branch segment then simply needs to be chosen uniformly. Hence, it is possible to simulate the entire *SMC* by Prune-Regraft Operations on labelled trees.

3.4 The Neutral Theory and methods of statistical genetics

All population models of the previous chapters feature state transitions that arguably rely on a minimal number of assumptions with respect to evolutionary influences, and where the direction of evolution is determined purely by chance, i.e., random genetic drift. In nature, however, it is hard to believe that organisms die, reproduce and evolve completely by chance. One aim of statistical genetics is to identify the mechanisms influencing and driving evolution.
It is quite easy to imagine mathematical population models incorporating some sort of preferential reproduction or death, e.g. due to the specific genotype of an individual. Mechanisms and influences like that are often referred to as *natural selection*, which was one of Darwin's central postulates. Natural selection is the notion that the evolutionary success (i.e. its ability to survive and/or reproduce) of an individual is, in some way, influenced by its phenotype, which is in turn considered to be largely determined by its genetic material.
One way to represent selective mechanisms in population models is to assign a value called *fitness* to an individual, often depending, possibly among other variables, on an individual's genotype, and making the reproductive mechanism work in favour of individuals of high fitness. A situation where the fitness of one genotype is higher than that of the others (often represented by a fitness value of $1+s$ of this genotype compared to 1 for all others) is referred to as *directional selection*. Models incorporating directional selection may also incorporate mutations to the beneficial genotype and back, leading to some equilibrium between genotypes, or, excluding mutation, be used to address questions like how likely it is that the advantageous genotype is either lost by chance or the entire population is replaced by individuals of this type, which is called a *selective sweep*. Often, a distinction is also made between *hard sweeps*, where the fitness advantages and disadvantages are present from the beginning, and *soft sweeps*, where they are established after an initial period governed by

drift, which is supposed to resemble a situation in nature where a sudden ecological change causes different fitnesses of genotypes. A slightly different approach is to consider genotype fitnesses which depend on their frequency. Models incorporating this can, for instance, be used to create a situation called *balancing selection*, where two or more different genotypes keep each other at some equilibrium frequency.
It may seem more appropriate to evaluate data in such contexts, with the aim of explaining the evolutionary history of a population by a selective mechanism. However, the theory becomes a lot more complex, and models of natural selection are usually much harder to analyze mathematically. Furthermore, it turns out that examples are hard to find in which there is clear statistical evidence that a certain model of selection is the best representative of a population's evolutionary background. The variety of existing models also forces competition between theoretical approaches, and it is hard to find statistical grounds upon which one can decide what approach better explains the situation observed in reality.
On the other hand, much of population-genetical data seems to fit the simple, drift-dominated models discussed above surprisingly well. These models aim to represent what is called *neutral evolution*. The fact that they seem, in general, superior to models incorporating more intricate mechanisms than plain uniformity, has lead to a hypothesis called *Neutral Theory*, a term coined by Kimura, stating that almost all evolution in nature is determined by chance [Kim83].
Of course, extreme cases have indeed been observed in nature, where it is likely that a genotype-dependent force is present, and which the neutral theory is not suited well to explain. One very prominent example of this is the abundance of dark phenotypes of peppered moths in England that grew during the 20th century, which was suspected to be caused by increasing levels of pollution [Ket58]; recently, a candidate gene responsible for this color variation was identified [Hof16]. The *LCT* region on chromosome 2 of *H. sapiens* is also often mentioned in that regard, mainly because of drastically reduced variation in the vicinity. In Chapter 5, we will discuss this in more detail. The neutral theory, however, is still very useful as a null hypothesis; for instance, concerning the *LCT* region, first it is necessary to have an expectation of the amount of genetic variation, before one can determine whether variation is reduced. One approach of statistical genetics is to compare data to what would be expected under a neutral model, with the intention of rejecting the neutral theory and providing statistical evidence of non-neutral evolution; the actual description of the mechanism causing non-neutrality is attempted a posteriori.
In order to do this, mathematical properties of various quantities derived under neutrality are utilized. For instance, the expected frequency spectrum (3.4) of mutations in a neutral population is known; a simple test for non-neutral evolution could exploit this by performing a goodness-of-fit test of polymorphism data to this distribution. Over the years, many intricate statistics have been defined based on the neutral frequency spectrum. One approach developed by F. Tajima involves the calculation of the mean number of pairwise differences

$$\hat{\theta}_\pi := \frac{\#\text{pairwise differences}}{\binom{n}{2}}$$

for a sample of size n. It is possible to show that $\hat{\theta}_\pi$ is another maximum-likelihood estimator of θ. The argument of Tajima was that if the population evolved neutrally, the estimators $\hat{\theta}_\pi$ and $\hat{\theta}_W$ would largely agree. The statistic

$$d := \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{Var(\hat{\theta}_\pi - \hat{\theta}_W)}}$$

is called *Tajima's D* [Taj89] and enjoyed much use in the last decades as a test for selection. The usual way of interpreting it is that directional selection, where the population is quickly filled up with descendants of individuals of a beneficial genotype, would cause an excess of low- and high-frequency variants, and therefore a tendency of d to be negative; on the other hand, an evolutionary mechanism resembling balancing selection would cause a tendency of d to be positive. Other frequency-based tests for neutrality have been developed by Fay and Wu, and Fu and Li [FL93; FW00].
In the context of recombination, an important concept is *Linkage Disequilibrium* (*LD*, [Gil98]), which will be the subject of Chapter 5. It is defined as the covariance

$$D := f(ab) - f(a)f(b)$$

of the co-incidence of two alleles a, b at two loci. It is possible to estimate recombination rates with *LD*, but also again nonneutral evolution, e.g. directional selection. In a neutral population, D is supposed to be kept at low levels due to recombination shuffling alleles. Due to the short time an individual with a beneficial allele needs to replace the entire population with its offspring, alleles carried by this individual which are "close" to the beneficial allele will also reach high frequency, because recombination will have less time to uncouple them. As a result, the genetic material around the beneficial allele will become very homogeneous and D will attain comparatively high values. This phenomenon is known as *genetic hitchhiking*. *LD* may also reveal links across distances on the chromosome and epistatic interaction between loci.
With the advent of more advanced sequencing techniques, recently haplotype-based methods of detecting natural selection have become popular, such as *IHS* and *EHH* [Voi+06], which exploit increased levels of haplotype homogeneity, similarly to *LD*.
A little more inspired by a perspective of molecular biology is the *McDonald-Kreitman Test*, which is founded upon the ratio of nonsynonymous to synonymous substitutions, and related statistics [MK91].

Chapter 4

Trees evolving in time

4.1 The Evolving Moran Genealogy

Kingman's Coalescent describes the genealogy of a sample taken from a neutrally evolving population at a single point in time and can be nicely represented in a discrete way by a Yule Tree. What we will investigate in the following, is how a genealogy changes with time in such a model. Most importantly, it turns out that similar methods to the already described apply to such "evolving" genealogies; for instance, we still recover the Yule Process as the mechanism generating the genealogies and providing the probability distribution on them. Our approach is different to e.g. that of [PWW09] in that we assume a strictly finite population size; later on, we carefully examine the behaviour of certain quantities with respect to large populations.
The model we are going to consider is a Moran Model of some size n. The obvious advantage this model has over the Wright-Fisher Model with respect to genealogical properties is that the transitions in the Moran Model, i.e., one parent individual duplicating, and its offspring replacing the killed individual, can be interpreted as binary splits similarly to an iteration of the Yule Process. Such binary splits we would also like to see in a genealogy, as Coalescent Trees are also binary. But in the Wright-Fisher Model, one individual may give rise to several ones in the next generation, which may be impossible to represent by a binary split; Kingman's Coalescent only arises in the infinite-population limit of this model.
Recall that the Moran Process denotes a sequence $M := (P_i)_{i\in\mathbb{N}}$ of populations, where in each step, procedure 6 is applied. We already established in Chapter 3 that with probability 1, there is a finite time i^{MRCA} at which M will enter a state in which the population consists only of the copies (descendants) of some $x_k \in \mathcal{P}_0$, while all other $x_l \in \mathcal{P}_0, l \neq k$, and their copies have been removed from the population. The individual x_k, or one of its descendants, is thus the *MRCA* dating back to at most time 0, and looking backwards in time, there exists a branching pattern describing how the current population of x_k-copies has been created.
We assume in the following that in each time step, the copy of the duplicated individual is placed at its side (instead of replacing the killed individual), and other individuals are shifted to the left or right depending on whether $l < k$ or $l > k$. This is achieved by using the following procedure 7 to generate the population of the next time step:

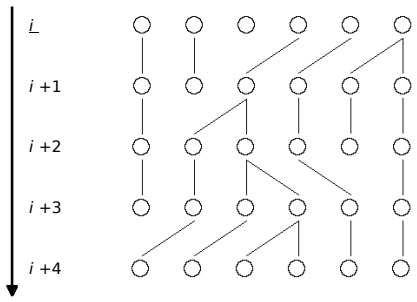


FIGURE 4.1: Iterations of a Moran Process, where the order is maintained according to 7.

PROCEDURE 7: Moran with planar order maintenance in M

Input: Population P of size N
1: Choose one element x_k of P uniformly and create a copy $x'_k = x_k$
2: Choose one element x_l of P uniformly
3: **if** $l < k$ **then**
4: Lower the position of individuals $x_{l+1}, \ldots, x_{k-1}$ by one;
5: Assign the possible positions $k-1, k$ to individual x_k and its copy with probability $1/2$;
6: $P' \leftarrow \{x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{k-1}, x_k \mathrm{v} x'_k, x_k \mathrm{v} x'_k, x_{k+1}, \ldots, x_n\}$
7: **else if** $l > k$ **then**
8: Increase the position of individuals $x_{k+1}, \ldots, x_{l-1}$ by one;
9: Assign the possible positions $k, k+1$ to individual x_k and its copy with probability $1/2$;
10: $P' \leftarrow \{x_1, \ldots, x_k \mathrm{v} x'_k, x_k \mathrm{v} x'_k, x_{k+1}, \ldots, x_{l-1}, x_{l+1}, \ldots, x_n\}$
11: **else**
12: Replace x_k by x'_k;
13: $P' \leftarrow \{x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n\}$
14: **end if**
Output: New population P' of size N

It is assumed that original and copy randomly choose one of the two possible neighbouring positions. The purpose of this is to stress the "memorylessness" of the Moran Model, in that after a duplication, both are indistinguishable and it is not possible to tell which one was present in the previous step by looking at the population. Figure 4.1 illustrates some Moran steps with planar order maintenance.

Note that procedure 7 provides just a minor additional feature and the process M with procedure 7 replacing procedure 6 is still the usual Moran Process. With this in mind, we proceed to give a description of the branching pattern from the *MRCA* to the present generation.
Since the time i^{MRCA} is almost surely finite, we consider a slightly modified definition of the lineage

$$Y(x_k) := (y_i(x_k))_{0 \le i \le i^{MRCA}}$$

of an individual $x_k \in P_{i^{MRCA}}$ as the *finite* sequence of its ancestors. The set $L =$

$\{Y(x_k), k = 1, \dots, n\}$ denotes all the lineages of individuals in the population $P_{i^{MRCA}}$. The time i^* denotes the time at which all lineages $Y(x_k) \in L$ become congruent (i.e., the *MRCA* is established backward in time), and the sequence of times $i_1 = i^* + 1, \dots, i_{n-1} = i^{MRCA}$ collects all the first time steps at which there are exactly $2, \dots, n$ lineages that are pairwise noncongruent. The times i_l are the ones in which a set C of congruent lineages splits into two sets of congruent lineages C_1, C_2; the times $i_l - 1$ would be respectively those at which sets of pairwise noncongruent lineages "coalesce". The sets of pairwise congruent lineages at time i_l define a partition on L into l sets.
Now, at some time i_l we let all sets of pairwise congruent lineages be represented by a leaf ι. Additionally, at time 0, all lineages in L are neccessarily congruent and are represented by a single leaf ι. We assume that at the times i_l, the number of leaves is increased by turning the leaf representing the splitting set C into a node ν labelled by the integer $l - 1$, and two leaves representing C_1, C_2 are appended below ν. Planary order maintenance in M implies that the indices of the individuals whose lineages are contained in C_1, C_2 form two successive blocks $l_1, \dots, l_2, r_1, \dots, r_2$. We assume that the leaf representing C_1 is placed to the left and the one representing C_2 to the right.
Thus, in the end the genealogical history is represented by a Yule Tree, as the instructions given above are equivalent to the construction of a Yule Tree given in Chapter 1. Furthermore, for $l = 1, \dots, n-1$, each leaf $\iota^s, s = 1, \dots, l$ that represents a set of lineages in time step i_l corresponds to a sequence of individuals

$$X(\iota^s) = \left(x(\iota^s)_k^j\right)_{i_l \leq j < i_{l+1}}$$

which are ancestral to some portion of the population at time i^{MRCA}. For all $i_l \leq j < i_{l+1} - 1$, it holds that either $x(\iota^s)_k^j \equiv x(\iota^s)_k^{j+1}$ or $x(\iota^s)_k^{j+1}$ is a duplicate of $x(\iota^s)_k^j$. In time step i_{l+1}, since the number of leaves is increased by one, one of the l different $x(\iota^s)_k^{i_{l+1}-1}$, $s = 1, \dots, l$ is necessarily subject to a duplication in the associated iteration of M. Since M is neutral, each of the $x(\iota^s)_k^{i_{l+1}-1}$ has the same chance of being selected for this, and therefore each of the ι^s has the same chance of being turned into an internal node. The genealogy of the population $P_{i^{MRCA}}$ is therefore given by a Yule Tree $T_{i^{MRCA}}$ resulting from the discrete Yule Process (Procedure 1).
It is worth mentioning that the observation that the genealogy of a planarily ordered Moran Process can be represented by a Yule tree is essentially a reformulation of a result from [AP96]. In this work, the same is shown for labelled trees, and since we know that labelled trees are equivalent with respect to topology to trees generated under the Yule Process, it should not come as a surprise that this construction is possible.
We assume in what follows that M already starts in a state where the *MRCA* is already established, such that at $i = 0$ there already exists a genealogy $T_0 \in \mathcal{T}_n$. M can then be emulated by observing the genealogy $T_i, i \geq 0$ directly, where T_i is modified according to procedure 8.

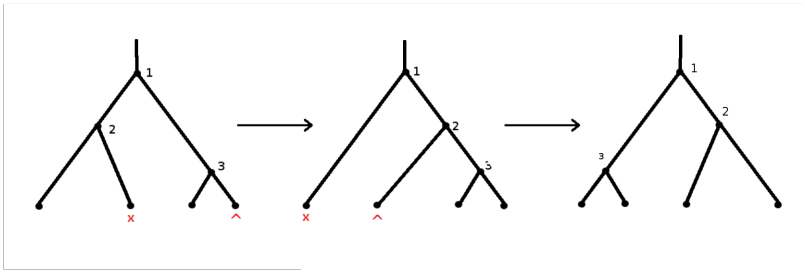


FIGURE 4.2: Two steps in an *EMG* of size 4. Each step, one individual is killed ("X") and one duplicated ("Λ").

PROCEDURE 8: Evolving Moran Genealogy given T_i

if $k = l$ **then**
    $T_{i+1} \leftarrow T_i$
else
    Remove the leaf representing the killed individual x_l alongside the branch connecting it to the remainder of T_i, and the internal node v_j at the position in T_i that branch is attached at;
    Merge the two branch segments b, b' connected to v_j into one;
    Merge all pairs of branch segments b, b' in layers $j, j+1$ belonging to the same branch into single branch segments;
                                ▷ 4-6 "remove" layer $j+1$
    Decrease the labels of $v_{j'}, j' > j$ by one in T_i;
    Turn the leaf representing the duplicated individual x_k in T_i into an internal node with two new leaves appended to it and label it by $n-1$; $T_{i+1} \rightarrow T_i$;
end if
return T_{i+1}

This procedure modifies a Yule Tree by removing a random leaf and the branch connecting this leaf to the tree, and by introducing a new "split" at the bottom similarly to procedure 1, symbolizing the removal of one and the duplication of another individual. Let $\Phi_{k,l}(T_i)$ denote the output of procedure 8 given k, l. Then, we define:

Definition 11.

- The process $(\mathcal{T}_i)_{i\in\mathbb{N}}$ with $T_{i+1} = \Phi_{k,l}(T_i)$ for uniform k, l and uniformly chosen $T_0 \in \mathcal{T}_n$ is called *Evolving Moran Genealogy*, for short *EMG*.
- For $T, T' \in \mathcal{T}_n$, we define the notation

$$T \rightharpoonup T' \Leftrightarrow \exists k, l \in \{1, \dots, n\} : \Phi_{k,l}(T) = T'$$

- We will identify a leaf ι of any Yule Tree $T_i, i \geq 0$ arising under the *EMG* with the individual $x \in P_i$ of the associated Moran Process represented by ι, and write $x \in T_i$ if an individual x is represented in T_i.

The *EMG* (see Figure 4.2) is a Markov Chain on the set $\mathcal{T}_n$ of n-sized Yule Trees, where, only considering the duplications and removals of leaves, one recovers the Moran Model. It has a transition matrix E with nonzero diagonal entries, since $l = k$ entails $T_{i+1} = T_i$. It is recurrent, because at most $n-1$ transition steps of the above

form are needed to transform T_i into some arbitrary $T' \in \mathcal{T}_n$ (to see this, recall that any Yule Tree of size n can be generated in $n-1$ steps). Since $T_{i+1} = T_i$ may occur arbitrarily many successive times, aperiodicity of the *EMG* follows as well.
As a consequence, there exists a stationary distribution P^* of the *EMG* on $\mathcal{T}_n$. It follows that P^* is the uniform distribution, i.e. $P^*(T) = \frac{1}{(n-1)!}$ for all $T \in \mathcal{T}_n$, because a genealogy T_i observed at some large time i is stochastically the same object as T_0, and we can consider it as a tree generated by procedure 1. A similar statement was derived and discussed in [AP96] and [Ald00].
The reason to draw T_0 uniformly from $\mathcal{T}_n$ is that T_0 can be thought of as generated by procedure 1. In mathematical terms, the uniform distribution represents a natural *entry law* of the *EMG*. Because the uniform distribution is also the stationary distribution, the *EMG* is a stationary process by definition.
The relation $T \rightharpoonup T'$, indicating that T can be transformed into T' by some duplication/remove combination in M, can be used to give a formal description of the entries of the transition matrix E of the *EMG*. Importantly,

$$T = T' \Rightarrow T \rightharpoonup T';\ T \rightharpoonup T' \not\Rightarrow T' \rightharpoonup T$$

Then the T, T'-entry of E can be denoted in the following way:

$$\Pr(T_{i+1} = T' | T_i = T) = \begin{cases} 0 & T \not\rightharpoonup T' \\ \frac{|\{(k,l) \in \{1,\dots,n\}^2 : \Phi_{k,l}(T) = T'\}|}{n^2} & \text{otherwise} \end{cases}$$

In particular, the diagonal entries of E depend on T and do not have to be equal. For example, suppose $n = 2^k$ for some $k \geq 2$, and consider the *caterpillar* $C \in \mathcal{T}_n$ obtained under the Yule process by always choosing the leftmost leaf split, and a *complete binary search tree* $B \in \mathcal{T}_n$, which is a tree characterized by the fact that there is an equal number of leaves on both subtrees below each internal node. Then $\Pr(T_{i+1} = C | T_i = C) = \frac{2n}{n^2}$, whereas $\Pr(T_{i+1} = B | T_i = B) = \frac{n+2}{n^2}$.

Application: Tree Balance Since a Yule tree T in the *EMG* is plane and individuals ordered from left to right, we may consider the left and right subtrees T^l, T^r below the root node ν_1. Essentially, T^l can be thought of as the induced subtree of all leaves on the left side below ν_1 (the same holds for T^r). Suppose we are interested in the dynamics of the number of leafs on the left, i.e. $|T^l|$.

Definition 12. The process $TB := (|T_i^l|)_{i \in \mathbb{N}}$ is called Tree Balance Process of the Evolving Moran genealogy.

The choice between observing left and right subtree size is arbitrary, since always $|T^r| = n - |T^l|$. A closely related quantity is the Ω_1-statistic [**feretti:srecevents**; Ald01], where $\Omega_1(T_i) := \min(|T_i^l|, |T_i^r|)$, and one observes $(\Omega_1)_{i \in \mathbb{N}}$. In some applications, it is more convenient to consider the minimum of both sides instead of insisting on the notion of left and right; for instance, the Ω_1-process is still well-defined if considered on labelled trees. However, there is little difference between *TB* and the Ω_1-process, as paths of *TB* are essentially mirrored at $\lfloor \frac{n}{2} \rfloor$ when considering Ω_1. Determining the dynamics of *TB* thus suffices to also obtain those of Ω_1.
By Proposition 2, tree balance is uniform under the Yule Process. This allows us to calculate:

Proposition 3. *The transition probabilities of TB are as follows:*

If $2 \leq |T_i^l| \leq n-2$,

$$\Pr(|T_{i+1}^l| = \omega \mid |T_i^l|) = \begin{cases} \frac{|T_i^l|(n-|T_i^l|)}{n^2} & \omega = |T_i^l| + 1 \\ \frac{|T_i^l|^2+(n-|T_i^l|)^2}{n^2} & \omega = |T_i^l| \\ \frac{|T_i^l|(n-|T_i^l|)}{n^2} & \omega = |T_i^l| - 1 \end{cases}$$

If $|T_i^l| = 1$,

$$\Pr(|T_{i+1}^l| = \omega \mid |T_i^l| = 1) = \begin{cases} \frac{1}{n} & \omega = 2 \\ \frac{(n-1)^2+2}{n^2} & \omega = 1 \\ \frac{1}{n^2} & \text{otherwise} \end{cases}$$

And if $|T_i^l| = n-1$,

$$\Pr(|T_{i+1}^l| = \omega \mid |T_i^l| = n-1) = \begin{cases} \frac{1}{n} & \omega = n-2 \\ \frac{(n-1)^2+2}{n^2} & \omega = n-1 \\ \frac{1}{n^2} & \text{otherwise} \end{cases}$$

Proof. Suppose $2 \leq |T_i^l| \leq n-2$. $|T_{i+1}^l| = |T_i^l| - 1$ is the case if one individual on the left side is removed and one on the right is duplicated. This happens with probability $\frac{|T_i^l|(n-|T_i^l|)}{n^2}$. We obtain the same probability for the case $|T_{i+1}^l| = |T_i^l| + 1$.
Finally, we have $|T_{i+1}^l| = |T_i^l|$ if removal and duplication take place on the same side. The probability of this is $\frac{|T_i^l|^2+(n-|T_i^l|)^2}{n^2}$.
The only difference in the cases $|T_i^l| = 1, n-1$ is that one has to include the possibility of a complete removal of T_i^l in the first and T_i^r in the latter case. If this happens, $|T_{i+1}^l|$ and $|T_{i+1}^r|$ are independent of $|T_i^l|$ and $|T_i^r|$. In fact, if the left side of T_i is completely removed in the transition to time $i+1$, T_{i+1} can be interpreted as a random Yule Tree. By Proposition 2, $|T_{i+1}^l|$ thus assumes any value $1, \ldots, n-1$ with uniform probability.
Therefore, considering $|T_i^l| = 1$, the total probability $\Pr(|T_{i+1}^l| = 2 \mid |T_i^l| = 1)$ is the sum of the probability that the individual on the left is duplicated and one on the right is removed, which amounts to $\frac{(n-1)}{n^2}$, and the probability that it is removed and $|T_{i+1}^l| = 2$ by chance, which happens with probability $\frac{n-1}{n^2}\frac{1}{n-1} = \frac{1}{n^2}$. This sum equals $\frac{1}{n}$.
For $\Pr(|T_{i+1}^l| = 1 \mid |T_i^l| = 1)$ and $\Pr(|T_{i+1}^l| = \omega > 2 \mid |T_i^l| = 1)$ the calculation is similar, and of course the case $|T_i^l| = n-1$ can be treated analogously. □

We may refer to the phases between complete removals of T_i^l or T_i^r as *episodes* of the process *TB*. We notice that the transition probabilities within an episode are identical to those of the allele frequency $f_P(a)$ in a two-allele Moran Model (see section 3.2). Therefore, in the large-population limit, tree balance can be thought of as a Wright-Fisher Diffusion (see equation 3.1). Also, if n is large and $\frac{|T_i^l|}{n}$ is either close to 0 or 1 (the genealogy is "unbalanced"), the strength of diffusion is weakest. Consequently, if the Evolving Moran Genealogy enters an unbalanced state, genealogies in the following generations are expected to remain rather unbalanced, whereas balanced trees entail a higher volatility of *TB* in next steps. Interestingly, a similar behaviour can be observed in trees along recombining chromosomes (see Chapter 6 and [**feretti:srecevents**]).
At last, we note that the complete removals of T_i^l or T_i^r, i.e. starting and ending

times of episodes in *TB*, are precisely the times of *MRCA* jumps in the *EMG*, to be discussed in section 4.3.

4.2 Going backwards in time

We already established by taking a close look on procedure 8, that the *EMG* features a step equivalent to the appending of leaves in the Yule Process, which represents the birth of a new individual. To motivate the constructions in this section, we recall that Yule Trees are equivalently described by successive random graftings (procedure 2). Thus, one might ask whether a process on $\mathcal{T}_n$ exists where the birth of an individual not contained in the previous population is represented by a grafting event on the current tree. Such a process, it turns out, exists, and can be interpreted in a very intuitive way.
Let $T \in \mathcal{T}_n$ denote a random Yule Tree of size n. On T, we perform a *Merge-Regraft* operation:

PROCEDURE 9: Merge-Regraft on given T

1: Choose one branch segment b of T from the set $\{b_1, \dots, b_{\frac{n(n+1)}{2}}\}$ with probability
$$\Pr(b = b_k) = \begin{cases} \frac{1}{n^2} & b_k \text{ ends in a leaf} \\ \frac{2}{n^2} & \text{otherwise} \end{cases}$$
and χ from $\{left, right\}$ with equal probability;
2: **if** b ends in a leaf **then**
3: $T' \leftarrow T$
4: **else**
5: Remove the n-th layer of T; remove v_{n-1}; place leaves at the tips of the branch segments that extend across layer $n-1$;
6: ▷ the position of v_{n-1} is then occupied by some leaf
7: Regraft a new leaf at branch b with orientation χ in T according to Procedure 2 (skipping step 1);
8: $T' \leftarrow T$
9: **end if**
10: **return** T'

One may imagine that in step 5 all leafs are moved up by one layer, such that two of them must "merge". Let the result of this operation be denoted by $\Phi'_{b,\chi}(T)$. $\Phi'_{b,\chi}(T)$ is itself an object of the set $\mathcal{T}_n$. The function Φ' can be thought of as a combinatorial inversion of Φ (see Section 4.1): The split event facilitated by Φ is revoked by Φ', and the leaf that is removed under Φ can be recovered ("revived") by Φ' by regrafting; in fact, we have $T \rightharpoonup T' \Leftrightarrow \exists (b, \chi) : \Phi'_{b,\chi}(T') = T$.
We consider the process
$$R := (\tilde{T}_i)_{i \in \mathbb{N}},$$
where $\tilde{T}_0$ is uniformly chosen and, given $\tilde{T}_i$, $\tilde{T}_{i+1}$ is generated by the mechanism described above, i.e. $\tilde{T}_{i+1} = \Phi'_{b,\chi}(\tilde{T}_i)$ for some random choice of b and χ (See Figure 4.3).
In what follows, we will show that the notion that R represents a sequence of reversed *EMG*-steps is justified mathematically, and that this process indeed represents a *time-reversal* of the *EMG*, where the term time-reversal is used in the sense of

e.g. [LW98]:

Definition 13 (Time Reversal of a Markov Chain)**.** Given a Markov Chain $(S_i)_{i\in\mathbb{N}_0}$ on a state set $\{1,\dots,m\}$ with transition probabilities $p_{j,k}, j,k \in \{1,\dots,m\}$ and stationary distribution $\pi = (\pi_1,\dots,\pi_m)$, the *reverse chain* is defined as the Markov Chain $(\hat{S}_i)_{i\in\mathbb{N}_0}$ on the state set $\{1,\dots,m\}$ with transition probabilities

$$\hat{p}_{jk} = \frac{\pi_k p_{kj}}{\pi_j}$$

The reason to define a reversed Markov Chain this way is that if the original chain runs for a very long time, i.e., i is large, it is possible to revert time by exchanging states and weighting according to the stationary distribution. This is expressed by the following expression:

$$\Pr(S_i = j | S_{i+1} = k) = \frac{\Pr(S_{i+1} = k | S_i = j)\Pr(S_{i+1} = k)}{\Pr(S_i = j)} \approx \frac{p_{kj}\pi_k}{\pi_j} = \hat{p}_{jk}$$

In the limit, i.e., assuming that S starts at $-\infty$ instead of 0, this appoximation becomes an equality. In particular, $\forall j : \sum_{k=1}^{m} \hat{p}_{jk} = 1$, and the stationary distribution of the reversed chain $\hat{S}$ is given by π.

Lemma 4. *For all $T, T' \in \mathcal{T}_n$:*

$$\Pr_{EMG}(T_{j+1} = T' | T_j = T) = \Pr_R(\tilde{T}_{i+1} = T | \tilde{T}_i = T') \tag{4.1}$$

with $\Pr_{EMG}$ denoting the transition probability of the EMG-process, and $\Pr_R$ that of the process R.

Proof. Recall that Φ was dependent on the choice of k, l, which were both chosen uniformly. The probability of $k = l$ in one step of the *EMG*, which for all $T \in \mathbb{T}_n$ entails $\Phi_{k,k}(T) = T$, is $\frac{1}{n}$. The probability that the branch segment b chosen in a transition of the process R is inside layer n, which always leads to $\Phi'_{b,\chi}(\tilde{T}) = \tilde{T}$, is also $\frac{1}{n}$ in total for any T.
We define for arbitrary $T, T' \in \mathcal{T}_n, T \rightharpoonup T'$:

$$S_1 := \{(k,l) \in \{1,\dots,n\}^2 : \Phi_{k,l}(T) = T', k \neq l\}$$
$$S_2 := \{(b,\chi) \in \{b_1,\dots,b_{\frac{n(n-1)}{2}}\} \times \{\textit{left},\textit{right}\} : \Phi'_{b,\chi}(T') = T\}$$

If we can show that $|S_1| = |S_2|$, we are done. Let $(k,l) \in S_1$. Let ν denote the internal node deleted by $\Phi_{k,l}(T)$. Choosing the regrafting site b as the branch segment generated by merging the two segments connected to ν (compare step 5 in procedure 8), and χ according to whether the branch of x_l extends to the left or right in T, we obtain a unique $(b,\chi) \in S_2$, which yields a mapping $\mu : S_1 \to S_2$. Since by definition of the Yule process there cannot be two or more tuples k, l and k', l with $k \neq k'$ such that $\Phi_{k,l}(T) = \Phi_{k',l}(T)$, μ is injective.
On the other hand, for any $(b,\chi) \in S_2$ such that $\Phi'_{b,\chi}(T') = T$, choosing l such that x_l is the leaf regrafted in T' by $\Phi'_{b,\chi}(T')$ and k such that x_k is the leaf replacing the highest-labeled internal node in T' by $\Phi'_{b,\chi}(T')$ (see step 5 of procedure 9), we obtain $(k,l) \in S_1$ such that $\mu((k,l)) = (b,\chi)$. Therefore, μ is a bijection and both sets are equally large. □

Corollary 2. *The process R represents the time-reversed process of the EMG.*

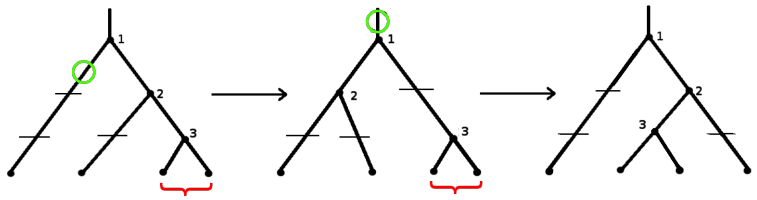


FIGURE 4.3: Two possible transitions of the process R, i.e. the $EMG^\flat$ of size 4. Branch segment chosen for regrafting marked by "o".

Proof. The existence of a time-reversed process $R(EMG)$ on $\mathcal{T}_n$ is provided by the fact that it is a recurrent Markov Chain with nonzero stationary distribution. The transition probabilities of this process are

$$\begin{aligned}\Pr_{R(EMG)}(T_{j+1} = T' | T_j = T) &= \Pr_{EMG}(T_i = T' | T_{i+1} = T) \\ &= \Pr_{EMG}(T_{i+1} = T | T_i = T') \frac{\Pr_{EMG}(T_i = T')}{\Pr_{EMG}(T_{i+1} = T)}\end{aligned}$$

Since the stationary distribution of the EMG is the uniform distribution, we have $\Pr_{EMG}(T_i = T') = \Pr_{EMG}(T_{i+1} = T)$. Therefore,

$$\Pr_{R(EMG)}(T_{j+1} = T' | T_j = T) = \Pr_{EMG}(T_{i+1} = T | T_i = T') \tag{4.2}$$

and these transition probabilities are exactly the ones provided by the process R (compare equation (4.1)). □

Definition 14. We call the process R the *Evolving Moran Genealogy backwards in time*, for short $EMG^\flat$.

We end this section with the remark that the $EMG^\flat$ resembles the *Aldous Chain* on *cladograms* [Ald00]. However, the two processes act on different state spaces. A cladogram is a tree with labelled leaves and unlabelled internal nodes, of which there exist $\Pi_{m=1}^{n}(2m-1)$ of size n. In the Aldous Chain, a cladogram is modified by cutting off a leaf and reattaching it somewhere in the cladogram, an operation known as the *Aldous Move*, which is similar to the regrafting operation that is part of the $EMG^\flat$. Letting $n \to \infty$, one obtains a Markov Process on *continuum random trees* as the limiting process [LMW18].

4.3 The *MRCA* Process

Besides the technical aspects, there are some reasons why the $EMG^\flat$ as a stochastic process can prove useful in theoretical and practical regard. While the transitions in the EMG rely on two random mechanics (duplication and removal), in the $EMG^\flat$ they are unified within the regrafting operation. Because of that, aspects about the genealogy itself may become more tractable to analytic investigation. One good example of this is the *MRCA*-Process.
Let x^* denote the *MRCA* of a genealogy generated by a neutral Moran Process. With probability 1, after some finite time a descendant of x^* will become ancestral to the entire population, establishing a new *MRCA*. In the EMG, this corresponds to cases

where the left or right side of the current genealogy T_i consists of only one individual, and this individual happens to be removed without replacement. Taking a look at procedure 8, this means that the topmost internal node ν_1 of T is removed, and its place is taken by the node ν_2. The establishment of a new *MRCA* is therefore represented in the *EMG* by the eventual obliteration and repositioning of the root node; hence, the *MRCA* "jumps". Because of that, we will synonymously call a *MRCA* jump a "root jump" in the following. Also, a root jump necessarily coincides with the end of an episode in the tree balance process (see Section 4.1).
Defining $\chi_{MRCA}(i) = 1$ if at time i a new *MRCA* of the population is established, and $\chi_{MRCA}(i) = 0$ otherwise, we call $(\chi_{MRCA}(i))_{i\geq 0}$ the *MRCA*-Process.

Lemma 5. *$(\chi_{MRCA}(i))_{i\geq 0}$ is a geometric jump process of intensity $\frac{2}{n^2}$.*

Proof. In the $EMG^\flat$, the root of the genealogy T_i changes if and only if the imaginary branch b_1 is chosen for regrafting. This happens with probability $\frac{2}{n^2}$ in each step (see also Figure 4.4). □

In [PW06], the *MRCA*-Process in the infinite-population limit is identified as a Poisson-Process of intensity 1. There, this is achieved making use of the *Lookdown-Construction* ([DK96]) of the Moran Process, another tree-like construction that has proven useful for the investigation of genealogical traits of a population in the past. Lemma 5 agrees with this result, because the limit of the geometric jump process as $n \to \infty$, with time rescaled in units of $\frac{n^2}{2}$, is also a Poisson process of intensity 1.
By Lemma 5, we also easily conclude that the number of steps needed to observe any number $r \in \mathbb{N}$ of root jumps follows a negative binomial distribution $NB(r, \frac{2}{n^2})$.
However, exploiting the structure of the $EMG^\flat$, we even get to investigate the properties of the *MRCA*-Process during specific intervals, such as ongoing fixations in the underlying Moran Process.

Definition 15. Suppose a member $\tilde{x}$ in a neutral Moran Process was generated as the result of some duplication at time $i^* > 0$.

- A (neutral) *fixation* refers to the case of $\tilde{x}$ becoming ancestral to the population at some time $i^{fix} > i^*$, i.e., in the transition from $i^{fix} - 1$ to i^{fix}, the last remaining individual $\tilde{x}$ is not ancestral of is removed without replacement.
- i^{fix} is called *fixation time* of the individual $\tilde{x}$
- i^* is called *birth time* of $\tilde{x}$

If the descendants of $\tilde{x}$ "fix" in the population, it follows that $\tilde{x}$ is a common ancestor of the population. However, it should be noted that it is not necessarily the most recent one.

Lemma 6. *In a Moran Population of size $n \geq 2$, we expect to observe $2 - \frac{2}{n}$ MRCA-jumps between (and including) i^* and i^{fix}.*

Proof. By our assumptions, we know that one *MRCA*-jump necessarily happens at the transition of $T_{i^{fix}-1}$ to $T_{i^{fix}}$. We claim that we expect another one during the remainder of the fixation time.
Let $l = i^{fix} - 1 - i^*$. We know $l \geq n - 2$, since the minimal number of steps necessary to fix the descendants of $\tilde{x}$ is $n - 1$. The sequence of genealogies $(T_{i^*}, \dots, T_{i^{fix}-1})$ in reverse order is a path $y = (T'_0, \dots, T'_l)$ of the $EMG^\flat$, where $T'_0 = T_{i^{fix}-1}, \dots, T'_l = T_{i^*}$.

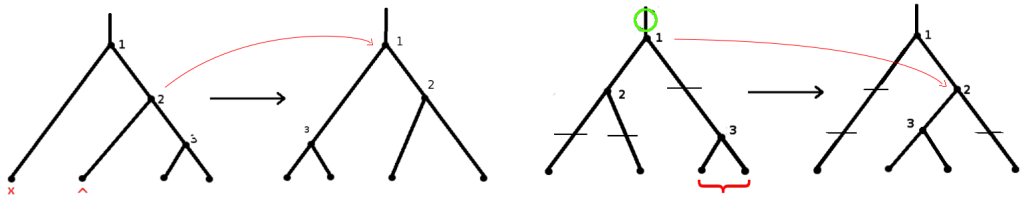


FIGURE 4.4: *MRCA* jumps in the *EMG* and $EMG^\flat$ (See also Figures 4.2, 4.3).

The set of $EMG^\flat$-time steps $\{1,\dots,l\}$ contains a subset $I = \{i_1,\dots,i_{n-1}\}$, $i_l \leq i_{l+1}$, where $i \in I$ if and only if $x \in T'_l$ holds for the individual x regrafted at time i; i.e., x is also present in the population at the time l, which represents the birth time of $\tilde{x}$ in the $EMG^\flat$. I thus consists of exactly the times where individuals of the population at the birth time of $\tilde{x}$ are revived. In particular, $i_1 = 0$ and $i_{n-1} = l$. For $i \in I$, let $S_i := \{x \in T'_i : x \in T'_l\}$ denote the set of individuals that will be members of the population at the birth time of $\tilde{x}$.
Starting from T'_1, a root jump can only occur in some step i if $i \in I$. For any $i_j \in I$, we know that regrafting must take place in some layer $k \leq j+1$. We therefore consider the sequence

$$\hat{T}^{(1)} = (T'_l)_{S_{i_1}}, \dots, \hat{T}^{(n-2)} = (T'_l)_{S_{i_{n-2}}}, \hat{T}^{(n-1)} = T'_l$$

of S_i-induced subtrees of T'_l for $i \in I$. Since T'_l is a random Yule tree, by Lemma 2 we may assume that each $\hat{T}^{(j)}$, $j = 2,\dots,n-1$ is obtained from a random grafting operation 2 performed on $\hat{T}^{(j-1)}$. The probability of a root jump in step i_j is therefore the probability of regrafting at the imaginary branch of $\hat{T}^{(j)}$, which equals $\frac{2}{j(j+1)}$.
The total expected number of root jumps along the $EMG^\flat$-path y is then

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

This expression equals $1 - \frac{2}{n}$. Adding the root jump that necessarily occurs in step 1, we end up with an expectation of $2 - \frac{2}{n}$.

□

Considering the infinite-population limit, we conclude that between birth and fixation time of an individual, there are 2 expected *MRCA* jumps in total. There are other ways of showing this; for instance, one may calculate the expected number of root jumps conditioned on the length l of the fixation period, and arrive at the expectation of 2 for large n by approximating the distribution of l. However, while the above approach is seemingly less refined and a litte more spadework is necessary, it enables us to obtain a closed formula for *all* n.

The distribution of root jump frequency In a Moran Process with population size n, the number of root jumps during a neutral fixation period is at least 1 and smaller than $n-1$. Therefore, there exists a probability distribution of that number for given n, which may be calculated along similar lines as the expectation. This distribution converges as $n \to \infty$, yielding a root jump distribution in the infinite-population

case.
For $n \geq 2$, let $\mathrm{Pr}_n(k)$ denote the probability of observing k root jumps during a neutral fixation in an *EMG* of size n, and $\mathrm{Pr}_\infty(k)$ the same probability in the infinite-population limit. $\mathrm{Pr}_n(k)$ can be written as follows:

$$\mathrm{Pr}_n(k) := \sum_{2 \leq i_1, \ldots, i_{k-1} \leq n-1} \Pi_1^k \frac{2}{i_k(i_k+1)} \Pi_{j \neq i_1, \ldots, i_{k-1}} \left(1 - \frac{2}{j(j+1)}\right)$$

This is obtained by multiplying the probabilities of regrafting at the imaginary branches of $\hat{T}^{(i_1)}, \ldots, \hat{T}^{(i_k-1)}$ (in the sense of the notation used in Lemma 6) and not regrafting at the imaginary branches of all other $\hat{T}^{(j)}$, summed up over all possible choices of $i_1, \ldots, i_{k-1}$. For $k = 1$, in which case the imaginary branch is never chosen for regrafting, we have simply

$$\mathrm{Pr}_n(1) = \Pi_{j=2}^{n-1} \left(1 - \frac{2}{j(j+1)}\right)$$

and this expression can be simplified to $\frac{n+1}{3(n-1)}$. By reordering the factors, we obtain the following expressions for $k = 2, 3, \ldots$:

$$\begin{aligned}
\mathrm{Pr}_n(2) &= \Pi_{j=2}^{n-1} \left(1 - \frac{2}{j(j+1)}\right) \left[\sum_{k=2}^{n-1} \frac{2}{k(k+1)-2}\right] \\
\mathrm{Pr}_n(3) &= \Pi_{j=2}^{n-1} \left(1 - \frac{2}{j(j+1)}\right) \left[\sum_{k=2}^{n-2} \frac{2}{k(k+1)-2} \cdot \left(\sum_{l=k+1}^{n-1} \frac{2}{l(l+1)-2}\right)\right] \\
&\ldots
\end{aligned} \tag{4.3}$$

For $n = 2$, $\mathrm{Pr}_n(1) = 1$, and as $n \to \infty$, $\mathrm{Pr}_n(1) = \frac{n+1}{3(n-1)} \to \frac{1}{3} = \mathrm{Pr}_\infty(1)$. In particular, $\mathrm{Pr}_n(1)$ decreases monotonously with n. Note that this can be interpreted as an analogon to a result in [PW06] about the infinite-population limit. In the terminology of this work, the value $\frac{1}{3}$ corresponds to the probability that the "next fixation curve has not yet started" at the time i^*; a fixation curve begins in the Lookdown-Construction at the birth time of the individual which will become the *MRCA* in the future, and climbs its way up from the bottom "immortal line" to infinity.
The other probabilities in the infinite-population limit can be calculated numerically by evaluating the infinite-sum expressions on the right-hand sides of (4.3). By continuity, the probabilities $\sum_{k=1}^{\infty} \mathrm{Pr}_\infty(k)$ sum up to 1. The largest contribution comes from $\mathrm{Pr}_\infty(2) = \frac{11}{27} = \frac{11}{9} \cdot \frac{1}{3}$. As a side note, since $\mathrm{Pr}_n(2)$ increases monotonously with n, we can calculate that for $n \leq 9$, the distribution is dominated by $\mathrm{Pr}_n(1)$, whereas for $n \geq 10$, the probability $\mathrm{Pr}_n(2)$ provides the largest value. Figure 4.5 outlines some of the distributions for different population sizes.

4.4 The age and lifetime of coalescent events

Another implication of the $EMG^\flat$ is that coalescent events (i.e., internal nodes) are "visible" in the genealogy for a certain average number of steps. We can determine an "age structure" of coalescent events in that we may calculate how long a node with a certain label is expected to remain part of the genealogy.

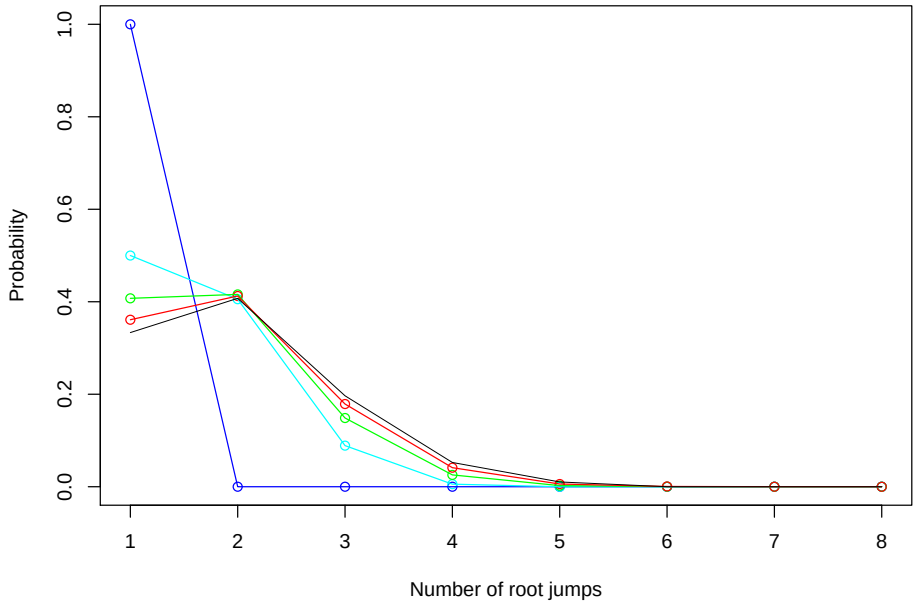


FIGURE 4.5: The distributions of $P_n(k)$, $k = 1, \ldots, 8$, $n = 2$ (blue), 5 (turquoise), 10 (green), 25 (red) and ∞ (black).

In general, the time until the internal node labeled k is moved down by one layer is geometrically distributed with parameter $\frac{k(k+1)}{n^2}$, because $\frac{k(k+1)}{2}$ branches exist above this internal node, each providing a probability mass of $\frac{2}{n^2}$. In the case of the root node, this expectation is $\frac{2}{n^2}$, as stated before.
The expected time l_1 ("life-time") until the current root node of T_i vanishes under the $EMG^\flat$ is therefore

$$\sum_{k=1}^{n-1} \frac{n^2}{k(k+1)} = n^2 \left(1 - \frac{1}{n}\right)$$

In the large-population limit, this corresponds to a rate of 2, known as the "rate of fixations". This can be extended to arbitrary internal nodes: The expectation of the time l_k until the node with label k vanishes in the $EMG^\flat$ is given by

$$\mathbb{E}(l_k) = \sum_{j=k}^{n-1} \frac{n^2}{j(j+1)} = n^2 \left(\frac{1}{k} - \frac{1}{n}\right) \tag{4.4}$$

corresponding to a rate of $\frac{2}{k}$ in the limit. The expected life times of coalescent events therefore form a harmonic series. We also find

$$Var(l_k) = \sum_{j=k}^{n-1} \frac{n^4}{(j(j+1))^2} - \frac{n^2}{j(j+1)}$$

which, surprisingly, does not converge in rescaled time. Note that this holds for individual k; life-times of two or more coalescent events considered at the same time are certainly correlated.

The average over all internal nodes becomes

$$\begin{aligned}\frac{1}{n-1}\mathbb{E}(l_k) &= \frac{1}{n-1}\sum_{k=1}^{n} n^2 \sum_{j=k}^{n-1}\frac{1}{j(j+1)} \\ &= \frac{n^2}{n-1}\sum_{k=1}^{n-1}\left(\frac{1}{k}-\frac{1}{n-1}\right) \\ &= \frac{n^2}{n-1}a_{n-1}-\frac{n^2}{(n-1)^2} \\ &\approx n\log(n)\end{aligned} \tag{4.5}$$

Rescaling time by $\frac{2}{n^2}$, we obtain an average life time of coalescent events of approximately $\frac{2\log(n)}{n}$, which converges to 0 as $n \to \infty$. We conclude that in large populations, most coalescent events only exist for a short time. In the *EMG*, "new" coalescent events are always introduced at the bottom, in contrast to the *EMG*$^\flat$, where they may be generated at arbitrary branch segments in the current genealogy. But since the processes are time-reversals of each other, the above result holds for both. In the context of the *EMG*, one possible interpretation of this is that, given a recent duplication event, all descendants of at least one of the two individuals, original and duplicate, will be removed from the population within short time.
However, an average over n is difficult to interpret in the limit, since there is no uniform distribution on $\mathbb{N}$. Looking at the lifetime distribution from a slightly different perspective, we come to a conclusion that is more sound, but slightly less general. Let $m \in \mathbb{N}$ denote a (large) integer and consider the probability $\Pr(l_{k-1} > m | l_k \leq m)$. Note that we have $\{l_k \geq m\} \subset \{l_{k-1} \geq m\}$. Using the definition of a conditional probability, we may rewrite this in the following way:

$$\Pr(l_{k-1} > m | l_k \leq m) = 1 - \frac{\Pr(l_{k-1} \leq m)}{\Pr(l_k \leq m)}$$

Next, note that

$$\begin{aligned}\Pr(l_{k-1} \leq m) &= \sum_{j=n-k}^{m-1}\Pr(l_k = j)\left[\sum_{i=0}^{m-1-j}\left(1-\frac{k(k-1)}{n^2}\right)\frac{k(k-1)}{n^2}\right] \\ &= \sum_{j=n-k}^{m-1}\Pr(l_k = j)\left[1-\left(1-\frac{k(k-1)}{n^2}\right)^{m-j}\right] \\ &= \Pr(l_k \leq m)\mathbb{E}\left[1-\left(1-\frac{k(k-1)}{n^2}\right)^{m-j} | l_k \leq m\right]\end{aligned} \tag{4.6}$$

which we achieve by combining all possibilities of removing the node labelled k in exactly j steps with all possibilities of moving the node labelled $k-1$ down by one layer before, and interpreting the internal summation on the right of 4.6 as a geometric series. Using this, the term above becomes

$$\Pr(l_{k-1} > m | l_k \leq m) = \mathbb{E}\left[\left(1-\frac{k(k-1)}{n^2}\right)^{m-j} | l_k \leq m\right] \tag{4.7}$$

This expression offers a way of relating the life times of coalescent events with each other, such that we may put them into context with the Borel-Cantelli Lemma. As

an example application, we assume $m = n^2$ and consider $k \geq \lceil \sqrt{\epsilon} n \rceil + 1, \epsilon > 0$. We may calculate that for $j \leq n^2 + \frac{2\log(k)}{\log\left(1-\frac{k(k-1)}{n^2}\right)}$, we have $\left(1 - \frac{k(k-1)}{n^2}\right)^{m-j} \leq \frac{1}{k^2}$, and importantly,

$$n^2 + \frac{2\log(k)}{\log\left(1 - \frac{k(k-1)}{n^2}\right)} \geq n^2 + \frac{2\log(k)}{\log(1-\epsilon)} := c(n,k) \tag{4.8}$$

Thus we may estimate

$$\begin{aligned} & \mathbb{E}\left[\left(1 - \frac{k(k-1)}{n^2}\right)^{m-j} | l_k \leq n^2\right] \qquad (4.9) \\ \leq \quad & \frac{1}{k^2} \cdot \Pr(l_k \leq c(n,k) | l_k \leq n^2) + 1 \cdot \Pr(l_k \geq c(n,k) | l_k \leq n^2) \end{aligned}$$

We obtain an upper bound for the sum of all the conditional expectations for $k > \sqrt{\epsilon} n$:

$$\sum_{k=\lceil \sqrt{\epsilon} n \rceil + 1}^{n-1} \mathbb{E}\left[\left(1 - \frac{k(k-1)}{n^2}\right)^{m-j} | l_k \leq m\right] \tag{4.10}$$

$$\leq \sum_{k=\lceil \sqrt{\epsilon} n \rceil + 1}^{n-1} \frac{1}{k^2} \Pr(l_k \leq c(n,k) | l_k \leq n^2) + 1 \cdot \Pr(l_k \geq c(n,k) | l_k \leq n^2) \tag{4.11}$$

Furthermore, clearly $\Pr(l_k \geq c(n,k) | l_k \geq n^2) \leq \Pr(l_{\lceil \sqrt{\epsilon} n \rceil + 1} \geq c(n, n-1) | l_k \leq n^2)$. Markov's inequality and Equation 4.4 yield

$$\begin{aligned} & \Pr(l_{\lceil \sqrt{\epsilon} n \rceil + 1} \leq c(n, n-1) | l_k \leq n^2) \qquad (4.12) \\ \leq \quad & \mathbb{E}\left[l_{\lceil \sqrt{\epsilon} n \rceil + 1} \mid l_k \leq n^2\right] c(n, n-1)^{-1} \\ \leq \quad & \mathbb{E}\left[l_{\lceil \sqrt{\epsilon} n \rceil + 1}\right] c(n, n-1)^{-1} \\ \leq \quad & \frac{n^2}{\sqrt{\epsilon} n} \frac{1}{\alpha n} \end{aligned}$$

for some $\alpha > 0$. Hence, the sum necessarily converges.

Theorem 2. *In the infinite-population limit of the EMG, all nodes of label $k \geq \sqrt{\epsilon} n$ for $1 > \epsilon > 0$ are removed with positive probability within two units of coalescent time.*

Proof. The "counterpart of the Borel-Cantelli Lemma" [Bru80] states that for a sequence $(\mathcal{A}_k)_{k \in \mathbb{N}}$, $\mathcal{A}_{k-1} \subseteq \mathcal{A}_k$ of events, the probability of infinitely many $\mathcal{A}_i$ occurring equals 1 if and only if there exists an increasing sequence $(t_i)_{i \in \mathbb{N}}$ such that the sum $\sum_{k=1}^{\infty} \Pr(\mathcal{A}_{t_{i+1}} \mid \bar{\mathcal{A}}_{t_i})$ diverges. Letting $n \to \infty$, the sequence

$$\mathcal{A}_1 := \{l_{n-1} \geq n^2\}, \ldots, \mathcal{A}_{n - \lceil \sqrt{\epsilon} n \rceil - 1} := \{l_{\lceil \sqrt{\epsilon} n \rceil + 1} \geq n^2\}$$

meets the requirements of this statement, but we have shown already that the sum does converge. In rescaled time, n^2 Moran steps correspond to two units of coalescent time. □

4.5 Conclusion I

In this chapter, we have considered a process which can be considered an extension of the classical Moran Model, in that a genealogical structure of the population exists at all times and the change in this genealogy over time reflects the change the population undergoes. It suffices to consider Yule Trees of size n to represent such a genealogy, which allowed us to define the *EMG* as a Markov Chain on Yule Trees.
From a combinatorial perspective, it is not surprising that reverting the two operations performed on the genealogy under the *EMG* (splitting and removing) is equivalent to a time-reversal of this chain. Inspection of the transition probabilities in a Markov Chain featuring "mergings" and "revivals" of individuals reveals that this indeed constitutes a time-reversal of the *EMG*, which we have denoted by $EMG^\flat$.
Both processes are defined upon a finite population, and while our calculations often allow a glimpse at the situation if n is taken to infinity, those results are to be interpreted with some caution, as a limiting process would need to be precisely determined. The underlying Moran Model does converge in rescaled time, and so the genealogy does (to Kingman's Coalescent); relying on these facts, our predictions in graph-theoretical sense still hold for the Coalescent genealogy of the entire population.
The consideration of these processes allows access to statistical properties of the genealogy that can not be recovered in the Moran Model itself. For the *EMG*, the tree balance-process serves as an example; since its transition probabilities are equal to a two-allele Moran Model, we obtain an accurate description in the limit in terms of a diffusion process. The $EMG^\flat$ appears to be extremely useful to study the life-times of tree nodes, since we know the probabilities with which nodes change their label, i.e., are moved down in the tree due to regraftings. In particular, we gain access to the probabilities of root jumps, and may calculate the expected number of such jumps during neutral fixations. The result obtained considering $n \to \infty$, which yields the number 2 for this expectation, is certainly not new, as the Lookdown-Construction can be utilized to obtain the same. To our knowledge, however, there is no corresponding formula for finite population sizes that has previously been derived.
Under the $EMG^\flat$, we can also make more general predictions about the life-times of the internal nodes of the genealogy. Especially Equation 4.7 seems useful, because it offers a way of calculating the survival probability of an infinite number of such nodes according to the Borel-Cantelli principle. With that, it becomes possible to assess how much of the genealogy is removed and restructured during a given time interval. A point of interest is whether Theorem 2 is already optimal or it can be extended to "all" nodes, and whether $m = n^2$ is necessary. Further possibilities of future research will be discussed in section 6.3.

Chapter 5

Trees in space and Linkage Disequilibrium

5.1 Motivation: Linkage Disequilibrium in finite populations

Switching gears slightly, we will consider the change in trees that is observable along the chromosome in recombination models, instead of the change of a whole-population genealogy over time in this chapter. The Ancestral Recombination Graph (*ARG*) and its Markovian approximation, the Sequential Markov Coalescent (*SMC*), have already been introduced briefly in Section 3.3 and provide a cornerstone construction of the theory of neutrally evolving and recombining populations. In much the same way the Coalescent facilitates this in non-recombining populations, *ARG* and *SMC* can be used to develop an understanding of the ancestral process and to provide a "neutral expectation" (i.e., in accordance with the neutral theory) of quantities and statistics observable in data.
In Section 3.4, we also encountered the concept of *Linkage Disequilibrium* (*LD*). *LD* is one quantity of interest under recombination, because it provides a measure of the evolutionary connection between loci which may be physically far apart. The statistic associated to this term considers two loci; usually, it is calculated for pairs of polymorphic sites, where a polymorphic site refers to a nucleotide position sporting a *SNP*. Most definitions, however, make use of the terminology of loci and alleles.
Consider a Wright-Fisher Population of $2N$ chromosomes and let α, β be two loci with alleles a, A and b, B with allele frequencies $f(a)$, $f(A)$, $f(b)$ and $f(B)$. Assume that the the four haplotypes ab, aB, Ab and AB occur with frequencies $f(ab) = x_1$, $f(aB) = x_2$, $f(Ab) = x_3$ and $f(AB) = x_4$ (Figure 5.1). Two-locus linkage disequilibrium is

$$D_{\alpha,\beta} := x_1 x_4 - x_2 x_3 \tag{5.1}$$

This can also be written as $D_{\alpha,\beta} = x_1 - (x_1 + x_2)(x_1 + x_3) = f(ab) - f(a)f(b)$. The quantity D can be interpreted mathematically as the covariance of presence/absence of the alleles a and b. A haplotype configuration such that $D_{\alpha,\beta} = 0$ is called *linkage equilibrium*. In this case all haplotype frequencies are identical to the product of the involved allele frequencies. The term "disequilibrium" indicates that *LD* is a measure of the deviation from this; the intention behind that will be explained below.
The value D is dependent on the allele frequencies, which is in some sense unfavourable. For instance, the maximum value, $1/4$, can only be attained if all allele frequencies are at $1/2$. But if we only want to observe whether co-occurrence of certain alleles is significant in some sense, allele frequencies should not matter for the measure we use. To remedy this, several standardizations have been introduced.

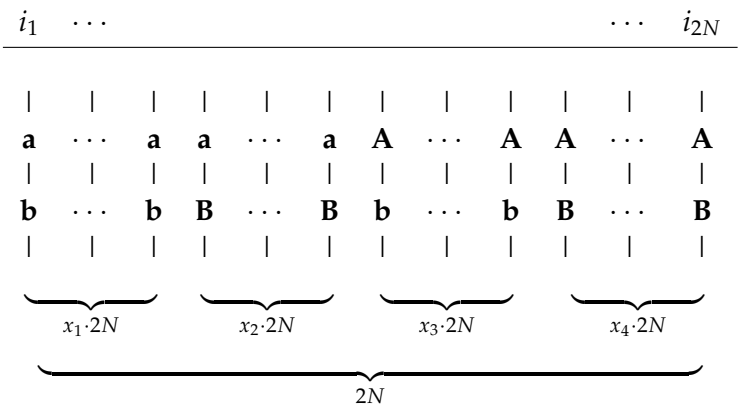


FIGURE 5.1: Alignment of $2N$ haplotypes $i_1, \cdots, i_{2N}$, indicated by vertical lines, with two bi-allelic loci. Counts given below the braces.

One is Pearson's correlation coefficient of the allelic association

$$r_{\alpha,\beta} := \frac{D_{\alpha,\beta}}{\sqrt{f(a)f(A)f(b)f(B)}} \tag{5.2}$$

which can take values between -1 and 1; also, the sign only depends on the labelling of alleles. We have

$$D_{\alpha,\beta} = 0 \iff r_{\alpha,\beta} = 0$$

Since arguably the sign is not of particular interest, we define another measure of linkage by

$$r^2_{\alpha,\beta} = \frac{D^2_{\alpha,\beta}}{f(a)f(A)f(b)f(B)} \tag{5.3}$$

which also has a mathematical interpretation in the form of the χ^2-value of the 2×2-table of the allelic association. An alternative, frequently used, standardization

$$D'_{\alpha,\beta} = \frac{|D_{\alpha,\beta}|}{\min\{f(A)(1-f(B)), f(B)(1-f(a))\}}, \tag{5.4}$$

known as "Lewontin's D" [Lew64] should also be mentioned in this context. For mathematical reasons, we focus on $r^2_{\alpha,\beta}$.

One primary field of application of *LD* is the estimation of recombination rates, as we will see that in neutral populations, there is a direct connection between the recombination rate on a chromosome segment, and the value of $D_{\alpha,\beta}$, $r_{\alpha,\beta}$ and $r^2_{\alpha,\beta}$ for two loci α, β at opposite sides of this segment. There are several other applications of *LD*; two major ones will be outlined in the following. The first is the so-called "hitchhiking effect" that was also mentioned in Section 3.4; the other is (epistatic) interaction.

In the standard Wright-Fisher Model without recombination and with mutation assuming infinite sites, D can never be 0, as this model of mutation does not allow a situation where all four possible haplotypes exist for two polymorphic loci. To see this, think of a Coalescent Tree: Two different mutations can either be "nested" (one below the other) or "disjoint" (affecting non-overlapping parts of the tree; see also [Fu95; KF18]); but the same mutation cannot occur twice. Recombination, on the other hand, provides a way of generating missing haplotypes by "shuffling" the

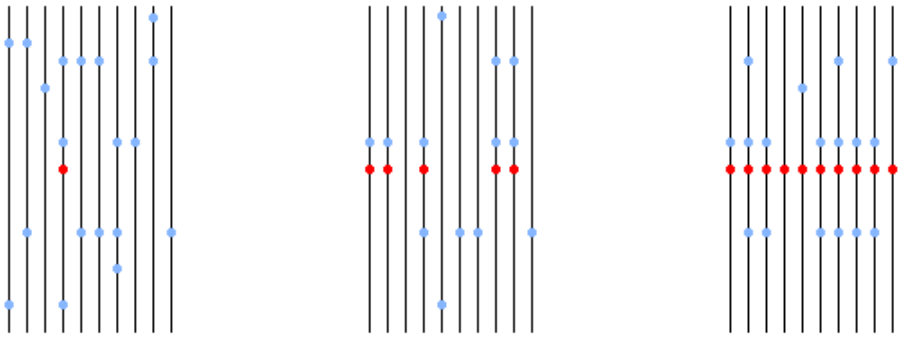


FIGURE 5.2: Illustration of the hitchhiking effect. Haplotypes are represented as vertical lines. In some generation (*left*), a beneficial allele (red variant) appears, which quickly spreads to the population until it gets fixed (*middle, right*). In the process, much of the variation in the region is removed; only variants which are initially linked to or recombine into a haplotype carrying the selected allele can typically be found. New variants are rare, because of the swiftness of a sweep. Linkage among remaining variants tends to be high.

alleles present on chromosomes. In large populations, recombination of sufficient rate in fact inevitably leads to a situation over time where haplotypes occur almost exactly at a frequency given by the product of the frequencies of the required alleles. There exists an equilibrium between drift and recombination, and of course, polymorphisms may be even entirely removed from the population, but the general rule-of-thumb is that the larger the recombination rate is, the more we expect values of $D_{\alpha,\beta}$ around zero [Gil98]; and a similar statement can be made for $r^2_{\alpha,\beta}$.
However, in a population model having experienced a recent selective sweep due to the introduction of an allele with a selective advantage (see Section 3.4), the quick spread of this allele compared to a neutral fixation means that recombination has less time to achieve that. Neutral variants at other loci that initially co-occur with this allele by chance, will attain high frequencies in the population, which is known as "genetic hitchhiking", or the hitchhiking effect [SH74]. Pairs of such alleles tend to yield a high value of $r^2_{\alpha,\beta}$. With increasing distance on the chromosome, and thereby increasing recombination rate between loci, this effect becomes less visible. Still, one expects to see elevated levels of $r^2_{\alpha,\beta}$ in the vicinity of a locus that has experienced or is in the late stages of a selective sweep. This suggests to look for such elevated levels of $r^2_{\alpha,\beta}$ between pairs of *SNP*'s in chromosomal regions of restricted size to identify potentially selected loci in an organisms genetic composition.
On the other hand, also loci that are distant may yield high values of $r^2_{\alpha,\beta}$. Of course, this can also happen by chance in a neutral setting, but it may also be due to some form of interaction, which may e.g. dictate that individuals require a certain combination of alleles at those loci to be fit; such a situation is commonly referred to as *epistatic interaction* and will indeed induce a tendency of alleles at those loci to become correlated. Using *LD* to detect such long-range interactions is in the spirit of so-called *Genome-wide Association Studies* (*GWAS*), which concern themselves with the "interplay" between loci and aim at identifying epigenetic influences on evolution.
To carry out such analyses, it is necessary to develop a "neutral expectation" of *LD* (in particular, $r^2_{\alpha,\beta}$). For instance, to be able to tell what an "elevated" level of *LD*

is, we need an estimate of the quantity $\mathbb{E}(r^2_{\alpha,\beta})$ with respect to distance between α and β in a neutral population. In models of uncountably infinite population size (e. g. Hardy-Weinberg), it is possible to derive a differential equation for D, and an expectation of $r^2_{\alpha,\beta}$ based on that. In finite populations and their limiting processes, however, this is a notoriously difficult problem; and yet another layer of complexity is added by the fact that one needs to consider $r^2_{\alpha,\beta}$-values observed in samples, since it is unrealistic to assume that whole populations can be sampled.
Using the theory developed in the following sections, we will at least obtain a partial solution. In this analysis, trees will enter the picture once more, as we know that the genealogical history of a sample from a recombining population, given by an *ARG*, is composed of a sequence of Coalescent Trees, and mutations found on the branches of such trees are precisely the two-allelic loci to choose α and β from.

5.2 The limiting value of $r^2_{\alpha,\beta}$

In fact, one part of the solution to the problem discussed above was given long ago by Haldane (see [Hal40]), but has only rarely found entry into the literature since; one reason may be, as we will discuss, that there is a certain difficulty of interpreting it. On a broader scale, in this work an expectation of the χ^2-value of "random $m \times m$-tables" was derived; incidentally, setting $m = 2$ yields the correct value $1/(n-1)$ for the expectation $\mathbb{E}(r^2_{\alpha,\beta})$, where α and β are "unlinked", and n may be either interpreted as the sample or population size, if the population is finite.
The problem lies in the interpretation of the word "unlinked". Recall that the recombination rate ρ in the *ARG* is the limit of the product $4Nc$, where c is the individual chance of each haplotype of the next generation being subject to a crossover. Haldane's value turns out to be correct if $c = 1$; this means that each locus-α-allele in the population chooses its partnering β-locus allele uniformly and independently of its partner in the previous generation. However, $c \to 0$ is required in the derivation of the *ARG*, and a recombination probability of 1 is usually dismissed for practical reasons - $c = 1/2$ intuitively makes much more sense as the maximal value. After all, recombination is not a "requirement" during meiosis for the resulting haplotypes to be able to form viable offspring.
Still, this statement is of some significance. We will see in the following sections that $\mathbb{E}(r^2_{\alpha,\beta})$ of a sample from the infinite-population limit of a Wright-Fisher Model converges to this value as ρ becomes large. For now, we offer a proof adapted for the current setting:

Lemma 7 (Haldane)**.** *Consider a Wright-Fisher population of $2N$ chromosomes, two unlinked loci α, β and a random sample of size $n \leq 2N$. Let the sample allele frequencies be $f(a) = s/n > 0$, $f(A) = (n-s)/n > 0$ and $f(b) = u/n > 0$, $f(B) = (n-u)/n > 0$. Then, sample mean and variance of $r^2_{\alpha,\beta}$ are*

$$\mathbb{E}(r^2_{\alpha,\beta}) = \frac{1}{n-1} \tag{5.5}$$

$$Var(r^2_{\alpha,\beta}) = \frac{\kappa - 1}{(n-1)^2}, \tag{5.6}$$

where κ is the fourth standardized moment (kurtosis) of a hypergeometric random variable.

Proof. Consider chromosomes carrying allele a. Among those $0 \leq k \leq s$ may also carry allele b. In fact, $k := n \cdot f(a,b)$ is a hypergeometrically distributed random

variable with parameters n, s, u, since alleles at the loci α and β choose their partner independently under the assumption of $c = 1$. So

$$\Pr(f(a,b) = k/n) = \frac{\binom{u}{k}\binom{n-u}{s-k}}{\binom{u}{s}}$$

Since

$$r^2_{\alpha,\beta} = \frac{\left(\frac{f(a,b)}{n} - \frac{s}{n}\frac{u}{n}\right)^2}{\frac{s}{n}(1-\frac{s}{n})\frac{u}{n}(1-\frac{u}{n})}$$

we have for the expectation

$$\mathbb{E}(r^2_{\alpha,\beta}) = \sum_{k=0}^{s} \frac{\binom{u}{k}\binom{n-u}{s-k}}{\binom{n}{u}} \frac{(nk-su)^2}{s(n-s)u(n-u)}$$

The denominator of the term $\frac{n^2(k-s\frac{u}{n})^2}{s(n-s)u(n-u)}$ is independent of k and can be extracted from the summation. The remainder of the summation can then be written as

$$\sum_{k=0}^{s} \frac{\binom{u}{k}\binom{n-u}{s-k}}{\binom{n}{u}} n^2 \left(k - s\frac{u}{n}\right)^2 = n^2\, Var(K) = n^2 \frac{s(n-s)u(n-u)}{n^2(n-1)}$$

such that the entire equation simplifies to

$$\mathbb{E}(r^2_{S,U}) = \frac{s(n-s)u(n-u)}{s(n-s)u(n-u)(n-1)} = \frac{1}{n-1}$$

independently of u and s. To obtain the variance, we first write

$$Var(r^2_{\alpha,\beta}) = \mathbb{E}(r^4_{\alpha,\beta}) - \mathbb{E}(r^2_{\alpha,\beta})^2 = \mathbb{E}(r^4_{\alpha,\beta}) - \frac{1}{(n-1)^2}$$

and

$$\mathbb{E}(r^4_{\alpha,\beta}) = \sum_{k=0}^{s} \frac{\binom{u}{k}\binom{n-u}{s-k}}{\binom{n}{u}} \frac{(nk-su)^4}{(s(n-s)u(n-u))^2}$$

The last expression simplifies to

$$\frac{1}{(n-1)^2}\mathbb{E}\left(\frac{f(a,b)-\mu}{\sigma}\right)^4$$

with $\mu = \frac{su}{n}$ and $\sigma^2 = \frac{s(n-s)u(n-u)}{n^2(n-1)}$. Therefore,

$$Var(r^2_{\alpha,\beta}) = \frac{1}{(n-1)^2}\kappa$$

where κ is the kurtosis of $f(a,b)$. □

Note that these results also do not depend on the allele frequencies s or u. Those might therefore be assumed to follow an arbitrary distribution, e.g. the uniform distribution, or, with *SNP*'s in mind, the neutral frequency spectrum.
Lemma 7 covers the limiting case of infinite recombination rate. If the recombination rate between α and β is given by a real number $0 < \rho < \infty$, no closed analytic expression for $\mathbb{E}(r^2_{\alpha,\beta})$ is known. There have been some attempts to derive approximations

in the past; one popular approach by Sved [Sve71] relies on comparing $\mathbb{E}(r^2_{\alpha,\beta})$ to the *conditional probability of linked identity by descent* (*LIBD*), because the expectation of the latter can be determined recursively. In this work, a finite Wright-Fisher model is assumed, and $r^2_{\alpha,\beta}$ is calculated across the whole population. In the end, the following approximation is obtained:

$$\mathbb{E}(r_{\alpha,\beta}(c)^2) \approx \frac{1}{1+4Nc\frac{1-\frac{c}{2}}{(1-c)^2}} \approx \frac{1}{1+4Nc}, \tag{5.7}$$

Compared to simulations (see also [BL07]), this formula seems to do quite well, and there have been few attempts over the years to improve this formula directly (e.g. [Obe+13], where the resulting formula is a composite of the term $1/(n-1)$ and Sved's). However, no sensible value of c exists such that it becomes equal to the limiting value predicted by Haldane, and, a fortiori, this is not achieved for the supposed limit value $c = 1/2$. Thus, clearly a discrepancy between eqs (5.5) and (5.7) exists, a fact which even Sved himself has expressed his discomfort with.[1]

5.3 Correlation between trees

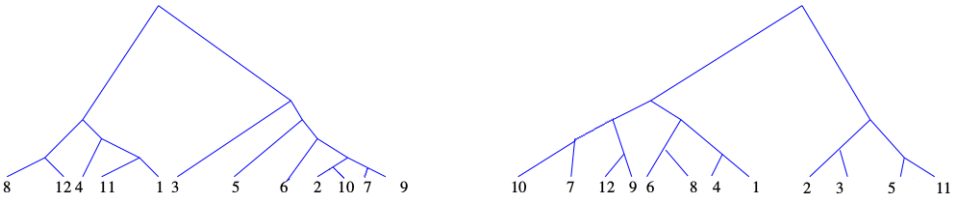
Within the framework of coalescent theory and ancestral recombination graphs, it is possible to define a slightly altered concept of linkage disequilibrium by considering trees instead of polymorphic sites. Consider a Wright-Fisher Model with recombination in the limit, i.e. $N \to \infty$ and $4Nc \to \rho$, where each individual is a haplotype of infinite sites embedded in the unit interval $[0,1]$. We already established in Section 3.3 that the ancestral process of a sample is given by the *ARG*. We introduce the following conventions:

Definition 16.

1. We denote by A_n the ancestral recombination graph of a sample of size n.
2. A real number $\gamma \in [0,1]$ is called a *site* of the chromosome.
3. An interval $S = [a,b) \subset [0,1]$ of maximal length such that no γ, $a < \gamma < b$ is a crossover site in A_n is called a chromosomal *segment*.
4. We denote by G_S the Coalescent Tree obtained as the restriction of A_n to segment S, and by G_γ the Coalescent tree obtained as the restriction of A_n to a single site $\gamma \in [0,1]$.

If γ and γ' are two different sites, but contained in the same segment S, the valid genealogy is G_S in each case. Two trees G_S and $G_{S'}$ at different segments S, S' may, but need not, be different, because recombination may, or may not, alter the tree. For a sample of size n we expect a number of ρa_{n-1} recombination events in A_n, and therefore $\rho a_{n-1} + 1$ segments.
The family $(G_\gamma)_{\gamma\in[0,1]}$ can be viewed as a (non-Markovian), continuous stochastic process on the set of Kingman coalescent genealogies with state changes caused by recombination events. A Markovian approximation of $(G_\gamma)_{\gamma\in[0,1]}$ exists in the form of the *Sequential Markov Coalescent* (see Section 3.3). In the *SMC*, genealogies change along the chromosome by uniformly choosing a branch of the current genealogy, removing the subtree below and re-inserting it somewhere else in the tree, which is

[1] See his report "Linkage Disequilibrium" on www.handsongenetics.com, labelling his endeavour in this regard a "sorry saga".



	S_1	S_2	U_1	U_2
Label	1,4,8,11,12	2,3,5,6,7,9,10	1,4,6,7,8,9,10,12	2,3,5,11
Frequency	$p(S_1) = 0.4166\cdots$	$p(S_2) = 0.5833\cdots$	$p(U_1) = 0.6666\cdots$	$p(U_2) = 0.3333\cdots$
	$S_1 \cap U_1$	$S_1 \cap U_2$	$S_2 \cap U_1$	$S_2 \cap U_2$
Label	1,4,8,12	11	6,7,9,10	2,3,5
Frequency	$x_1 = 0.3333\cdots$	$x_2 = 0.0833\cdots$	$x_3 = 0.4166\cdots$	$x_4 = 0.25$

FIGURE 5.3: Example for the calculation of *tLD* between two segments (not shown) S and U, with two coalescent genealogies G_S and G_U for a sample of size $n = 12$. The leaves of the two trees are labelled from 1 to 12. In this case, we have $r^2_{S,U} = 0.044$.

called a *prune-regraft* operation.
The *SMC*-construction is considered a reasonably accurate approximation of the *ARG*. Throughout the next sections, we assume that $(G_\gamma)_{\gamma\in[0,1]}$ is a realization of the *SMC*. $(G_\gamma)_{\gamma\in[0,1]}$ thus becomes a continuous Markov process, with changes in G_γ occurring at random positions in the unit interval with the same rate as in the *ARG*, which again define chromosomal segments in the sense of Definition 16.
The genealogy G_S extracted from $(G_\gamma)_{\gamma\in[0,1]}$ for any segment S provides a natural classification of the sample into two disjoint sets S_1 and S_2: those chromosomes which belong to the "left", and those which belong to the "right" subtree under the root node of G_S, respectively. This classification into "left" and "right" can also be interpreted as two different alleles, originating by a point mutation along one of the root branches. Moving from segment S to another segment U, the tree G_S changes to tree G_U, which may be different as a result of recombination. As a consequence, also the left (U_1) and right (U_2) descendants below the root of G_U may differ from S_1 and S_2 (Figure 5.3).

To measure correlation, let

$$f(S_i) = |S_i|/n,\ f(U_i) = |U_i|/n\ (i = 1,2)$$

and

$$x_1 = |S_1 \cap U_1|/n,\ x_2 = |S_1 \cap U_2|/n,\ x_3 = |S_2 \cap U_1|/n,\ x_4 = |S_2 \cap U_2|/n\,.$$

With this, we formulate the following

Definition 17.

$$r^2_{S,U} = \frac{(x_1x_4 - x_2x_3)^2}{p(S_1)p(S_2)p(U_1)p(U_2)} \tag{5.8}$$

is called *topological linkage disequilibrium* of the segments S and U, in short *tLD*.

By this definition, we obtain a measure of the correlation between the Coalescent Trees themselves, and extend the concept of *LD* beyond the consideration of polymorphic sites. However, as the reader may remember, Coalescent Trees, while topologically equivalent to Yule Trees on the whole, feature neither an implicit planar embedding nor an orientation of branches; therefore, classifying the subtrees below the root into "left" and "right" is problematic. But this problem can be worked

around: In fact, one of the main reasons to consider the squared correlation r^2 is that it is independent of the specific labelling of alleles, and by the same logic, also of the classification of the root subtrees. Because of that, we are free to label any subtree of a given Coalescent Tree as the left, without affecting *tLD*. Note that $r^2_{S,U} = 1$, if and only if $S_1 = U_1$ or $S_1 = U_2$. Such a configuration is called *complete linkage*.
Originally, this idea resulted from the application of the T_3 statistic [LW13; Rau18], a topology-based statistic to test for the hypothesis of neutral evolution. It relies on tree balance (e.g., Ω_1 is involved in the calculation), and on a partial resolution of Coalescent Tree topology. As a by-product, one obtains an estimation of root subtree clusters S_1, S_2 at a segment, and the observed change of such clusters along the genome gave rise to the hypothesis that there may be some information contained in the way individuals associate themselves in the trees that are encoded by the *ARG*.
Conceptually, *tLD* can be analyzed directly within the framework of Coalescent Theory. There are also some quantitative differences between *tLD* and *SNP*-based (classical) *LD*. For instance, *tLD* is affected only by topological changes at the root of a coalescent tree. This is in contrast to classical *LD*, which may (but need not) be affected by any recombination event on the tree. We explore this in more detail in Section 5.5.

5.4 *tLD* over large distances

In the following, we disregard branch lengths of individual coalescent trees G_γ of the *SMC* and consider the sequence

$$(T_\gamma)_{\gamma\in[0,1]} = (E(G_\gamma))_{\gamma\in[0,1]}$$

of canonical embeddings into the set of labelled trees; i.e. $T_\gamma \in \mathcal{L}_n$ for $\gamma \in [0,1]$. Note that $(T_\gamma)_{\gamma\in[0,1]}$ is a finite-state Markov process, as the transitions between labelled trees initiated by the *SMC* are random prune-regraft operations (see Procedure 4). We will reiterate this shortly.
Assume $T_\gamma = T_1$ for some labelled tree $T_1 \in \mathcal{L}_n$. Select a branch segment at layer k, say, to place a pruning site on the tree; cut the underlying subtree; select a branch segment at a layer less than or equal to k to place a re-grafting site; re-attach the cut subtree at the re-grafting site. This, in short, describes a prune-regraft operation and generates a second tree $T_2 \in \mathcal{L}_n$; we call the tuple (T_1, T_2) a *single recombination SMC*, for short *srSMC*. T_1 and T_2 are discrete representatives of two coalescent genealogies G_1, G_2 generated in succession under the *SMC*. The complete *SMC* can be viewed as a sequence of *srSMC*s.

An *srSMC* can also be viewed as a triplet (T_1, b_i, b_j), where T_1 is a random labelled tree, b_i and b_j denote the branch segments of the pruning and re-grafting events, respectively. Given T_1, the probability of (T_1, b_i, b_j) can be calculated as follows: The probability of pruning in layer $k > 1$ is given by

$$\Pr(\text{pruning in level } k) = \frac{(k-1)^{-1}}{a_{n-1}}. \tag{5.9}$$

(See [FDW13]). This probability is obtained by averaging over the duration of the layers in the Kingman coalescent. The branch segment b_i on which the pruning event is placed is chosen uniformly from all branch segments in layer k. By Theorem 1, the regrafting site is placed uniformly on any branch segment of layer $l \leq k$.

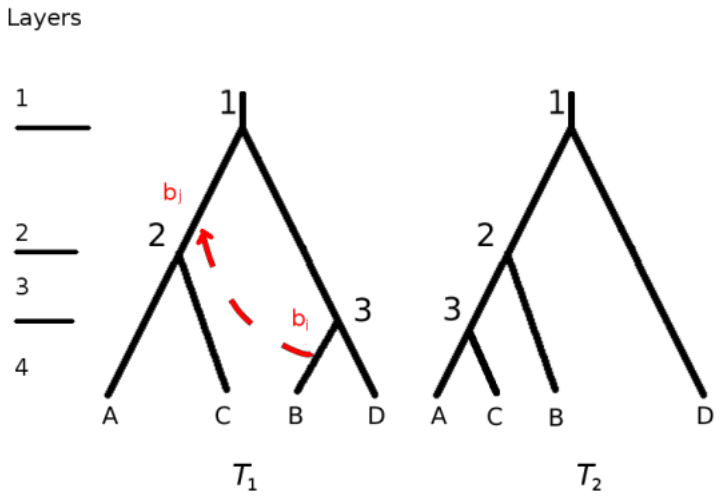


FIGURE 5.4: A *srSMC* of size 4, represented by its associated labelled trees T_1 and T_2. A subtree of T_1 is selected for pruning (b_i) and is regrafted (b_j) at a layer smaller or equal to the pruning site b_i. Note that the prune-regraft operation affects the internal labelling of T_2.

The prune-regraft operations relate to the state transition probabilities in the Markov process $(T_\gamma)_{\gamma\in[0,1]}$. An explicit formulation of this process is not as easy as e.g. in the case of the *EMG*, due to the vast number of possibilities the prune-regraft operation provides. However, it is easily seen that the set of all *Aldous moves* on a tree T_γ (see [Ald00]) is a subset of the possible transformations T_γ can undergo. Therefore, $(T_\gamma)_{\gamma\in[0,1]}$ is recurrent and aperiodic.

Consider the segments $S = [0, \gamma_1)$ and $U = [\gamma_2, 1]$, with $\gamma_1 < \gamma_2$, which are farthest apart in the *SMC*. For the expectation of their squared correlation $r^2_{S,U}$, we state the following:

Theorem 3. *Let segments S and U have trees T_S and T_U and the topological groupings (S_1, S_2) and (U_1, U_2). Then,*

$$\lim_{\rho\to\infty} \mathbb{E}(r^2_{S,U}(\rho)) \to \frac{1}{n-1}$$

Proof. $(T_\gamma)_{\gamma\in[0,1]}$ is a Markov chain with a uniform stationary distribution $\Pr^*(T) = \frac{2^{n-1}}{n!(n-1)!}$ for all labelled trees T. Since it is recurrent, for any $\epsilon > 0$ there exists an integer $M \in \mathbb{N}$ (*mixing time*), depending only on the sample size n and on ϵ, such that after M, or more, state changes in $(T_\gamma)_{\gamma\in[0,1]}$

$$\| \Pr(T_U = T|T_S) - \Pr^*(T)\|_{\mathbb{L}^1} = \sum_{T\in\mathcal{T}_n} |\Pr(T|T_S) - \Pr^*(T)| \leq \epsilon\,.$$

By choosing ρ sufficiently large, the probability of M or more changes is arbitrarily close to 1. Therefore, $\Pr(T_U = T|T_S)$ may be brought arbitrarily close to $\Pr^*(T)$. Consider the random variable $k_\rho = |S_1 \cap U_1|$ of chromosomes that are on the left of both trees T_S and T_U under the *SMC* with recombination rate ρ. As $\rho \to \infty$, T_U can be treated as generated almost independently from T_S. Say the number of individuals in U_1 is given by $|U_1| = u, u \in \{1, \ldots, n-1\}$. Then, U_1 can be treated as almost

uniformly chosen from all $\binom{n}{u}$ possible u-sized sets of individuals.
Therefore, k_ρ converges in distribution to a random variable X which, given $|S_1|$ and $|U_1|$, is distributed hypergeometrically: $X_{|S_1|,|U_1|} \sim Hyp(n, |S_1|, |U_1|)$. $|S_1|$ and $|U_1|$ themselves are uniformly distributed on $\{1, ..., n-1\}$. Additionally, for some individual $x \in S_1$, the probability that $x \in U_1$ converges to $|U_1|/n$ as $\rho \to \infty$. The assortment of one individual of the sample in the limit is therefore unlinked in the sense of Section 5.2.
Lemma 7 then applies to the situation in the limit. $\mathbb{E}(r^2_{S,U})$ therefore must converge to $\frac{1}{n-1}$, since $r^2_{S,U}$ can be treated as a continuous function of k_ρ. □

Over large distances, we recover the limiting value originally given by Haldane [Hal40] (eq (5.5)) as the expected value of *tLD*. Note that by Lemma 7 one can also obtain an exact expression for the variance of *tLD* in this situation. While the criticism of the requirements necessary to apply Lemma 7 is understandable, the formula derived there indeed approximates the expectation of *tLD* for loci which are far apart. Distance on the chromosome therefore affects a sample of an infinite population in a similar way the value c affects a finite recombining population. The simplicity of the calculation suggests that a similar statement can be made for classical *LD* as well, although the number of individuals affected by a *SNP* follows the neutral frequency spectrum instead of the uniform distribution.

5.5 Behaviour with distance and numerical approximation

For two segments S, U on the chromosome, *tLD* equals 1 if $S = U$ and declines toward $1/(n-1)$ with increasing distance between S and U. In this section, we will point out a way to approximate the behaviour in between. The stategy we will employing can be roughly described as approximating the "average change" in $r^2_{S,U}$ in an *srSMC*, that is, if S and U are neighbouring segments, and extrapolating with this to obtain estimates for larger distances. We start with the following observation:

Lemma 8. *Let T_1, T_2 be the two labelled trees resulting from a srSMC. The probability of a topological change, i.e. breaking of complete linkage, between T_1 and T_2 is asymptotically $\frac{1}{3} + \mathcal{O}\left(\frac{1}{\log(n)}\right)$.*

Proof. Recombination events that force breaking of complete linkage between neighbouring segments can be subdivided into two groups: Events which shift a non-root branch above the root or events which move a branch from the left to the right (or vice versa) root-subtree without changing the root. We call the latter *switching events*. The probability Pr(root-change) of root-changing events is $\mathcal{O}(\frac{1}{\log(n)})$ [FDW13]. To calculate the probability Pr(switch) of switching events, assume without loss of generality that a branch is moved from left to right. Suppose pruning takes place in layer k.
The right side has $1 \leq j \leq k-1$ branch segments in layer k, where each number j has probability $1/(k-1)$. The probability of selecting a k-layer branch segment on the left for pruning is $(k-j)/k$, and the probability of selecting a branch segment on the right for re-grafting is $(k-1)(j+1)/(k(k+1))$ when averaged over all k-sized labelled trees (see Proposition 4 below for a derivation). This needs to be multiplied by $1/((k-1)a_{n-1})$ (see equation (5.9)), and then summed over all levels k. We obtain

$$\Pr(\text{switch}) = 2\sum_{k=2}^{n}\sum_{j=1}^{k-1} \frac{k-j}{k}\frac{(k-1)(j+1)}{k(k+1)}\frac{1}{k-1}\frac{1}{(k-1)a_{n-1}},$$

where the factor 2 accounts for the two possibilities, switching from left to right or vice versa. After some simplifications, this can be rewritten as

$$\Pr(\text{switch}) = \frac{1}{3} + \frac{1}{3na_{n-1}} + \frac{1}{6a_{n-1}} - \frac{1}{(n+1)a_{n-1}},$$

where all terms, except the constant $1/3$, are of $\mathcal{O}\left(\frac{1}{\log(n)}\right)$ or smaller. □

We still have to supply the proof of one missing piece, namely, that the probability of choosing a branch segment on the right side, given the number j of branch segments on the right side in layer k, is $(k-1)(j+1)/(k(k+1))$ (averaged over all trees in $\mathcal{L}_n$). Since the regrafting segment is chosen uniformly, it suffices to calculate the expected number of segments on the right side, given the number of segments on the right side in layer k.

Proposition 4. *Consider the set $\mathcal{L}_n$ of all labelled trees T of size k with $1 \leq j \leq k-1$ leaves on one (referred to as the "right") side of T. Let j_T denote the number of branch segments on the right side of T. Then, we have*

$$\mathbb{E}(j_T) = \frac{(j+1)(k-1)}{2}.$$

Proof. There are $k-1$ internal nodes in T, $j-1$ht side is given by $k+1-i_{j-1}, \ldots, k+1-i_1$. Then, $j_T + j_{T'}$ equals

$$\begin{aligned}
& 1 \cdot (i_1 - 1) + j \cdot (k - k - 1 + i_1) \\
& +2 \cdot (i_2 - i_1) + (j-1) \cdot (k + 1 - i_1 - k - 1 + i_2) \\
& + \ldots \\
& +(j-1) \cdot (i_{j-1} - i_{j-2}) + 2 \cdot (k + 1 - i_{j-2} - k - 1 + i_{j-1}) \\
& +j \cdot (k - i_{j-1}) + 1 \cdot (k + 1 - i_{j-1} - 1) \\
= \quad & (j+1)(k-1).
\end{aligned}$$

From this observation, and from the fact that labelled trees are uniform under the coalescent, we conclude

$$\mathbb{E}(j_T) = \frac{(j+1)(k-1)}{2}$$

□

Knowing the proportion of recombination events contributing to the decay of $r^2_{S,U}$, we proceed to calculating the expected proportion of chromosomes affected by a switching event between two segments. Let $L_{S,U}$ denote the number of chromosomes whose assignment to either the left or right class is *not* affected by switching. The quantity $L_{S,U}$ may be interpreted as a *probability of linked identity by descent*, similarly to the parameter Q in [Sve71], which denotes the same probability conditioned on linked identity at one of the two loci. $L_{S,U} = n$ is equivalent to complete linkage and $L_{S,U} < n$ means that between S and U, there has been at least one switching or root-changing event (see Figure 5.5). Note that root-changing events entail $L_{S,U} = 0$, even if S and U are neighbouring segments. Most of the following calculatons rely on the fact that the probability of such events converges to 0 with increasing n.

Lemma 9. $\mathbb{E}(L_{S,U})$ *declines exponentially with respect to the number of recombination events separating S and U.*

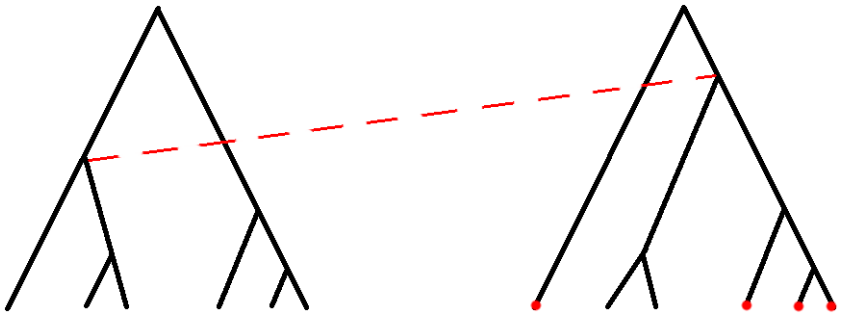


FIGURE 5.5: A single recombination event moves a branch of the left root-subtree to the right side. In the resulting tree, chromosomes marked by red dots remain in the same left/right-grouping as before recombination took place. Their number is denoted by the quantity $L_{S,U}$.

Proof. By Lemma 8, we have an estimate of Pr(switch). Recombination events are distributed uniformly over the branches of a given genealogy T in the *SMC*, meaning that the size of the subtree T_{b_i} below a recombination event is distributed according to the neutral frequency spectrum (see Section 3.2). Thus,

$$\Pr(|T_{b_i}| = k) = \frac{1}{k \cdot a_{n-1}}. \tag{5.10}$$

If we take the average over all *srSMC*s, the expected proportion of chromosomes affected by a recombination event is the expectation of the above distribution divided by n, i.e.

$$\frac{n-1}{n\, a_{n-1}} \approx \frac{1}{a_{n-1}} \approx (\gamma + \log(n))^{-1},$$

where γ is the Euler-Mascheroni constant. We use this to derive an approximation for $\mathbb{E}(L_{S,U})$ recursively. Let U_1, U_2 be two neighbouring segments. Then, we have

$$\begin{aligned}
\mathbb{E}(L_{U_1,U_2}) &= \mathbb{E}(L_{U_1,U_1})\,(1 - \Pr(\text{switch}) - \Pr(\text{root-change})) \\
&\quad + \mathbb{E}(L_{U_1,U_1})\Pr(\text{switch})\frac{n-1}{n\, a_{n-1}} \\
&= \mathbb{E}(L_{U_1,U_1})\left(1 - \Pr(\text{root-change}) - \Pr(\text{switch})\frac{n-1}{n\, a_{n-1}}\right).
\end{aligned}$$

Iterating this formula with initial value

$$\mathbb{E}(L_{S,S}) = n$$

shows that

$$\begin{aligned}
\mathbb{E}(L_{S,U}) &= n\left(1 - \Pr(\text{root-change}) - \Pr(\text{switch})\frac{n-1}{n\, a_{n-1}}\right)^{c_{S,U}} \\
&\approx n\left(1 - \Pr(\text{root-change}) - \frac{1}{3(\gamma + \log(n))}\right)^{c_{S,U}}
\end{aligned} \tag{5.11}$$

with $c_{S,U}$ representing the number of recombination events between S and U, which depends only on ρ. □

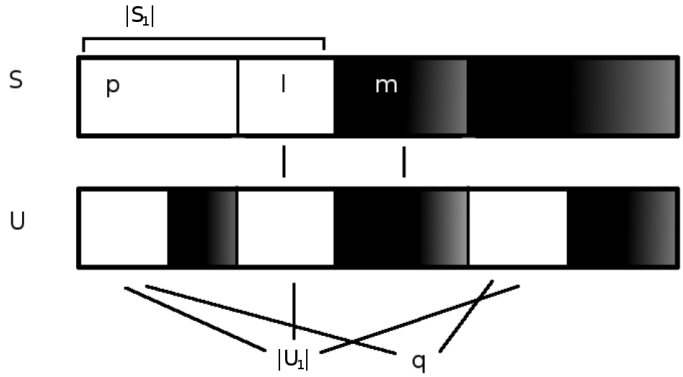


FIGURE 5.6: At segments S and U there are chromosomes on the left and right side of the trees, indicated as white and black bars, respectively. When moving from S to U, $L_{S,U} = l + m$ chromosomes remain on their sides, and the rest switches sides.

We close this section with the remark that all terms in eq (5.11) can be calculated explicitly for given sample size n, and that by Lemma 8 (cf. [FDW13]), we have the approximation

$$\mathbb{E}(L_{S,U}) \approx n\left(1 - \mathcal{O}\left(\frac{1}{\log(n)}\right)\right)^{c_{S,U}}.$$

The above results about the decline of $L_{S,U}$ suggest an approximation scheme for the expectation of *tLD* with respect to the number of recombination events separating two segments. $L_{S,U}$ can be written as $L_{S,U} = l + m$, where l (m, respectively) is the number of $L_{S,U}$-chromosomes on the left (right) side of both trees. There are $p = |S_1| - l$ additional individuals on the left side of T_S and $q = |U_1| - l$ on the left side of T_U. See Figure 5.6 for a sketch. To calculate $r^2_{S,U}$, one needs to determine how many of the additional p chromosomes on the left side of T_S are also on the left side of T_U by chance. We choose to approximate this number by a hypergeometric random variable, with the remark that this approximation becomes more accurate with increasing recombinational distance between S and U, since we have seen in Section 5.4 that the assignment in the limit is ideed hypergeometric.

Let k denote the number of chromosomes which are on the left side of T_S and of T_U by this hypergeometric assignment. Thus, in total there are $k + l$ chromosomes on the left side of both trees. Under these assumptions, the expected *tLD* is

$$\mathbb{E}(r^2_{S,U}|p,q,l,m) = \sum_{k=0}^{q} \frac{\binom{p}{k}\binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot \frac{\left(\frac{k+l}{n} - \frac{(p+l)(q+l)}{n^2}\right)^2}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)}.$$

Note, that the term

$$\sum_{k=0}^{q} \frac{\binom{p}{k}\binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot k^2$$

is the second moment of a hypergeometric random variable with parameters $n - L_{S,U}, p, q$ and with expectation $qp/(n - L_{S,U})$. Then,

$$\frac{(p+l)(q+l)}{n^2} - \frac{qp}{n \cdot (n - L_{S,U})} = \frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2}$$

and

$$\begin{aligned}
\mathbb{E}(r^2_{S,U}|p,q,l,m) &= \sum_{k=0}^{q} \frac{\binom{p}{k}\binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \frac{\left(\frac{k}{n} - \frac{qp}{n\cdot(n-L_{S,U})} - \frac{\frac{npq}{L_{S,U}-n}+(l+p)(l+q)}{n^2} + \frac{l}{n}\right)^2}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)} \\
&= \sum_{k=0}^{q} \frac{\binom{p}{k}\binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot \left(\frac{\left(\frac{k}{n} - \frac{qp}{n\cdot(n-L_{S,U})}\right)^2}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)} \right. \\
&\quad - 2\frac{\left(\frac{k}{n} - \frac{qp}{n\cdot(n-L_{S,U})}\right)\cdot\left(\frac{\frac{npq}{L_{S,U}-n}+(l+p)(l+q)}{n^2} - \frac{l}{n}\right)}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)} \qquad (\#) \\
&\quad \left. + \frac{\left(\frac{\frac{npq}{L_{S,U}-n}+(l+p)(l+q)}{n^2} - \frac{l}{n}\right)^2}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)} \right).
\end{aligned}$$

We may simplify this expression significantly, and in a meaningful way. The middle term of the summation (line #) vanishes because of symmetry; and the first summand contains the variance of a $Hyp(n-l-m, p, q)$ random variable divided by some constants. Therefore,

$$\begin{aligned}
\mathbb{E}(r^2_{S,U}|p,q,l,m) = & \qquad (5.12) \\
& \frac{\frac{pq}{n-L_{S,U}}\cdot\left(1-\frac{p}{n-L_{S,U}}\right)\frac{n-L_{S,U}-q}{n-L_{S,U}-1}\cdot\frac{1}{n^2}}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)} \qquad \text{(HYP)} \\
& + \frac{\left(\frac{\frac{npq}{L_{S,U}-n}+(l+p)(l+q)}{n^2} - \frac{l}{n}\right)^2}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)}. \qquad \text{(PSP)}
\end{aligned}$$

In this form, the contribution that arises from the hypergeometric random assignment, labelled HYP, and the remaining parameter-specific (PSP) terms are separated. This decomposition is useful in two ways. First, under the above assumptions, an upper bound can be obtained for $\mathbb{E}(r^2_{S,U})$, at least if $L_{S,U}$ is small in relation to n (see Lemma 10). Second, by averaging over all configurations it is possible to calculate an average $\mathbb{E}(r^2_{S,U})$, independently of tree topologies at segments S and U. Since the size $|S_1|$ of the left side of T_S is uniform on $\{1, \cdots, n-1\}$ (see Section 2.5), we start by choosing $|S_1|$ randomly according to the uniform distribution. The $L_{S,U}$-sized portion of chromosomes not having undergone recombination when going from S to U is then subdivided into l individuals which are on the left side both in T_S and T_U, and m individuals which are on the right side in both trees by choosing hypergeometrically from the assignment at S, which implicitly determines the parameters

p, l and m. The number q of additional individuals on the left side of T_U is determined by drawing uniformly from $\{1, \ldots, n - L_{S,U}\}$.
These calculations are easily performed by computational algebra (Figure 5.9). Note that it is much more complicated to explicitly calculate the expectation of classical *LD* according to this scheme, because the sizes of the classes are not uniformly distributed. The resulting approximation of the expected *tLD*, based on $L_{S,U}$, has to be scaled with respect to the expected decay of $L_{S,U}$. Assuming that recombination events are uniformly distributed across a (finite) chromosome, the approximation of $\mathbb{E}(r^2_{S,U})$ can be expressed as a function of physical distance between segments S and U.

Lemma 10. *Assume that the recombinational distance between S and U is large, such that $L_{S,U}$ is small compared to sample size n. Then, we have:*

(a) The parameter-specific contribution ("PSP") in eq (5.12) is of order

$$\mathcal{O}\left(\left(\frac{L_{S,U}\log(n)}{n}\right)^2\right).$$

(b) The approximation of $\mathbb{E}(r^2_{S,U})$ in eq (5.12) is bounded from above by

$$\frac{1}{n-1} + \kappa_n\left(1 - \mathcal{O}\left(\frac{1}{\log(n)}\right)\right)^{c_{S,U}},$$

where $c_{S,U}$ denotes the number of recombination events separating S and U and κ_n is some constant of the order of $\mathcal{O}(\log(n)^2)$.

Proof. Under the assumption of $L_{S,U} < n$, it is possible to write $\frac{n}{n-L_{S,U}} = \sum_{i=0}^{\infty}\left(\frac{L_{S,U}}{n}\right)^i$. Furthermore, $l \cdot (n - p - q - l) \in [-l \cdot n, l \cdot n]$. This allows us to rewrite the numerator of the PSP term in the following way:

$$\begin{aligned}
&\left(\frac{(l+p)(l+q) - \frac{npq}{n-L_{S,U}}}{n^2} - \frac{l}{n}\right)^2 \\
=&\left(\frac{pq\sum_{i=1}^{\infty}(\frac{L_{S,U}}{n})^i + l(n-p-q-l)}{n^2}\right)^2 \\
\leq&\left(\sum_{i=1}^{\infty}\left(\frac{L_{S,U}}{n}\right)^i + \frac{l}{n}\right)^2 \\
\in&\mathcal{O}\left(\left(\frac{L_{S,U}}{n}\right)^2\right)
\end{aligned}$$

The last statement is true because $l/n \leq L_{S,U}/n$. Under the assumption that recombination distance between S and U is large, and $L_{S,U}$ is small compared to n, then the sizes $|S_1|$ and $|U_1|$ of the left sides of the genealogies are approximately independent and uniformly distributed on $\{1, \ldots, n-1\}$ (see Proposition 2 and Section 2.5).

Thus, the expectation of the denominator in

$$\mathbb{E}\left(\frac{1}{\frac{|S_1|}{n}\cdot\left(1-\frac{|S_1|}{n}\right)\frac{|U_1|}{n}\cdot\left(1-\frac{|U_1|}{n}\right)}\right)$$

converges to

$$\frac{n^4}{(n-1)^2}\sum_{k=1}^{n-1}\sum_{l=1}^{n-1}\left(\frac{1}{k(n-k)l(n-l)}\right)=\frac{4n^2}{(n-1)^2}\,a_{n-1}^2$$

as $L_{S,U}/n$ becomes small. This term is of order $\mathcal{O}(\log(n)^2)$, allowing us to conclude that

$$\mathbb{E}(\mathrm{PSP})\in\mathcal{O}\left(\left(\frac{L_{S,U}}{n}\right)^2\cdot\log(n)^2\right),$$

establishing claim (a).

To show (b), we recall Hölder's inequality

$$\mathbb{E}(X^2)\leq\mathbb{E}(X)\cdot\max X,$$

for a non-negative random variable X. Let $X=\frac{L_{S,U}}{n}$, assuming exactly $c_{S,U}$ recombination events between S and U. The maximal value of this random variable is 1 (no decline at all). The expectation of X, given the number of recombination events between S and U, has been calculated in this section and is approximated by $\left(1-\mathcal{O}\left(\frac{1}{\log(n)}\right)\right)^{c_{S,U}}$. Thus,

$$\mathbb{E}\left(\left(\frac{L_{S,U}}{n}\right)^2\mid c_{S,U}\right)\leq 1\cdot\left(1-\mathcal{O}\left(\frac{1}{\log(n)}\right)\right)^{c_{S,U}}$$

and therefore

$$\mathbb{E}(\mathrm{PSP}\mid c_{S,U})\leq\delta_n\left(1-\mathcal{O}\left(\frac{1}{\log(n)}\right)\right)^{c_{S,U}}$$

with some constant $\delta_n>0$ depending on n.
The expectation of the hypergeometric contribution in Equation 5.12 is 0 for $L_{S,U}=n$ and converges to $\frac{1}{n-1}$ from below for $L_{S,U}/n\to 0$, which establishes claim (b). □

It should be stressed, however, that Lemma 10 only applies to large distances. It makes use of the same assumptions that were made to derive Equation 5.12 and is therefore only an approximation of the situation in the *SMC*. Furthermore, the term $4a_{n-1}^2$ is not bounded from above, and because of that this upper bound is only of relevance for large n. In that case, however, because $L_{S,U}/n$ declines exponentially nevertheless, this statement may still be of use, because it allows to determine confidence intervals, for instance by the help of Markov's Inequality.

5.6 In Data

The calculation of *tLD* usually requires two steps. In the first, at the two segments S,U considered, the tree topology valid on these segments is used to obtain the respective topological groupings (S_1,S_2 and U_1,U_2) of the sample. The second one is the actual calculation of $r_{S,U}^2$.

In practice, particularly in the first step one is faced by several challenges. Most importantly, tree topology is unknown and has to be, at least partly, resolved by using *SNP* data. Ideally, since we are only interested in resolving the first split, we construct the groupings based on *SNP*s which lie on the root branches of the Coalescent Trees, as such *SNP*s would serve as perfect indicators of this grouping. However, since it is often impossible to tell whether a *SNP* has this property, the entirety of *SNP*s on a segment is taken into account and phylogenetic means are employed to resolve the upper topology.
The other major problem is that the boundaries of segments can usually not be determined reliably; one way to resolve this is to cut the chromosome into segments ("windows") of some specified length, with the hope that such a window does not contain too many recombination events (and hence, multiple topologies), but, on the other hand, enough *SNP*s to resolve the tree. These issues therefore need to be taken into account when simulating *tLD*.
Large parts of the work discussed in this section, including all programming and the generation of heatmaps, were carried out by Martina Rauscher ([WRW18; Rau18]).

Simulations The program `ms` [Hud02] can be used to generate samples of n chromosomes with recombination rate ρ and mutation rate θ. Importantly, `ms` generates data assuming an *ARG* instead of the *SMC*, which is expected to yield a closer depiction of the situation in reality.
Given a dataset simulated with `ms`, we consider two cases: (a) the true (simulated) tree structure provided by the *ARG* (option "-T" selected) is used for the calculation of *tLD*; or (b) the clusters are estimated from the polymorphism data provided by `ms`. The latter case reflects the situation in real data, where tree estimation is necessary and segment boundaries are unknown. The purpose of the first is to obtain a benchmark of the method, which can be compared to our theoretical predictions. In the case of unknown tree topology, the following clustering strategy is employed to estimate the topological grouping: First, on a given window the two most diverged haplotypes ('antipodes') are determined; then, the remaining haplotypes are assigned to either of the antipodes based on minimal Hamming distance. The estimated clusters agree well with the actual clusters provided by the Coalescent Trees if several *SNP*s are used jointly for estimation. However, to avoid the confounding effect of multiple recombination events on cluster estimation, window size should be as small as possible, which means that the number of *SNP*s to use per window can not be arbitrarily large. By simulations, it was found that as few as about ten SNPs give good results (An analysis performed by M. Rauscher, see [Rau18]; the good agreement between estimated and actual *tLD* shown in the two heatmaps in Figure 5.7 gives evidence of this). This result is also supported by the excellent agreement of the summary statistics (average and variance) of *tLD* determined from actual and estimated clusters (Figure 5.8).
Comparing *tLD* with classical *LD* in simulated data, we find that both average and variance of *tLD* are larger than those of classical *LD*. On the same distance scale, classical *LD* vanishes much more quickly (Figure 5.8). Although average *LD* is small, its variance is high compared to the average, in particular for short distances. In this regime the variance of *tLD* is much smaller relative to the average than that of *LD*, best seen in the difference of the coefficients of variation, σ/μ, of the two statistics (Figure 5.8). These observations are theoretically supported by Lemma 8.

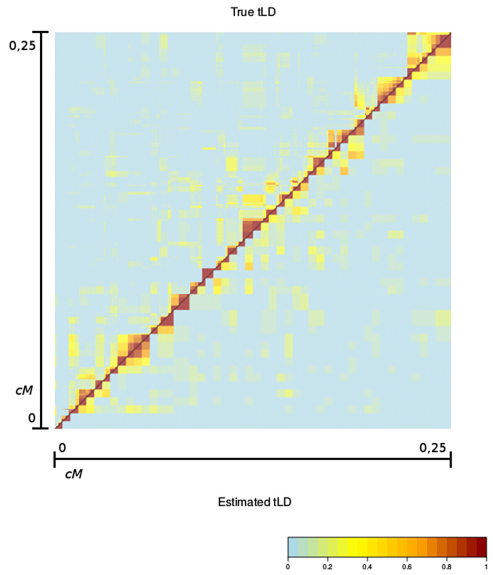


FIGURE 5.7: Heatmaps of actual (upper triangle) and estimated (lower triangle) *tLD* for a sample size of $n = 200$ across a region corresponding to 0.25cM. Both are calculated from the same `ms`-simulation with parameters $theta = 100, \rho = 100$ `ms 200 1 -t 100 -r 100 1000 -T`. Assuming a recombination rate of $c = 10^{-8}$ per *bp* per generation, and a population size of $N = 10^4$,[2] the size of the region considered is $\approx 2.5 \cdot 10^5$ *bp*. With these parameters one expects $a_{n-1}\,\rho \approx 600$ recombination events across the entire chromosome, i.e. $600/1000 = 0.6$ events per `ms`-fragment.

Application to experimental data In practice tree topology is unknown. To estimate tree topology at the tree root, we apply the same clustering approach as on simulated data. As an example *tLD* across the "*LCT* region" ($\approx$ *bp* 135900000 to *bp* 136700000) on chromosome 2 in the CEU (Americans of Central European descent) and YRI (Yorouban ancestry in Nigeria, Africa) populations (data from [Aut+15]) of *H. sapiens* was determined. We estimated *tLD* using chromosomal segments of size 5*kb* ('window size') and a step size of 2.5*kb*. Most of these windows contain ten or more SNPs. First, and unsurprisingly, we find a strongly elevated level of *tLD* in the *CEU* population compared to *YRI*. Second, there is a much higher, and a longer-ranging level of correlation to be observed for *tLD* than for conventional *LD*. Third, *tLD* is contiguously high in the regions containing the DARS and the MCM6 genes in the CEU population. Remnants of elevated *tLD* in these regions are visible also in the YRI population (Figure 5.10).

On the whole, the elevated level of *tLD* supports the well-known hypothesis that the *LCT* locus is, or has been, under positive selection in European populations of *H. sapiens*. The enzyme encoded by the *LCT* gene itself is lactase, which in turn facilitates breakdown of lactose (milk sugar) molecules. In later stages of the human life cycle, expression of this gene is regulated down, such that lactose can no more be digested, which is termed *lactose nonpersistence* (and can, in severe form, lead to the condition of lactose intolerance).

It has been hypothesized that the neighbouring *MCM6* gene harbours an enhancer of the transcription of the *LCT* gene. In this gene, a number of *SNP*s have been

[2] For human populations, values like these are typically assumed.

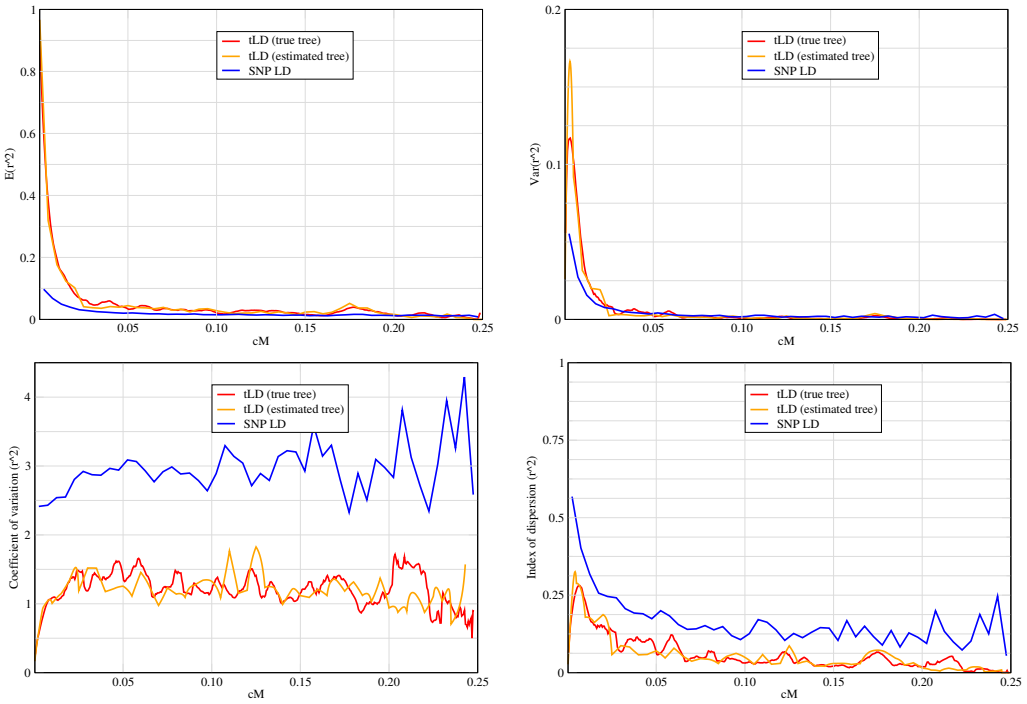


FIGURE 5.8: *tLD vs.* classical SNP-*LD*. Average (top left), variance (top right), coefficient of variation (bottom left) and index of dispersion (bottom right) of r^2. Data from a single simulation run performed with the program `ms` [Hud02]. Parameter settings: `ms 200 1 -t 100 -r 100 1000 -T`. For a poulation size of $N = 10^4$ and a recombination rate of $1cM/Mb$ the simulated region corresponds to 0.25cM or 250kb physical distance. Red: *tLD* calculated from the actual coalescent trees (i.e., using the trees obtained by setting the parameter $-T$). Orange: *tLD* calculated from estimated tree topology (see text). Blue: Classical *LD* calculated from SNP pairs. Coefficient of variation: σ/μ; index of dispersion σ^2/μ.

detected which seem to be linked to *LCT* expression at a later age; one of these variants is found at extremely high frequencies in European populations [Ena+02]. This, together with the fact that humans of northern-European descent seem to be under comparably low risk of becoming lactose-intolerant, suggests that the *LCT* and *MCM6* loci have been selected for this variant in the demographic past of northern Europe. One ecological explanation of this proposes that due to e. g. scarcity of other food in winter, it would have been advantageous for humans to be able to digest dairy products in adulthood.
One of the first studies of genetic variability in this region was conducted in 1973 [Cav73], reporting a high level of an $F_{S,T}$-related measure.[3] Later, haplotype homozygosity [Ber+04] and Tajima's D [Kor+13] were used to provide statistical evidence. Our results obtained by calculating *tLD* seem to reinforce this. Furthermore, the strong linkage between neighbouring loci (*DARS* and *MCM6*) may be interpreted as a result of the interaction between *LCT* and *MCM6*, which would fit to the hypothesis that the polymorphisms influencing fitness are actually found in the latter.
Interpreting *tLD* as a consensus value for pairs (w_1, w_2) of windows, one might raise the question how much it deviates from a simple "pooling" approach, i.e. averaging the values of classical *LD* for all pairs of *SNP*s with the property that one *SNP* lies in w_1 and the second in w_2. Since tree estimation relies on the *SNP*s just like classical *LD*, it is to be expected that at least some of the values obtained for pairs of *SNP*s must be similar to the value of *tLD*. Taking again the *LCT* region as an example and looking at the colouring patterns in the heatmaps of *tLD* and classical *LD*, one can see a somewhat similar structure, but also different magnitudes and a trend of *tLD* to stay elevated for longer distances. To quantify this, we calculated the correlation between *tLD*-values and pooled classical *LD* on this region, which turned out to be ≈ 0.3193. A positive correlation of some magnitued is expected for the reasons given above. It should be kept in mind here that tree topology is not known explicitly and that the *tLD*-values we obtain are therefore estimates of the true *tLD*.
As an example of how *tLD* may be used to estimate recombination rates, we chose another region, expanding over position $5.5\ 10^6$ to $5.6\ 10^6$ of the human chromosome 5 ([Aut+15]). The region was divided into overlapping windows of size 5kb as before, and the average of estimated *tLD* was calculated over all window pairs of a given distance, which is represented on the x-axis of the right graph in Figure 5.9.
The decline of this average with distance should resemble the expected decline of *tLD*, of which we have obtained an estimate in Section 5.5. We calculated this expectation for $n = 198$ (the CEU subsample consists of 99 diploid genotypes), scaled it to the expected decay of $L_{S,U}$, and then adjusted it with respect to the number of observed recombination events inside this region, by fitting the curves of observed and expected decline according to the method of least squares. The estimated recombination rate on this segment is $\approx 0.508 cM/Mb$.

5.7 Conclusion II

Topological Linkage Disequilibrium appears to be useful in both theoretical and practical regard. On the theoretical side, *LD* defined by tree topology can be integrated into a framework of Coalescent Theory very elegantly, where for classical *LD* the conditional two-site frequency spectrum would need to be employed, admits an

[3] $F_{S,T}$ is a measure of population subdivision; since it is not further covered in this work, we refer to [Gil98]

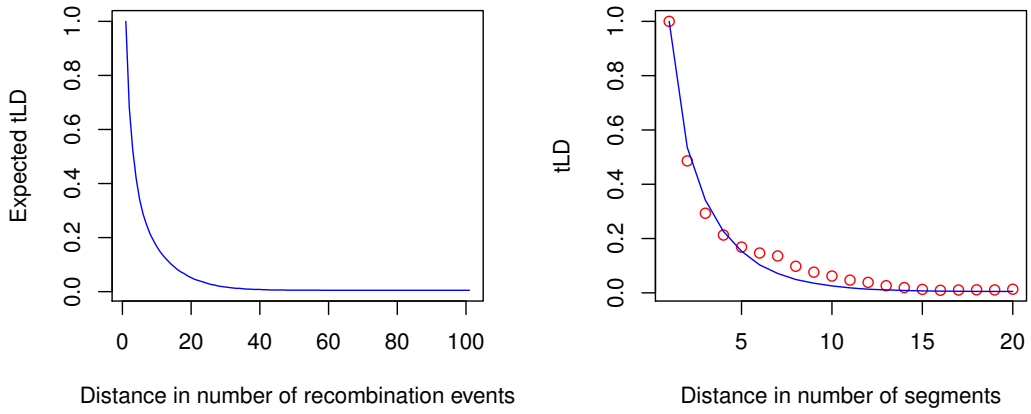


FIGURE 5.9: *Left:* Theoretical result according to eq (5.12) for a sample of size $n = 100$. 100 recombination events correspond roughly to a genomic distance of 50kb (for a calculation see Figure 5.7). *Right:* Experimental data from the 1k genomes project [Aut+15], CEU population, chromosome 5. We have randomly selected a region of about 100kb (at position 5.5 10^6 to 5.6 10^6) and calculated *tLD* for segments of size 5kb, spaced up to a distance of 50kb. Dots represent the average *tLD* between pairs of segments of a given distance (x-axis, in multiples of 2.5 *kb*). Blue line: Least-squares fit of expected *tLD*.

approximation of its expected behaviour over long ranges, and at least in part helps to resolve some of the problems that have persisted throughout the history of discussion and application of this concept. Furthermore, the methods with which *tLD* are analyzed can help to derive a more refined understanding of classical *LD* as well; Lemma 7 serves as an example, which, since it makes no assumptions on the allele frequency distributions, in principle admits reformulation to match the situation of two distant *SNP*s.

Regarding applications, the biggest advantage that *tLD* provides is the fact that it provides a sensible "consensus value" for (pairs of) genomic regions, whereas classical *LD* is extremely unstable and leads to rather noisy data. Another fact on the plus side is that we know, from theoretical investigation, the proportion of recombination events that have an effect on *tLD*, while for classical *LD*, this is as of yet unknown and would probably require extensive averaging arguments. As a result, *tLD* is expected to behave more consistently, and also to decline more slowly with distance, such that it might be possible to detect interactions between distant regions more reliably by making use of this measure. So far, however, evidence of the latter point is only provided by simulations. Mathematically, it remains an open problem to show this.

Of course, *tLD* essentially uses the same data classical *LD* is calculated from, which means that these advantages must also come with disadvantages. The trade-off one has to agree upon is that errors may happen in the estimation of tree topology, and recombination within windows may dilute the true picture even more. In fact, in many species recombination rate and mutation rate are of similar magnitude, such that it is not to be expected that there are many *SNP*s available on a segment for tree estimation. Therefore, the cost of the theoretically advantageous properties of *tLD* comes in the form of loss of precision, an effect one should seek to minimize when

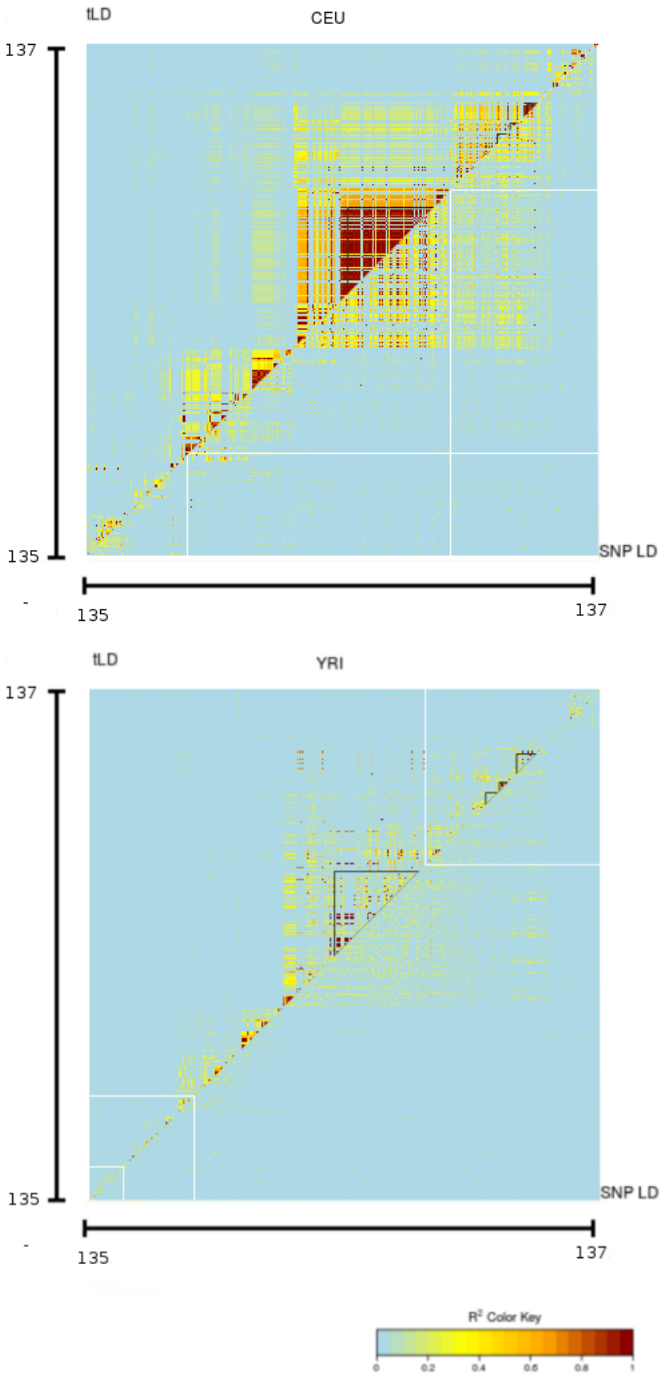


FIGURE 5.10: Heatmaps of classical (lower triangle) and topological (upper triangle) *LD* based on experimental data collected from the 1k genomes project [Aut+15]. Upper picture: CEU population. Lower picture: YRI population. Shown is a region of 2Mb containing the LCT-locus on chromosome 2 (from position 135Mb to position 137Mb in coordinates of the hg19 assembly). Dark blue triangles within the plot indicate the positions of the ZRANB3, LCT, MCM6 and DARS genes (from bottom left to top right).

implementing the method.
Computationally, *tLD* does not require much more ressources than classical *LD*. Depending on the clustering approach, the time it takes to estimate a tree topology on a window may vary, but is usually of polynomial size in n and the number S of *SNP*s per window. For example, in the case of the *two-means* approach described in Section 5.6, the runtime is of order $\mathcal{O}(n^2 S)$. Since *SNP*s may be pooled into a window, the correlation matrix obtained for *tLD* is actually smaller that that of classical *LD*. Room for improvement of the method is suspected to be found especially at the estimation of tree topology, because this is the most error-prone step of the entire procedure. It may provide a possibilitiy for future research to determine the most accurate clustering strategy for estimation of coalescent tree topology.
A correlation test on the *LCT* region reveals a positive correlation between *tLD* and pooled classical *LD*, as expected. Keeping in mind that tree topology estimation is not exact, we hypothesize that the true correlation would be at least of the same magnitude. This is certainly just to be considered an example; in general, under near perfect conditions with respect to topology estimation, we expect a consistently high positive correlation. On the other hand, a correlation below one would also indicate that the qualitative differences between the two approaches we postulate are significant. A mathematical analysis of this issue and calculation of correlations for larger regions still need to be carried out.

Chapter 6

Outro

6.1 Summary

As we have seen, the tree construction invented by Yule tends to surface in theoretical biology whenever neutral Moran-type models are considered. The genealogy of a finite Moran population can be represented by a Yule Tree, and a sample taken from the infinite limit of such a population can be represented by an object of the class of labelled trees, which is topologically equivalent to the class of Yule Trees. The relation of topological equivalence between all tree classes $\mathcal{G}_n$, $\mathcal{L}_n$ and $\mathcal{T}_n$ is founded upon the observations of Theorem 1 and is summarized by the following diagram:

$$\mathcal{G}_n \overset{E}{\longleftrightarrow} \mathcal{L}_n \overset{\text{Lemma } 3}{\longleftrightarrow} \mathcal{T}_n$$

with E denoting the canonical embedding of a coalescent tree into $\mathcal{L}_n$.
Interestingly, $\mathcal{T}_n$ is a finite class of objects, but encodes the same topological complexity the Coalescent does. As Aldous, Steel (e.g. [SM01; Ald00]) and others have demonstrated, taking a step back from the Coalescent Process to a set of discrete objects is a very fruitful way of investigating this process with respect to its graph-theoretical properties. We have concerned ourselves with the question to what extent it is possible to transfer this approach from a "static" setting (i.e., the Coalescent Process in its basic form) to a "dynamic" one, where the dynamics are brought about by either considering the population over time under the action of drift, or by considering the evolutionary history along the chromosome, subject to changes induced by recombination.
Regarding the first one, it turns out that this can be achieved in a very intuitive way if a finite Moran Model in discrete time is assumed. Then, a Markov Chain (the *EMG*) on the set $\mathcal{T}_n$ can be readily defined such that changes in the population are translated into operations on the population genealogy $T_i \in \mathcal{T}_n$, with the consequence that the entire Moran Model becomes a "sub-process" of the *EMG*. The *EMG* itself, then, reveals interesting properties of its own. Most notably, the fact that it is time-reversible allows the construction of the $EMG^\flat$-process, of which we find that we may translate the changes in the genealogy back into the setting of a pure Moran Model: Instead of splitting and killing, which is what happens forward in time, we merge individuals who are duplicates of each other in the order dictated by the tree, and "revive" individuals of the past at the same time. The $EMG^\flat$ facilitates an observation of a couple of features of the tree over time; the *MRCA* process is one example of these, and an attempt to apply the same methods to other nodes than the root uncovers some facts about the average life-time of coalescent events.
The sequence of trees encoded in an *ARG* was already an object of intensive study throughout the early 2000's. The formal description of the prune-regraft operation [EW06] allowed for the treatment of this sequence as a stochastic process, and the

SMC was constructed to provide a Markovian approximation [MC05]. In contrast to the *EMG*, one considers a sample *taken* from a population which is already in the infinite limit, instead of the entire, finite population. We used this setting to analyze the quantity called *Linkage Disequilibrium* and modified it slightly such that it depends only on the tree instead of the mutational process. It turns out that the existing knowledge about the *SMC*, along with some results derived by ourselves, is sufficient to enable a quite exact analysis of this concept, called *topological Linkage Disequilibrium*, as opposed to classical *LD*, where certain unsolved problems prevail; possibly, the presented type of analysis still offers some opportunities to also remedy that.
In the following Section 6.2, we will take a look at how the processes in time and space are unified on one, and set apart on the other hand. For the moment, we conclude that the Yule Process is a crucially intrinsic structure of large parts of neutral population-genetical theory, which causes many seperate aspects of it, which may seem to be far apart from each other, to be beautifully connected instead.

6.2 Cross-Links between time and space

We intend to take a quick look at similarities and differences between the Markov Chains encountered in Chapters 4 and 5. Most of these observations are of a preliminary nature, so this section is best seen as an attempt to point out possibilities of further research, more of which will be mentioned in Section 6.3.

Rates of topological change and Mixing Times Since both *EMG* and the chain $(T_\gamma)_{\gamma\in[0,1]}$, $T \in \mathcal{L}_n$ of labelled trees along the genome are recurrent, one may raise the question how long it takes for both to approximate their stationary distribution (uniform in both cases), and on the other hand, how strong their tendency of remaining in a certain state is with increasing n.
The latter question can be addressed rather quickly. In $(T_\gamma)_{\gamma\in[0,1]}$, one needs to sum up the probabilities of all possibilities of placing the regrafting site on the same branch segment as the pruning site, to find the probability of $T_S = T_S$ for two neighbouring segments S, S'. This can be done independently of the specific tree shape. Magnus Nordborg [Nor00] obtained the value

$$\Pr(\text{healing}) = \frac{2(n-1)}{3na_{n-1}} \approx \frac{2}{3\log(n)}$$

which he called the probability of a recombination event being "healed by coalescence" immediately.
In Section 4.1, we discovered that the probability $\Pr(T_{i+1} = C | T_i = C)$ in the *EMG*, if the current tree is a caterpillar, is $\frac{2n}{n^2} = \frac{2}{n}$. A formal proof that this probability is maximal for all Yule Trees would be requred; however, intuitively it makes sense to assume this for now because the caterpillar is the only tree where the removal of any leaf leads to the same object. Thus, in $(T_\gamma)_{\gamma\in[0,1]}$, we have a "sojourn probability" $p_s^{SMC} := \Pr(\text{healing}) \approx \frac{2}{3\log(n)}$, and at the same time an upper bound for the probability in the *EMG* given by $p_S^{EMG} \leq \frac{2}{n}$. In general, the probability of moving away from the current state is higher in the time-process than in the spatial one (incidentally, except $n < 10$), and converges to 1 faster in terms of magnitude.
This might seem a little bit counterintuitive, because a tree can undergo various changes under the *SMC*, entire subtrees are moved in contrast to the *EMG*, and in

fact, each transformation possible under the *EMG* has at least one corresponding subtree prune-regraft operation that would lead to a topologically equivalent tree under the *SMC*. We hypothesize therefore, that the total mixing time under the *SMC* is smaller than under the *EMG*. However, this will have to be explored in the future. In [Ald00], a mixing time for the *Aldous Chain* on cladograms is derived; quite possibly, by a similar logic the mixing time of the *EMG* becomes accessible. Deriving the mixing time of the *SMC* probably requires to consider the *rooted Subtree Prune-Regraft* problem, which is to determine the minimal number $n(L_1, L_2)$ of prune-regraft operations necessary to transform the labelled tree L_1 into another L_2 and is known to be *NP*-hard [BS05]. A suitable upper or lower bound for the average over all pairs of trees might help to resolve this issue.

The resilience of imbalance The process of tree balance (see Section 4.1), as a subprocess of the *EMG*, follows the dynamics of a Wright-Fisher diffusion. As such, its volatility is reduced if $|T_i^l|/n$ is either close to 0 or 1. In $(T_\gamma)_{\gamma\in[0,1]}$, a similar phenomenon can be observed.
Recall the probability of a switching event

$$\Pr(\text{switch}) = 2\sum_{k=2}^{n}\sum_{j=1}^{k-1}\frac{k-j}{k}\frac{(k-1)(j+1)}{k(k+1)}\frac{1}{k-1}\frac{1}{(k-1)a_{n-1}}$$

from our derivation of Lemma 8. If we condition this formula on, say, trees which have only $j = 1$ branch on the right side (and are thus highly unbalanced), we obtain

$$\Pr(\text{switch}|j=1) = 2\sum_{k=2}^{n}\frac{k-1}{k}\frac{(k-1)(2)}{k(k+1)}\frac{1}{k-1}\frac{1}{(k-1)a_{n-1}} \in \mathcal{O}(\frac{1}{\log(n)})$$

This accounts for the probability of moving a subtree from the "large" side of the tree to the single-branch side. The probability of shifting the other way is of order $\mathcal{O}(\frac{1}{n})$, and the probability of an immediate root change remains at $\mathcal{O}(\frac{1}{\log(n)})$. Thus, the overall chance of escaping the unbalanced state tends to 0 with increasing n.
On the other hand, it turns out that

$$\Pr(\text{switch}|j=k/2) \overset{n\to\infty}{\longrightarrow} \frac{1}{2}$$

if the tree is balanced ($j = k/2$). Of course, this is not the exact probability of changing the balance of the tree, but it is suggested that the probability of escaping unbalanced states is lower than balanced states in the *SMC*. This also agrees with the results in [FDW13], which lead to a similar conclusion. In general, tree balance seems to obey similar rules in time and space. Once the process enters an unbalanced state, it might take some time until this state is left. This fact might be of relevance in the analysis of population-genetical data, where unbalanced tree topology (which may happen purely by chance, since tree balance is uniform under neutrality; see Proposition 2) may have disruptive effects on statistics like Tajima's D, or T_3 [LW13].

6.3 Outlook

A couple of open problems that persist in the framework of Coalescent Theory in combination with the Yule Process have already been encountered througout this thesis. With regards to the application of the *tLD* measure, some additional work is

needed to quantify advantages, disadvantages and sources of errors in comparison to classical *LD*. An obvious point of interest is the determination of a correlation between *tLD* and pooled classical *LD*. This can be done considering an optimal setting with respect to knowledge about tree topology, but also in a setting assuming that topologies need to be estimated.
Following up on this, we suspect possible room for improvement of the method at the stage of tree estimation. We applied a *2-means* clustering approach based on Hamming distance; but since on a Coalescent Tree, some mutations may be more informative than others, making use of other distance measures, or a maximum-likelihood approach, might lead to more reliable results. It should be stressed that the method that was used still tended to outperform traditional methods of phylogeny (such as those provided by `RAxML` [Sta14]), which aim to determine the entire tree topology; an effect which is likely due to sparsity of data and uncertainty about ancestral and derived states.
From a graph-theoretical perspective, the problems mentioned in the previous section might provide interesting research opportunities. In particular, the prune-regraft operation is linked to an *NP*-hard computational problem. Knowledge about the *expected* number of operations to transform one tree into another should offer a way of deriving bounds for the mixing time of the *SMC*, connecting the computational aspects of the process with the stochastic ones. It is noteworthy in this regard that a "mixing"-probability of the *ARG* on an infinitely long chromosome has been proven [DPP15].
Finally, we want to point out some possibilities of taking this approach to a non-neutral setting. With respect to natural selection, many contributions to the literature have been made in the recent past. The *Ancestral Selection Graph* [KN97] is a prominent example, providing an ancestral process for a sample of an infinite population in equilibrium between natural selection, favouring some allele a over another A, and mutation between a and A alleles. A graphical representation [Len+15; BCH18] can be constructed that allows access to many implicit features of the model. However, the ancestral process in the case of a selective sweep is nontrivial; popular approaches to analyze this involve, e.g., the Lookdown-Construction; more recently, a promising method of *potential ancestry* [GS18] was proposed.
EMG-like constructions might help in this regard by looking at the problem from the perspective of a finite population. The biggest issue that needs to be overcome in this regard is the necessity to define the process "conditioned on fixation", a fact we have glossed over in Section 4.3. However, in the neutral case this problem can be worked around, because a root node vanishing under the $EMG^\flat$ always corresponds to a fixation (namely, of the descendants of the *MRCA* this root represents) in the *EMG* forward in time. For cases involving selection, considering a conditioned process seems unavoidable.
We discuss shortly how this can be achieved, adding the caveat that this is largely work in progress. In a finite two-allele (a, A) Moran Model, the transition probabilities for the frequency $f(a)$ of allele a are given by the following matrix:

$$T_{f(a)} := \begin{array}{c} \begin{array}{ccccccc} 0 & \dots & k/n & & \dots & 1 & \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ 0 & & \vdots & & & & \\ \vdots & & 0 & & & & \\ & & \frac{k(n-k)}{n^2} & & & & \\ & & \frac{k^2+(n-k)^2}{n^2} & & & & \\ & & \frac{k(n-k)}{n^2} & & & & \\ & & 0 & & & & \\ & & \vdots & & & \vdots & \\ & & & & & 0 & \\ & & & & 0 & 1 & \end{pmatrix} \end{array} \begin{array}{c} 0 \\ \vdots \\ \\ \\ \\ \\ k/n \\ \\ \\ \vdots \\ \\ 1 \end{array}$$

If a selection coefficient $s > 0$ is incorporated, this changes to

$$T_{f(a),s} := \begin{array}{c} \begin{array}{ccccccc} 0 & \dots & k/n & & \dots & 1 & \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ 0 & & \vdots & & & & \\ \vdots & & 0 & & & & \\ & & \frac{k(n-k)}{n(n+ks)} & & & & \\ & & \frac{k^2+(n-k)^2}{n(n+ks)} & & & & \\ & & \frac{k(1+s)(n-k)}{n(n+ks)} & & & & \\ & & 0 & & & & \\ & & \vdots & & & \vdots & \\ & & & & & 0 & \\ & & & & 0 & 1 & \end{pmatrix} \end{array} \begin{array}{c} 0 \\ \vdots \\ \\ \\ \\ \\ k/n \\ \\ \\ \vdots \\ \\ 1 \end{array}$$

The states 0 and n in these chains are absorbing, and there exists an *exit law* at these states. Suppose the *entry law* is one at $f(a) = 1/n$, meaning a single allele of type a exists at time 0. The results of [HP86] then allow us to reverse time in these chains with an entry law of one at 1 and an absorbing state $*$ (birth of the a-allele) replacing 0. For the neutral case, the transition matrix is given by

$$T_{f(a)^\flat} := \begin{array}{c} \begin{array}{ccccccc} * & \dots & k/n & & \dots & 1 & \end{array} \\ \begin{pmatrix} 1 & & & & & & * \\ 0 & & \vdots & & & & \vdots \\ \vdots & & 0 & & & & \\ & & \frac{(k+1)(n-k)}{n^2} & & & & \\ & & \frac{k^2+(n-k)^2}{n^2} & & & & k/n \\ & & \frac{(k-1)(n-k)}{n^2} & & & & \\ & & 0 & & & \vdots & \\ & & \vdots & & & 0 & \vdots \\ & & & & & 1 & \\ & & & & 0 & 0 & 1 \end{pmatrix} \end{array}$$

A similar matrix can be calculated for the selected case, which reflects the process backward in time after a selective sweep. Its entries, however, become quite complicated quickly and perhaps cannot be determined for general n. In any case, it is possible to define a process similar to the $EMG^\flat$ on labelled trees whose leaves are additionally classified into type a- and A-leaves, and whose transition probabilities respect the transition probabilities of the reversed chain $T_{f(a)^\flat}$. The same is possible in the case of $s > 0$. In particular, regrafting is not uniform over branch segments any more. Figure 6.1 outlines this process for $n = 4$ and under neutrality.

One benefit of performing these technically intense preparations is that in the reversed tree processes, one is allowed to consider again the *MRCA* process, but this time conditioned on the fixation of a beneficial allele. The way we have proven Lemma 6 (i.e., using Lemma 2) strongly suggests that the number of *MRCA* jumps is still expected to be 2 in the presence of selection. With the conditional process properly defined, we may look for a general proof.

Assuming this is possible, by such a time reversal one can assess the degree to which the genealogy deviates from the usual Yule distribution at intermediate stages of the selective sweep. Such a construction may therefore yield interesting insights on a process which is still rather mysterious. At last, let us propose a hypothesis with a practical consequence: Imagine a finite Moran Model with recombination, and a beneficial allele appears at one locus and becomes fixed. The selective fixation process is much faster than the neutral one, but the number of root jumps is always just around two regardless of selection. Then, moving away from the selected locus and observing the sequence of genealogies provided by the ancestral process with recombination, to both sides one would expect to observe around two prune-regraft operations that cause a root change, accounting for the around two *MRCA* jumps at the selected locus, in quick succession before neutrality governs the ancestral process again.

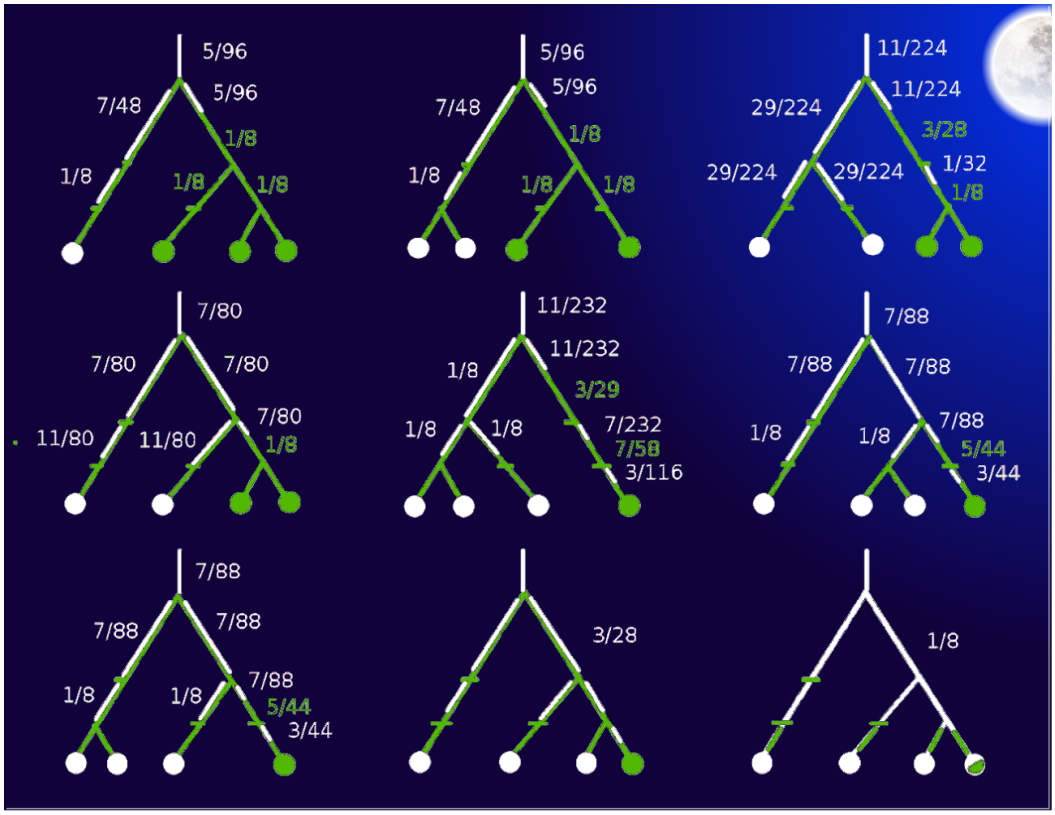


FIGURE 6.1: The fixation process in a Moran Model of size 4, backward in time, realized on the genealogy. The a-allele, represented by a black leaf, got fixed forward in time and will, backward in time, slowly vanish, until the birth of its first copy. On each branch segment, regrafting of a branch leading to a leaf of a particular allelic type is given by the number with the according color at that segment. Each tip has an additional $1/16$ probability for a leaf (to immediately vanish again, compare procedure 9) to be regrafted there.
For simplicity's sake, we have omitted trees which can be obtained from one of the depicted trees by exchanging subtrees below internal nodes. The probability distribution on the branches would remain the same after an exchange of subtrees anyways.
The system of equations neccessary to determine the probabilities of this tree process is huge ($\approx$ 33 independent variables for $n = 4$). To make this method practical, the complexity will have to be reduced.

Bibliography

[Ald00] D. J. Aldous. "Mixing time for a markov chain on cladograms". In: *Combinatiorics, Probability and Computing* 9.3 (2000), pp. 191–204.

[Ald01] D. J. Aldous. "Stochastic models and descriptive statistics for phylogenetic trees, from yule to today". In: *Statistical Science* 16.1 (2001), pp. 23–34. ISSN: 08834237.

[AP96] D. Aldous and R. Pemantle. *Random Discrete Structures*. Ed. by Springer IMA Volumes Math. Appl. 76. Berlin Heidelberg, 1996.

[Aut+15] A. Auton et al. "A global reference for human genetic variation". In: *Nature* 526 (2015), pp. 68–74. DOI: 10.1038/nature15393.

[BBS07] J. Berestycki, N. Berestycki, and J. Schweinsberg. "Beta-coalescents and continuous stable random trees". In: *Ann. Probab.* 35.5 (Sept. 2007), pp. 18-35-1887. DOI: 10.1214/009117906000001114.

[BCH18] E. Baake, F. Cordero, and S. Hummel. "A probabilistic view on the deterministic mutation–selection equation: dynamics, equilibria, and ancestry via individual lines of descent". In: *Journal of Mathematical Biology* (2018). ISSN: 1432-1416. DOI: 10.1007/s00285-018-1228-8.

[Ber+04] T. Bersaglieri et al. "Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene". In: *The American Journal of Human Genetics* 74.6 (2004), pp. 1111–1120. ISSN: 0002-9297. DOI: 10.1086/421051.

[Bio18] National Center for Biotechnology Information. *Database*. 2018. URL: https://www.ncbi.nlm.nih.gov/ (visited on 11/06/2018).

[BL07] S. Boitard and P. Loisel. "Probability distribution of haplotype frequencies under the two-locus Wright–Fisher model by diffusion approximation". In: *Theoretical Population Biology* 71.3 (2007), pp. 380–391. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2006.12.007.

[Bru80] F. T. Bruss. "A counterpart of the Borel-Cantelli lemma". In: *Journal of Applied Probability* 17.4 (1980), pp. 1094–1101. DOI: 10.2307/3213220.

[BS05] M. Bordewich and C. Semple. "On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance". In: *Annals of Combinatorics* 8.4 (Jan. 2005), pp. 409–423. ISSN: 0219-3094. DOI: 10.1007/s00026-004-0229-z.

[Cav73] L. L. Cavalli-Sforza. "Analytic review: some current problems of human population genetics." In: *American journal of human genetics* 25.1 (1973), pp. 82–104. DOI: https://doi.org/.

[CK70] J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Ed. by Blackburn Press. New Jersey, 1970.

[Dar59] C. Darwin. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London, UK: J. Murray, 1859.

[DK96] P. Donnelly and T. G. Kurtz. "A countable representation of the fleming-viot measure-valued diffusion". In: *Ann. Probab.* 24.2 (1996), pp. 698–742. DOI: 10.1214/aop/1039639359.

[DPP15] A. Depperschmidt, E. Pardoux, and P. Pfaffelhuber. "A mixing tree-valued process arising under neutral evolution with recombination". In: *Electronic Journal of Probability* 20 (2015), 22 pp. DOI: 10.1214/EJP.v20-4286.

[Ena+02] N. S. Enattah et al. "Identification of a variant associated with adult-type hypolactasia." In: *Nature Genetics* 30.2 (2002), p. 233. DOI: 10.1038/ng826.

[Eth11] A. Etheridge. *Some Mathematical Models from Population Genetics*. 2011.

[EW06] S. N. Evans and A. Winter. "Subtree prune and regraft: A reversible real tree-valued Markov process". In: *Ann. Probab.* 34.3 (May 2006), pp. 918–961. DOI: 10.1214/009117906000000034.

[FDW13] L. Ferretti, F. Disanto, and T. Wiehe. "The Effect of Single Recombination Events on Coalescent Tree Height and Shape". In: *PLOS ONE* 8.4 (Apr. 2013), pp. 1–15. DOI: 10.1371/journal.pone.0060123.

[FL93] Y. X. Fu and W. H. Li. "Statistical tests of neutrality of mutations." In: *Genetics* 133.3 (1993), pp. 693–709. ISSN: 0016-6731. eprint: http://www.genetics.org/content/133/3/693.full.pdf.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. 1st ed. New York, NY, USA: Cambridge University Press, 2009. ISBN: 05218980-64, 9780521898065.

[Fu95] Y. X. Fu. "Statistical Properties of Segregating Sites". In: *Theoretical Population Biology* 48.2 (1995), pp. 172–197. ISSN: 0040-5809. DOI: 10.1006/tpbi.1995.1025.

[FW00] J. C. Fay and C. I. Wu. "Hitchhiking under positive Darwinian selection." In: *Genetics* 155.3 (2000), pp. 1405–1413.

[Gil98] J. Gillespie. *Population Genetics: A concise guide.* Baltimode, Md: John Hopkins, 1998.

[GS18] A. González Casanova and D. Spanò. "Duality and fixation in Ξ-Wright-Fisher processes with frequency-dependent selection". In: *Ann. Appl. Probab.* 28.1 (Feb. 2018), pp. 250–284. DOI: 10.1214/17-AAP1305.

[GVL83] J. Greilhuber, M. Volleth, and J. Loidl. "Genome size of man and animals relative to the plant Allium cepa". In: *Canadian Journal of Genetics and Cytology* 25.6 (1983), pp. 554–560. DOI: 10.1139/g83-084.

[Hal40] J. B. S. Haldane. "The Mean and Variance of x2, When Used as a Test of Homogeneity, When Expectations are Small". In: *Biometrika* 31.3/4 (1940), pp. 346–355. ISSN: 00063444. URL: http://www.jstor.org/stable/2332614.

[Har08] G. H. Hardy. "Mendelian proportions in a mixed population". In: *Science* 28.706 (1908), pp. 49–50. ISSN: 0036-8075. DOI: 10.1126/science.28.706.49. eprint: http://science.sciencemag.org/content/28/706/49.full.pdf.

[HE10] A. Hodgkinson and A. Eyre-Walker. "Human triallelic sites: evidence for a new mutational mechanism?." In: *Genetics* 184.1 (2010), pp. 233–241. DOI: 10.1534/genetics.109.110510.

[Hof16] A. van't Hof et al. "The industrial melanism mutation in British peppered moths is a transposable element". In: *Nature* 534 (2016), pp. 102–105. DOI: 10.1038/nature17951.

[HP86] U. G. Haussmann and E. Pardoux. "Time Reversal of Diffusions". In: *Ann. Probab.* 14.4 (Oct. 1986), pp. 1188–1205. DOI: 10.1214/aop/1176992-362.

[Hud02] R. R. Hudson. "Generating samples under a Wright–Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002), pp. 337–338. DOI: 10.1093/bioinformatics/18.2.337. eprint: /oup/backfile/content_public/journal/bioinformatics/18/2/10.1093/bioinformatics/18.2.337/2/180337.pdf.

[Hud83] R. R. Hudson. "Properties of a neutral allele model with intragenic recombination". In: *Theoretical Population Biology* 23.2 (1983), pp. 183–201. ISSN: 0040-5809. DOI: 10.1016/0040-5809(83)90013-8.

[Hux42] J. Huxley. *Evolution: The Modern Synthesis*. Cambridge, Massachusettes: MIT Press, 1942.

[Jab+18] K. Jabbari et al. "Interdependence of linkage disequilibrium, chromatin architecture and compositional genome organization of mammals". In: *bioRxiv* (2018). DOI: 10.1101/293837. eprint: https://www.biorxiv.org/content/early/2018/04/04/293837.full.pdf.

[Ket58] H. B. D. Kettlewell. "A survey of the frequencies of Biston betularia (L.) (Lep.) and its melanic forms in Great Britain". In: *Heredity* 12 (1958), pp. 51–72. DOI: 10.1038/hdy.1958.4.

[KF18] A. Klassmann and L. Ferretti. "The third moments of the site frequency spectrum". In: *Theoretical Population Biology* 120 (2018), pp. 16–28. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2017.12.002.

[Kim83] M. Kimura. *The neutral theory of molecular evolution*. Ed. by Cambridge University Press. New York, 1983.

[Kin82] J. F. C. Kingman. "On the genealogy of large populations". In: *Journal of Applied Probability* 19(A) (1982), pp. 27–43. DOI: 10.2307/3213548.

[KN97] S. M. Krone and C. Neuhauser. "Ancestral processes with selection". In: *Theoretical Population Biology* 51.3 (1997), pp. 210–237. DOI: 10.1006/tpbi.1997.1299.

[Kön+09] W. König et al. "A two cities theorem for the parabolic Anderson model". In: *Ann. Probab.* 37.1 (Jan. 2009), pp. 347–392. DOI: 10.1214/08-AOP405.

[Kor+13] T. Korneliussen et al. "Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data". In: *BMC Bioinformatics* 14 (2013), p. 289. DOI: 10.1186/1471-2105-14-289.

[Len+15] U. Lenz et al. "Looking down in the ancestral selection graph: A probabilistic approach to the common ancestor type distribution". In: *Theoretical Population Biology* 103 (2015), pp. 27–37. DOI: 10.1016/j.tpb.2015.01.005.

[Lew64] R. C. Lewontin. "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models". In: *Genetics* 49.1 (1964), pp. 49–67.

[LMW18] W. Löhr, L. Mytnik, and A. Winter. "The Aldous chain on cladograms in the diffusion limit". In: *ArXiv e-prints* (2018).

[LW13] H. Li and T. Wiehe. "Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation". In: *PLOS Computational Biology* 9.5 (May 2013), pp. 1–14. DOI: 10.1371/journal.pcbi.1003060.

[LW98] L. Lovász and P. Winkler. "Reversal of markov chains and the forget time". In: *Combinatorics, Probability and Computing* 7.2 (1998), pp. 189–204.

[MC05] G. A. McVean and N. J. Cardin. "Approximating the coalescent with recombination." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1459 (2005), pp. 1387–1393.

[McA+18] A. McAvoy et al. "Public goods games in populations with fluctuating size". In: *Theoretical Population Biology* 121 (2018), pp. 72–84.

[MK91] J. H. McDonald and M. Kreitman. "Adaptive protein evolution at the Adh locus in Drosophila". In: *Nature* 351.652 (1991). DOI: 10.1038/351652a0.

[Mor58] P. A. P. Moran. "Random processes in genetics". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54.1 (1958), pp. 60–71. DOI: 10.1017/S0305004100033193.

[Mur84] F. Murtagh. "Counting dendrograms: A survey". In: *Discrete Applied Mathematics* 7.2 (1984), pp. 191–199. ISSN: 0166-218X. DOI: 10.1016/0166-218X(84)90066-0.

[Nor00] M. Nordborg. "Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization". In: *Genetics* 154.2 (2000), pp. 923–929. ISSN: 0016-6731. eprint: http://www.genetics.org/content/154/2/923.full.pdf.

[NT15] R. A. Neher and Bedford T. *nextflu: real-time tracking of seasonal influenza virus evolution in humans.* 2015. DOI: 10.1093/bioinformatics/btv381.

[Obe+13] U. Ober et al. "The expected linkage disequilibrium in finite populations revisited". In: *ArXiv E-Prints* (2013).

[PFL10] J. Pellicer, M. F. Fay, and I. J. Leitch. "The largest eukaryotic genome of them all?" In: *Botanical Journal of the Linnean Society* 164.1 (2010), pp. 10–15. DOI: 10.1111/j.1095-8339.2010.01072.x. eprint: /oup/backfile/content_public/journal/botlinnean/164/1/10.1111_j.1095-8339.2010.01072.x/2/j.1095-8339.2010.01072.x.pdf.

[PW06] P. Pfaffelhuber and A. Wakolbinger. "The process of most recent common ancestors in an evolving coalescent". In: *Stochastic Processes and their Applications* 116.12 (2006), pp. 1836–1859. DOI: 10.1016/j.spa.2006.04.015.

[PWW09] P. Pfaffelhuber, A. Wakolbinger, and H. Weisshaupt. "The tree length of an evolving coalescent". In: *ArXiv E-Prints* (2009).

[Rau18] M. Rauscher. "Topology of genealogical trees: Theory and application." PhD thesis. University of Cologne, 2018.

[SH74] J. M. Smith and J. Haigh. "The hitch-hiking effect of a favourable gene". In: *Genetical Research* 23.1 (1974), pp. 23–35. DOI: 10.1017/S0016672300014634.

[SM01] M. Steel and A. McKenzie. "Properties of phylogenetic trees generated by yule-type speciation models". In: *Mathematical Biosciences* 170.1 (2001), pp. 91–112. DOI: 10.1016/S0025-5564(00)00061-4.

[Son06] Y. S. Song. "Properties of Subtree-Prune-and-Regraft Operations on Totally-Ordered Phylogenetic Trees". In: *Annals of Combinatorics* 10.1 (June 2006), pp. 147–163. ISSN: 0219-3094. DOI: 10.1007/s00026-006-0279-5.

[Sta14] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (2014), pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033. eprint: /oup/backfile/content_public/journal/bioinformatics/30/9/10.1093_bioinformatics_btu033/3/btu033.pdf.

[Sve71] J. A. Sved. "Linkage disequilibrium and homozygosity of chromosome segments in finite populations". In: *Theoretical Population Biology* 2.2 (1971), pp. 125–141. ISSN: 0040-5809. DOI: 10.1016/0040-5809(71)90011-6.

[Taj89] F. Tajima. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." In: *Genetics* 123.3 (1989), pp. 585–595. ISSN: 0016-6731. eprint: http://www.genetics.org/content/123/3/585.full.pdf.

[Voi+06] B. F. Voight et al. "A Map of Recent Positive Selection in the Human Genome". In: *PLOS Biology* 4.3 (Mar. 2006). DOI: 10.1371/journal.pbio.0040072.

[Wak] J. Wakeley. *Coalescent Theory, an Introduction*.

[Wei08] W. Weinberg. "Über den Nachweis der Vererbung beim Menschen". In: *Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg* 64 (1908), pp. 369–382. DOI: 10.1017/S00166723000-14634.

[WRW18] J. Wirtz, M. Rauscher, and T. Wiehe. "Topological linkage disequilibrium calculated from coalescent genealogies". In: *Theoretical Population Biology* (2018). ISSN: 0040-5809. DOI: 10.1016/j.tpb.2018.09.001.

[WW18] J. Wirtz and T. Wiehe. "The evolving Moran Genealogy". In: *ArXiv E-Prints* (2018).

[Yul25] G. U. Yule. "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S." In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 213.402-410 (1925), pp. 21–87. DOI: 10.1098/rstb.1925.0002.

Epilogue

During the last days of writing this thesis, it somehow occurred to me that I might like to write something personal, perhaps at the end, without knowing exactly what it would be about. I asked my supervisor whether this would be ok or completely out of place; after considering it for some seconds, he said, "Yeah, go ahead and write an epilogue or something", remarking that this work was mine, MY book, and I could write whatever I wanted. If you don't mind, I'll do that in German. So here goes...

"Würde Wälder um Neersen pflanzen"

Das war's. Die letzten Feinschliffe sind getätigt. Figuren, Tabellen und dergleichen sind so gut an ihren Platz gerückt, wie es in LaTeX nur möglich ist. Fehlerhafte Zeilenumbrüche sind korrigiert, selbst im Literaturverzeichnis. Nach fast fünf Jahren, teils unter Bedingungen, die Uli wohl als "high difficulty" bezeichnen würde,[1] dazu zählt u. A. ein vor zwei Monaten gestohlener Laptop incl. 20 bereits geschriebener Seiten der Dissertation, die nicht gesichert waren, ist morgen Abgabetermin, und ich glaube, sogar den bürokratischen Teil geregelt zu haben, etwas, womit ich mir sonst gerne ein Bein stelle; selbst eine kleine Ungereimtheit mit "Docfile" wird mir wahrscheinlich nicht ernsthaft in die Quere kommen. Ich sollte mich wohl zufrieden zurücklehnen, und nicht genau jetzt, wo ich noch unter dem Eindruck der ganzen Ackerei stehe versuchen, die letzte Zeit einzuordnen, und vor allem nicht, etwaige prosaische/epische/lyrische/ Talente zu reaktivieren. Ich will es trotzdem versuchen, hauptsächlich aus zwei Gründen.
Zum einen wurde mir oft nachgesagt, ich könne über mathematische Sachverhalte, wie z. B. beim Zusammenstellen eines Seminarvortrags, einer Hausarbeit oder auch einer Abschlussarbeit, nahezu druckreif schreiben. Tatsächlich, das Schreiben über Mathematik hat mir nie große Probleme bereitet. Mir kam es immer vor, als gäbe es einen optimalen Weg, einen Sachverhalt in Worte zu fassen, ähnlich der Erdös'schen Auffassung vom "Book of Proofs", und darüber hinaus, als würde ich durch das Schreiben selbst zu diesem Optimalweg hin-konvergieren. Ich hatte geradezu Spaß beim Verfassen meiner Bachelor- und Masterarbeiten, war bereits lange vor dem Termin fertig und besserte hier, schnippelte da, kürzte, änderte Definitionen und konnte die Arbeit in einem Zustand übergeben, bei dem ich mir sicher war, jeden Buchstaben und jedes Symbol zweimal umgedreht zu haben. Hier, bei der Dissertation, ist dies weitestgehend unmöglich aus verschiedenen Gründen, der naheliegendste ist natürlich der Zeitdruck. Was ich geschrieben habe wird genügen müssen wie es ist, "frei schnauze", und von daher ist es passend, sich am Schluss noch ein wenig frei zu äußern. Aber abgesehen davon, habe ich zu der Dissertation bisher nicht so eine Beziehung entwickelt wie beispielsweise zu meiner Masterarbeit seinerzeit. Natürlich, ich scrolle über das PDF, ich bin überzeugt von meinen Ausführungen, ich finde sie teilweise richtig schön, ich glaube, der Text hat einen roten Faden und kann den

[1] Oder auch: "Note 1 schwer"

Leser überzeugen, vielleicht sogar in den Bann schlagen. Aber zu solch einer Vernarrtheit wie zu Bachelor- oder Masterzeiten habe ich nicht gefunden. Während des Schreibens schwankte meine Laune zwischen "Jawohl, wieder ein Absatz erledigt", und "Ich kann es nicht mehr sehen". Warum das so ist, kann ich mir nicht ganz erklären, schließlich täte ich das alles nicht wenn ich nicht der Meinung wäre, dass die Ergebnisse es wert sind eine Dissertation darüber zu schreiben. Aber bei Fragen bezüglich der Gemütsverfassung ist meiner Ansicht nach meist die einfachste Antwort die richtige, und die lautet nunmal, dass mir eine Pause wohl ganz gut täte. Denn diese Arbeit zu einem Ende zu führen ist mir nicht leichtgefallen.
Es ist mir nicht leichtgefallen, weil ich mich in die Biologie hineindenken musste, und dieses Fach studieren andere nicht grundlos 10 Semester lang bis zum Masterabschluss. Es ist es nicht, weil "Coalescent Theory" ein sehr weit verzweigtes, aber auch viel bearbeitetes Feld ist, und man sehr tief in die Materie eintauchen muss, bis man seine Nische findet und den Teil des Ganzen, an dem man "gerne" arbeiten möchte. Es ist auch aus persönlicher Sicht nicht leichtgefallen. Meine Promotionszeit kam mir privat oft anstrengend vor – ohne dabei ins Detail gehen zu wollen. Der Fachbereich tut wohl sein übriges dazu; bist du immer mit den Gedanken anderswo, insbesondere bei komplizierter mathematischer Theorie, verbeißt dich vielleicht darin, entwickelst gar eine Selbstgefälligkeit damit, findest du dich recht plötzlich alleine wieder, und es mag dir so vorkommen, als laufe das Leben gänzlich an dir vorbei. Man kann sich in der Mathematik leicht eine Art Elfenbeinturm zurechtzimmern, und sich von den Menschen um sich herum entfernen, und damit wohl auch von sich selbst.
Zeichnen wir kein allzu düsteres Bild und lassen diese Dinge ruhen. Dennoch: In Zukunft würde ich mich gerne bemühen, aus dem Turm auszubrechen und wieder den Kontakt herzustellen zu den Dingen, die mich umgeben und die mir neben der Wissenschaft noch am Herzen liegen. Allein, das ist ein hehrer Wunsch, denn die Wissenschaft ist wie die Wirtschaft, bloß schlimmer; Abstand zu nehmen bedeutet, das Tagesgeschäft aus dem Blick zu verlieren. Ich habe bereits Bewerbungen geschrieben, um direkt nach meiner Zeit hier in Köln weitermachen zu können, mich wieder in neues einzuarbeiten, ich bemerke, dass ich Artikel und Veröffentlichungen plane, dass ich Kontakte knüpfe um später Projekte starten zu können, dass ich in meinen Manuskripten bewusst Fragen offen lasse um Material zu haben für das nächste, und das darauf, und das über-über-übernächste... bis wann, weiß keiner. Viel Zeit lässt das augenscheinlich nicht, nochmal innezuhalten und sich zu fragen, was einem noch wichtig ist.
Im Jahre 1996, vermutlich, erschien im Lokalteil "Willich, Kreis Viersen, Tönisvorst" der Rheinischen Post ein Kommentar, in dem die örtliche Städteplanungs- und Umweltpolitik schwer kritisiert wurde. Die maue Begrünung, zunehmende Asphaltierung, Zerstörung natürlicher Lebensräume und der fahrlässige Umgang mit der Gesundheit der Bürger im Zusammenhang mit der Verstädtlichung wurden aufs Schärfste attackiert und ein Wechsel in den relevanten Führungspositionen zumindest nahegelegt, verbunden mit der Drohung, man werde sich selbst für das Bürgermeisteramt stark machen, sollte keine Abhilfe geschaffen werden. Der Autor: Ein gewisser Johannes Wirtz, 7 Jahre alt, Grundschüler.

AKTION Kinder planen

‚Würde Wälder um Neersen pflanzen'

KREIS VIERSEN. Wie planen Kinder ihre Heimatstadt, welche Wünsche haben sie an die Erwachsenen?, hatte die Rheinische Post am vergangenen Samstag gefragt. Der siebenjährige Johannes Wirtz aus Neersen, Am Roth 12, der die zweite Schulklasse besucht, hat sich dazu seine Gedanken gemacht und sie aufgeschrieben:

„Ich heiße Johannes Wirtz und wohne in Neersen. Es ist ein Stadtteil von Willich. Mir gefällt es überhaupt nicht mehr in dieser Stadt. Es wird zuviel gebaut. Straßen, Häuser. Doch keiner pflanzt Bäume. Nur kleines Gestrüpp. Wenn ich einmal Bürgermeister bin, pflanze ich um ganz Neersen Wälder. Birken, Buchen, Eichen, Akazien, Fichten, Zedern. Dann heißt mein Stadtteil Wald Neersen (Neersener Wald)."

Wer wie der siebenjährige Johannes Anregungen für seinen Heimatort hat, Ideen für Verbesserungen mitteilen will oder auch nur über etwas meckern möchte, was ihn ärgert, soll uns schreiben. Aufgerufen sind unsere ganz jungen Leser bis etwa 14 Jahre. Auch Schulklassen können bei unserer Aktion mitmachen. Zeichnungen sind willkommen.

Die Adresse lautet: **Redaktion Rheinische Post, Enger-**

Ich muss meine Aussagen von damals an dieser Stelle definitiv revidieren; ich habe keine Ambitionen, den lokalen Politikern (die man teilweise kennt und schätzt) in die Parade zu fahren, abgesehen davon ist Neersen und auch Willich wie der gesamte Niederrhein für eine Kulturlandschaft sehr grün und natürlich. Überhaupt habe ich dieses Dokument lange für eine dieser Peinlichkeiten gehalten, die man als Kind fabriziert, die einem als Jugendlichen bei der Erinnerung daran die Haare zu Berge stehen lassen (und das sage ich als jemand, der mit zugehaltenen Ohren schreiend in die nächste Etage rennt, sobald irgendeines meiner "Werke" als Kind zitiert wird oder Musik von mir gespielt wird oder irgendetwas, was ich früher mal gerne gemacht habe, ans Tageslicht gezerrt wird; Stichwort: Sich von sich selbst entfernen) und zu denen man als sozusagen Erwachsener keine rechte Bindung mehr hat. Es verwundert mich tatsächlich ein wenig, dass ich diesen Artikel jetzt hier mit einem Schmunzeln lesen kann. Mehr noch, erstaunt stelle ich fest, dass sich ein Kreis schließt; denn mein Promotionsfach heißt Computational Bio-logy. Kindern liegen neben vielen anderen Dingen oft die Natur am Herzen, es ist etwas magisches für sie. Wie es scheint, habe ich das nicht ganz aufgegeben, bin stattdessen seit Jahren wieder eher auf dem Weg dorthin zurück; immerhin bin ich heute tatsächlich eine Art Biologe. Und vor diesem Hintergrund macht es Sinn, dass ich mich wirklich über eine abgelehnte Bewerbung ärgere zu einer Stelle, bei der es hauptsächlich um Sequenzieren von Tannenbaum-DNA gegangen wäre, und dass ich mich tatsächlich von this.ven (eigentlich noch so einer für die Acknowledgements) zum Hambacher Forst zwecks Demonstration und Musizieren habe schleppen lassen (bleibt er?).

Dieser Text scheint nun langsam wirklich eine Art Sinn zu haben. Möglicherweise zeigen diese Überlegungen, dass das Vergangene manchmal in kurioser Art und Weise präsent bleibt, auch wenn man es zeitweise nicht wahrnimmt. Somit ist man frei, wieder aus dem Kreis herauszutreten, und sich mit der Fackel ins Dunkel aufzumachen.

Eine Professorin sagte einmal zu mir, Mathematik sei das Tasten in einem dunklen, engen Raum voll von umherliegenden Dingen, an denen man sich stoße, bis man den Lichtschalter finde und feststelle, alles sei so angeordnet wie es "richtig" ist. Das passt ins Bild. Und auch dazu fällt mir ein Satz aus irgendeinem Lehrbuch von früher ein, nämlich, dass es, ähnlich wie das Fliegen oder die Vergangenheit zu ändern ein menschlicher Wunsch ist, Licht in die Dunkelheit zu bringen. In der dunklen Jahreszeit, die mittlerweile unübersehbar angebrochen ist, kann man beobachten, dass dies vielen wichtig ist; man schaue nur in ein beliebiges Fenster. So ein schlechter Vorsatz für die Zukunft, denke ich, wird dies also nicht sein.

In den Acknowledgements habe ich einigen, die mich die letzten Jahre begleitet haben, schon ausgiebig gedankt. Zu guter Letzt, noch ein Wort an Sie, den unbeschwerten Leser. Wenn sie diese Arbeit von vorn bis hier komplett gelesen haben, kann ich nur sagen, dass es mir viel bedeutet, und ich hoffe es ist etwas für Sie darin

gewesen. Recht herzlichen Dank, und lott jonn.

JMW
geschrieben am Martinstag, November 2018.

Eidesstattliche Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit, einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie, abgesehen von unten angegebenen Teilpublikationen, noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Thomas Wiehe betreut worden.

Für die Arbeit relevante Teilpublikationen:

- *Topological Linkage Disequilibrium calculated from coalescent genealogies*; Rauscher, Wiehe, W., in: *Theoretical Population Biology*; 2018. [WRW18]
- *The Evolving Moran Genealogy*; Wiehe, W., in *ArXiv E-Prints*; 2018. [WW18]

In beiden Fällen war ich verantwortlich für die Herleitung des Großteils der mathematischen Theorie. In "Topological Linkage Disequilibrium" wurden Programmierung und Illustrierung hauptsächlich von Martina Rauscher übernommen; sofern Ergebnisse dieser Art in der vorliegenden Dissertation erwähnt und verwendet werden, ist dies an entsprechender Stelle kenntlich gemacht.

Unterzeichnet:

Datum:

Lebenslauf

Name:	Johannes Wirtz
Adresse:	Niehler Straße 330
Wohnort:	50735 Köln

Persönliches

Geburtsdatum:	01.07.1988
Geburtsort:	Willich, Kreis Viersen
Eltern:	Gabriele Wirtz, geb. Ditges; Wilhelm Wirtz
Familienstand:	ledig

Ausbildung

1998-2007:	Besuch des St. Bernhard-Gymnasiums in Willich Abschluss: Abitur (2.6)
2008-2011:	Bachelorstudium Wirschaftsmathematik, Universität zu Köln Abschluss: B.Sc. (2.4) Abschlussarbeit: "Über Verbandspolyeder" (1.0)
2011-2014:	Masterstudium Wirschaftsmathematik, Universität zu Köln Abschluss: M.Sc. (1.7) Abschlussarbeit: "Combinatorial aspects of trees in a population-genetical framework" (1.1)
Seit 2014:	Studium zur Erlangung des Doktorgrades im Fach Computational Biology

Unterzeichnet: _____

Datum: _____